Sound Matters - On the Influence of Audio Renderings on Socio-Cognitive Processing, Presence, and Fear in Social Situations in Virtual Reality

Inaugural-Dissertation zur Erlangung der Doktorwürde

der Fakultät für Humanwissenschaften

der Universität Regensburg

vorgelegt von

Sarah Roßkopf-Winderl

geboren in Zwiesel

Regensburg, 2025

Erstgutachter: Prof. Dr. Andreas Mühlberger

Zweitgutachter: Prof. Dr. Stephan Getzmann

Table of Contents

List of Abbreviations

3D ..... *three dimensional*
AI ..... *artificial intelligence*
ANOVA ..... *analysis of variance*
AOIs ..... *areas of interest*
BF ..... *Bayes factor*
bpm ..... *beats per minute*
BRIR ..... *binaural room impulse response*
BSI ..... *Brief Symptom Inventory*
CAVE ..... *cave automatic virtual environment*
DAS ..... *Dysfunctional Attribute Scale*
DFG ..... *Deutsche Forschungsgemeinschaft*
DiGA ..... *Digitale Gesundheitsanwendungen*
DRR ..... *direct-to-reverberation ratio*
ECG ..... *electrocardiogram*
EMG ..... *electromyography*
HATS ..... *head-and-torso simulator*
HMD ..... *head-mounted display*
HPEQ ..... *headphone equalization*
HRIRs ..... *head-related impulse responses*
HRTF ..... *head-related transfer function*
HTBAs ..... *head-tracked binaural auralizations*
ILD ..... *interaural level difference*
ITD ..... *interaural time difference*
M.I.N.I ..... *Mini International Neuropsychiatric Interview*
measHATS ..... *binaural auralizations based on measured BRIRs*
MPS ..... *Multimodal Presence Scale*
PANAS ..... *Positive And Negative Affect Schedule*
PHS ..... *prerecorded human speech*
RIRs ..... *room impulse responses*
RMS ..... *root mean square*
simHATS ..... *simulated BRIRs based on generic HRIRs*
simIndivHRIRs ..... *simulated BRIRs based on individually measured HRIRs*
SPIN ..... *Social Phobia Inventory*
SSQ ..... *Simulator Sickness Questionnaire*
STAI ..... *Stait-Trait-Anxiety Inventory*
SVF ..... *Stress Verarbeitungsfragebogen*
TOSTs ..... *two one-sided t-tests*
TSST ..... *Trier Social Stress Test*
TTS ..... *text-to-speech*
VR ..... *virtual reality*
VST ..... *virtual social stress scenario*
XR ..... *mixed reality*

List of figures

List of tables

Publications and contributions

The following articles, which fall within the scope of the dissertation project, have been published or are currently under review. All studies are published in or submitted to peer-reviewed journals as open-access articles under Creative Commons licenses. Accordingly, the authors retain copyright, and the articles may be shared and reused in accordance with the respective license terms, provided the original work is properly cited. The individual contributions to the study projects are outlined below:

Study 1

Roßkopf, S., Mühlberger, A., Stärz, F., Blau, M., Van de Par, S., & Kroczek, L. O. (2025a). Impact of Visual Virtual Scene and Localization Task on Auditory Distance Perception in Virtual Reality. *IEEE Transactions on Visualization and Computers Graphics. 5*, 31. 10.1109/TVCG.2025.3549855.

Contribution	*Contributor*
Conceptualization	Roßkopf, Kroczek, Mühlberger
Data curation	Roßkopf
Statistical analysis	Roßkopf, Kroczek
Manuscript – original draft	Roßkopf
Manuscript – review & editing	Roßkopf, Mühlberger, Stärz, Blau, van de Par, Kroczek

Study 2

Roßkopf, S., Kroczek, L. O., Stärz, F., Blau, M., Van de Par, S., & Mühlberger, A. (2024). The impact of binaural auralizations on sound source localization and social presence in audiovisual virtual reality: converging evidence from placement and eye-tracking paradigms. *Acta Acustica*, *8*, 72. https://doi.org/10.1051/aacus/2024064.

Contribution	*Contributor*
Conceptualization	Mühlberger, Blau, van de Par, Kroczek, Roßkopf
Data curation	Roßkopf
Statistical analysis	Roßkopf, Kroczek
Manuscript – original draft	Roßkopf
Manuscript – review & editing	Roßkopf, Mühlberger, Stärz, Blau, van de Par, Kroczek

Study 3

Roßkopf, S., Kroczek, L. O. H., Wechsler, T. F., Stärz, F., Blau, M., Van de Par, S., & Mühlberger, A. (2024). Hey AI: Can you trigger me? On the equivalency of Text-to-Speech synthesis and human speech in a virtual social stress paradigm. [*Manuscript under review*].

Contribution	*Contributor*
Conceptualization	Roßkopf, Mühlberger, Wechsler
Data curation	Roßkopf, Woltz
Statistical analysis	Roßkopf
Manuscript – original draft	Roßkopf
Manuscript – review & editing	Roßkopf, Mühlberger, Stärz, Blau, van de Par, Kroczek, Wechsler

Study 4

Roßkopf, S., Mühlberger, A., Stärz, F., Blau, M., Van de Par, S., & Kroczek, L. O. H. (2026). Aurally impressed, yet not more stressed: On the relationship between audiovisual realism, social anxiety, and presence in a virtual social stress scenario. *PLOS ONE,* 21(3): e0345565. https://doi.org/10.1371/journal.pone.0345565.

Contribution	*Contributor*
Conceptualization	Mühlberger, Blau, van de Par, Kroczek, Roßkopf
Data curation	Roßkopf
Statistical analysis	Roßkopf, Kroczek
Manuscript – original draft	Roßkopf
Manuscript – review & editing	Roßkopf, Mühlberger, Stärz, Blau, van de Par, Kroczek

Summary

Humans are inherently social beings, and social cues such as faces and voices play a central role in guiding attention and behavior. Auditory perception, particularly binaural hearing, is crucial for social cognition. It enables individuals to localize speakers and comprehend speech in noisy environments, which is essential for navigating complex social situations. Deficits in auditory processing often lead to difficulties in social situations. Moreover, mental disorders such as social anxiety are associated with impairments in social functioning. Since social functioning is closely linked to overall well-being, improving social behavior represents a key objective in psychological research.

Virtual reality (VR) is increasingly used to study and train social behavior due to its flexibility and ecological validity. However, users often report limited social presence in virtual interactions, which may reduce the effectiveness of VR-based interventions for social anxiety. This issue appears to be less pronounced in the context of non-social phobias. One contributing factor may be the dominance of visual over auditory realism in VR. Audio is often presented in mono or stereo formats, which can reduce the naturalness of interactions. This limitation may negatively impact both social and physical presence, potentially limiting the effectiveness of VR interventions. The implementation of binaural auralizations, which provide realistic, externalized audio renderings, has the potential to enhance presence by allowing voices to be perceived from their actual locations in space. This would support more realistic social virtual interactions and enable the investigation of key communication features, such as sound localization and synthetic speech, in a more realistic and controlled setting.

This thesis pursues four main research objectives. The first objective is to identify suitable behavioral and subjective evaluation methods for assessing the degree of realism achievable through binaural auralizations in audiovisual VR. Additionally, the interplay between these evaluation methods and the virtual visual scene is examined to address VR-specific perceptual challenges. The second objective is to evaluate immersion, realism, and perceptual audio quality across various binaural auralization techniques and to investigate their impact on presence in VR. The selection of auralizations for subsequent studies was informed by these findings. The third objective is to examine audio effects in a socially stressful VR scenario and to compare synthesized speech with recordings of natural human speech. The equivalence of synthesized speech to natural speech in eliciting social-evaluative threat and social presence was tested. The fourth objective is to investigate the effects of

binaural auralizations on affective responses, presence, and attention within the socially stressful VR scenario. Differential effects in low- vs. high-social stress conditions as well as the influence of social anxiety were explored.

In **Study 1**, the influence of the virtual visual scene and the measurement task on sound source localization and auditory distance perception of physical sound sources in VR was investigated. Two sequentially designed experiments were conducted with a total of 60 participants, who indicated the perceived source positions of sounds emitted from loudspeakers placed within the physical room but invisible to them. The findings from the first paradigm showed that audiovisual room incongruence negatively affected localization accuracy but had no impact on presence or perceived realism. In the second paradigm, participants localized sound sources either in a fully visible (congruent) virtual room or in a virtual room with reduced visibility. Distance estimation was performed using one of three methods: placing a virtual loudspeaker, walking, or verbal reporting. While room visibility did not directly affect perception, it interacted with the task. Specifically, distance overestimation was greater when using the placement task in the reduced-visibility scene. Additionally, presence and adverse effects were influenced by both variables.

In **Study 2**, two measurement paradigms were employed to investigate localization accuracy for loudspeakers (physical sound sources) and four audio renderings (virtual sound sources). Forty-nine participants indicated source positions using two methods: first, a placement task, and second, a naturalistic gaze-behavior paradigm. No differences were observed between binaural renderings and loudspeaker trials in ratings of social presence and subjective realism, although localization accuracy for renderings was slightly lower than for loudspeakers. Furthermore, the audio rendering that required the least time and technical equipment (simulated generic condition) did not perform worse than the other renderings. In contrast to the other conditions, the anchor audio was predominantly not perceived as externalized. Consequently, the anchor was inferior in localization accuracy, social presence, and perceived realism. A strong positive correlation was observed between social presence and subjective realism.

While the first two studies examined fundamental variables of auditory perception in VR and the hypothesized positive influence of auditory realism on quality of experience, particularly on social presence, the subsequent two studies aimed to transfer these findings to more complex virtual social interactions involving social-evaluative threat. In **Study 3**, artificial intelligence (AI)-generated text-to-speech (TTS) synthesis was compared to human voice recordings in eliciting psychosocial stress and social presence. To this end, the Trier

Social Stress Test (TSST) was used as the experimental gold standard to assess human stress responses. Both audio conditions provoked substantial stress responses in 40 participants, reflected in increased heart rate and subjective stress ratings. Furthermore, presence and affective state did not differ significantly between the audio conditions, highlighting the utility and practicality of synthetic speech in virtual social interactions.

In **Study 4**, the effects of audiovisual realism within a virtual social stress scenario were examined in a sample of 78 participants. Furthermore, the level of social-evaluative threat was varied. A control "low-stress" group performed a task with the prompt to "test" a virtual job interview and read aloud preformulated answers. In contrast, the high-stress group performed a self-promoting talk and answered demanding job interview questions. Consequently, the high-stress group showed higher stress responses, as indicated by increased salivary cortisol, heart rate, and self-reported stress, compared to the low-stress group. Realism and sound externalization were rated higher when head-tracked binaural auralizations (HTBAs) were used than when diotic renderings were used. However, neither social presence, nor stress responses, nor gaze behavior was affected by auditory realism. Throughout all stress groups, elevated levels of arousal were reported, potentially masking audio effects on stress and social presence.

Across all four studies, the effects of participants' social anxiety levels on experimental variables were examined. Results from **Studies 1** and **2** indicate that individual social anxiety did not influence auditory distance perception. However, effects on affective states were observed. In **Study 4**, but not in **Study 3**, higher social anxiety was associated with increased subjective stress and elevated heart rate. No consistent effect of social anxiety on social or physical presence was found. Nevertheless, the subjective experience of the virtual interaction, such as dominance and pleasantness, was influenced by social anxiety.

Overall, the findings underscore the importance of investigating VR-specific auditory perception and its role in immersion. The complex interaction between auditory and visual modalities should be considered when examining perceptual processes in virtual environments. Moreover, auditory realism positively influences both social and physical presence. Advanced audio rendering techniques have the potential to enhance the quality of virtual social interactions, though their impact varies considerably by context. Auditory immersion appears most effective in scenarios involving low to moderate affective arousal. Still, it may be less critical in highly affective VR applications, such as those used for treatments or training in the context of social anxiety. Nevertheless, practical improvements, such as leveraging TTS technology and simplifying the integration of HTBAs for applied use,

will facilitate the incorporation of advanced audiovisual virtual scenes into psychological research.

1 Motivation and outline

I envision my thesis defense as a meaningful yet nerve-wracking moment. Ideally, some of my loved ones will be there sharing in this pivotal moment, alongside curious minds drawn by the topic of my dissertation, which, in my admittedly biased opinion, is fascinating and refreshingly multifaceted. Naturally, the dissertation committee will be present as well, right in the front row, whose presence and evaluative intent make even seasoned researchers sweat. The room, maybe one of the university's beautiful brutalist-style auditoria, will be filled with quite a crowd.

I will be standing in front of an audience, probably delivering my presentation with a shaky voice, trying to sound confident, and hoping my heartbeat cannot be heard throughout the room. After my certainly convincing presentation, it is time for questions from the audience. Even without lifting my gaze from the slides, I can sense the presence of an engaged or perhaps critical questioner. From the tone of their voice alone, I might infer their mood, their intention, whether they are challenging or encouraging. Maybe I may even close my eyes to better comprehend the question, with (metaphorically) half of my brain being occupied with worries about failure.

In such a real-world scenario, my brain automatically localizes the speaker, drawing on finely tuned spatial cues processed through both ears. I would immediately know whether the voice comes from the nearby committee or the back of the room, suggesting a more personal connection and potentially a supportive question.

To prepare myself for this demanding situation and practice the talk as if I were really in that situation, the scenario could be recreated in VR. When discussing VR, images of VR glasses and futuristically designed worlds probably come to mind. Online searches for VR typically yield pictures of individuals wearing a head-mounted display (HMD), highlighting the dominant role of visual perception in humans (Blauert, 1997) and in psychological research (Hutmacher, 2019). VR research itself tends to emphasize visual aspects. During the initial phase of this doctoral project, a literature search in the PsycInfo database using the terms "auditory" or "auditory cognition" combined with "virtual reality" in the abstract returned no results. A subsequent search using "virtual reality" and "aud*" in the title produced 25 hits, 10 of which included the term "audiovisual". In contrast, replacing "aud" with "vis*" yielded 135 results.

To create the VR scene as realistically as possible, it would require implementing spatialized sound, e.g., via binaural auralizations. These auralizations are the auditory

counterpart presented over headphones to visualize a scene via an HMD. While the effects of visual realism in VR have been extensively studied, also in the context of affective scenes, the role of auditory realism in shaping virtual social interactions remains underexplored.

This research gap forms the foundation of my dissertation project, which investigates the role of audio renderings in virtual social interactions. The project explicitly examines their impact on socio-cognitive processing, the sense of presence, and affective responses. The broader context of this work lies in the investigation and treatment of social anxiety and anxiety disorders. Accordingly, particular attention was given to the interplay between social anxiety, auditory processing, and the influence of different audio rendering techniques.

To address these research goals, four empirical studies were conducted. This work was made possible through close interdisciplinary collaboration with acousticians from the Jade Hochschule Oldenburg and the Carl-von-Ossietzky University of Oldenburg. The collaboration took place within the framework of the Priority Program SPP2236 “Audictive” (https://gepris.dfg.de/gepris/projekt/422686707) of the Deutsche Forschungsgemeinschaft (DFG) as part of the project “Einfluss des Audio-Renderings in virtuellen Umgebungen auf Realismus, Präsenz und sozio-kognitive Verarbeitung” (https://gepris.dfg.de/gepris/projekt/444832396).

This thesis is structured into three main parts. The first part provides the general theoretical background on auditory perception, emphasizing its role in socio-cognitive processing. Furthermore, the principles of binaural hearing will be discussed, with a particular focus on sound source localization and distance perception, two variables examined in the first two studies of this dissertation. Then, an overview of the current state of research on audio rendering techniques is provided.

Audio renderings refer to simulations of auditory events designed to evoke realistic auditory impressions. In the context of this work, the term ‘audio renderings’ refers explicitly to either binaural auralizations, which are used to create spatially immersive hearing impressions, or text-to-speech (TTS) synthesis, which converts written text into spoken output. Both are relevant techniques for simulating convincing virtual environments with a focus on social interactions. Accordingly, the role of VR as a valuable tool in both basic and applied psychological research will be discussed, along with several key applications involving social virtual interactions. To support a comprehensive understanding of the dependent variables examined in the empirical studies, this section introduces key metrics used in audiovisual VR research on social interaction. These concepts are further explored in relation to social anxiety. Finally, the current state of research on the influence of acoustic

scenes on VR experiences will be reviewed, followed by an identification and discussion of the existing research gap. This section concludes with a derivation of the overarching research objectives that guide this thesis.

The central part comprises four empirical studies, three of which have been published in peer-reviewed journals with open access. The remaining study has been submitted and accepted for review in peer-reviewed journals. All four studies were formatted consistently to improve the dissertation's readability, with a combined reference section and continuous numbering of figures and tables.

In the final part of this thesis, a summary of the findings obtained throughout the dissertation project is provided. The scientific advancements achieved through the empirical studies are discussed, along with their contributions to the overarching research objectives. In addition, practical implications are outlined, and directions for future research on audiovisual virtual interactions in the context of mental health are proposed.

2 Theoretical background

2.1 Socio-cognitive processing and the role of binaural hearing

One of the objectives of this dissertation is to investigate the effects of audio renderings on socio-cognitive processing in VR; therefore, an initial definition of this construct will be provided. Social cognition concerns "how people make sense of other people and themselves in order to coordinate with their social world" (Fiske & Taylor, 2020). Humans are thoroughly social beings, and therefore, the human brain is adapted to process information about other human beings and social experiences. To create convincing artificial humans for virtual interactions intended to trigger social behavior, findings from basic social cognition research may be helpful.

People are not processed in the same way as things. What separates people from things are characteristics such as being intentional causal agents, mutuality of perception and evaluation, similarity to the perceiver, high complexity, and flexibility within time and situation. However, anthropomorphism can occur in response to robots and avatars, as well as towards animals and cars (Epley et al., 2007). Research is ongoing on how to reinforce thinking and feeling about an object as a human counterpart and behaving like that (Darling, 2015). Robots can even evoke romantic feelings, at least on an anecdotal basis (Ebner & Szczuka, 2025). Anthropomorphism is not surprising considering the flexibility with which people categorize fellow human beings. Cars can be humanized (Epley et al., 2007). At the same time, out-group beings such as homeless or drug-addicted people can be dehumanized, associated with decreased activation in brain areas involved in processing of social stimuli, and accompanied by increased activation of the disgust-related regions (Harris & Fiske, 2006).

A relevant contributor to the humanization, e.g., of cars, is the human tendency to recognize faces (Fiske & Taylor, 2020). The human brain is tuned to process social stimuli, such as faces. Car designers use this fact to create an aggressive or friendly appearance. Not just faces but also human voices trigger specialized socio-cognitive processing. Rich information can be retrieved from voices, including verbal content, speakers' gender, and age (Cacioppo & Decety, 2011). Even the affective state of the speaker and nonverbal content (e.g., sarcasm) can be extracted from the prosody (Schirmer & Kotz, 2006).

In general, auditory perception plays a crucial role in social cognition. Vocal social cues are an essential driver of human attention, and emotions modulate this effect. Angry and fearful voices are processed very quickly and can activate the limbic structures (Brück et al.,

2011). Also, social information from the auditory signal is contextualized with spatial information (Kroczek et al., 2024). The interconnection between social and spatial auditory processing is also reflected at a neurobiological level. The superior temporal sulcus is essential for the integration of audio(visual) signals and more generally for behaviorally relevant stimuli (Beauchamp et al., 2004). This region is also involved in social perception, which is conceptualized as the analysis of stimuli containing socially relevant information (Allison et al., 2000).

Additionally, proximity in sensory, temporal, or spatial terms makes a social stimulus more salient, meaning more likely to attract attention (Finisguerra et al., 2015; Fiske & Taylor, 2020; Latané et al., 1995), emphasizing the relevance of spatial perception for social cognition. A commonly drawn distinction in this context is between intimate, personal, social, and public space, with the first two referring to interactions within relatively close relationships (Latané et al., 1995; Lloyd, 2009). For this dissertation, only social and public spaces are considered relevant. These spaces begin at approximately 1.20 meters, and from this distance onward, the sound sources used in the subsequent studies are applied.

Salience can be driven both bottom-up by stimulus properties and top-down by context or prior knowledge. Two semi-motivational systems guide behavior: the behavioral inhibition system, linked to avoidance of aversive stimuli and negative affect, and the behavioral approach system, associated with approach toward rewarding stimuli and positive affect (Fiske & Taylor, 2020). These systems are highly relevant in social interactions (Kimbrel et al., 2010; Reichenberger et al., 2017). Socio-cognitive processing involves perceiving and responding to social information, including anthropomorphized objects. Affect plays a key role and is influenced by temporal and spatial factors (Fiske & Taylor, 2020), highlighting the role of spatial perception in speech and voice processing.

When interacting with others, individuals rely on binaural cues to localize the speaker and to analyze speech effectively in complex environments. Especially when visual cues are reduced or unavailable, binaural hearing is crucial for detecting social stimuli. In this case, the ears typically lead the eyes towards the speaker. By directing the gaze towards a speaker, we not only signal attention (Fiske & Taylor, 2020; Foulsham & Sanderson, 2013) but also enhance the discrimination of auditory spatial cues (Maddox et al., 2014). In group settings, or as illustrated in my initial example, when addressing an audience, binaural cues are essential for briefly identifying the speaker and therefore for effective communication.

Early work on selective auditory attention found that when two different speech streams are presented simultaneously (dichotic), listeners can selectively attend to one ear and

essentially ignore the other stream (Cherry, 1953). Known as the cocktail party phenomenon, this ability to filter auditory information at an early sensory stage underlines the importance of binaural hearing in complex social interactions (Bidelman et al., 2025; Bronkhorst, 2000; Getzmann et al., 2023). Given that the auditory system mainly processes mechanical stimuli, such as sound waves that cause vibrations of the eardrum and movement of the ossicles in the middle ear (Blauert, 1997), it is remarkable that speech can still be accurately perceived even in noisy environments (Bronkhorst, 2000). Reflecting its specialization in social cognition, the human brain is effective at processing human speech. In the relevant frequency band, which reaches from 250 Hz to 4000 Hz, the perception threshold is low, and the frequency resolution is high, enabling speech sounds to be finely differentiated (Krumbholz, 2009).

Further evidence for the crucial role of binaural processing for social cognition comes from research on neurodevelopmental and psychiatric disorders. For example, people with autism experience difficulties in social interactions, especially in interpreting social cues. Also, auditory processing is altered in people with autism; reduced performance in sound source localization (non-social stimuli) has been observed (Fujihira et al., 2022). Furthermore, individuals with schizophrenia show deficits in auditory processing, which contributes to the social disability associated with this disorder (Javitt & Sweet, 2015). Also, binaural processing is altered; specifically, the discrimination of interaural time differences is impaired (Matthews et al., 2013). To sum up, binaural hearing is crucial for social interactions because it enables speaker identification, improves speech processing in noisy environments, and likely supports social development and functioning. In the following section, the fundamentals of spatial hearing will be introduced.

2.2 Fundamentals of spatial hearing

Spatial hearing is an interdisciplinary research area that integrates psychology, psychophysics, physiology, otolaryngology, physics, engineering, and musicology. In a standard work on spatial hearing, the author outlines that there is no such thing as non-spatial hearing (Blauert, 1997). Each sound is automatically assigned to a location; only the accuracy varies. The auditory percept is shaped by the sound pressure wave at each eardrum and its physical attributes, such as energy, frequency, temporal structure, and density, with interaural differences playing a critical role in spatial hearing.

Auditory spatial perception is predominantly shaped by binaural cues, including interaural level and time differences, with additional contributions from binaural spectral cues (see Figure 1). The pinna, with its relief-like shape, filters sound spectrally and reduces wind

noise (Blauert, 1997). The auditory canal ends at the eardrum and shapes the perceived sound. Each person has an individual head-related transfer function (HRTF), which describes how their pinna, head, torso, and ear canals filter incoming waves based on the sound source's direction and distance.

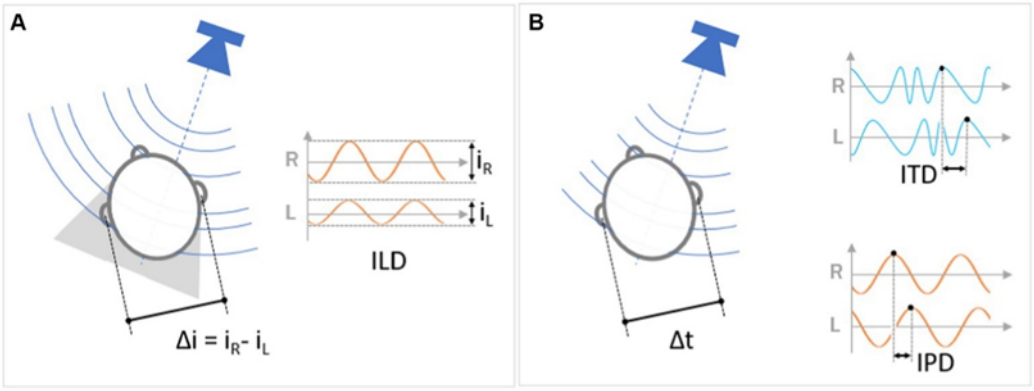

Figure 1: Binaural cues in spatial hearing. A: Interaural level difference (ILD) and B: Interaural time difference (ITD) when a sound source is presented frontally right to a perceiver. Reproduced from Carlini et al., 2024, licensed under CC BY.

Humans have spatial hearing limited to a head-centered coordinate system defined by azimuth, elevation, and distance, lacking moveable ears. Under optimal conditions, humans can localize sound sources in the frontal azimuth with an accuracy of approximately 1°, which is about half the spatial resolution of the visual system (Blauert, 1997). Under suboptimal conditions, accuracy is degraded significantly (Blauert & Braasch, 2020). However, localization accuracy can be improved by repeated exposure or speaker familiarity. Head movements furtherly enhance the spatial resolution (Blauert, 1997).

Auditory distance perception is a fundamental aspect of spatial hearing, with relevance to social behavior, as the characteristics of auditory distance cues vary with changes in peripersonal space (Sorokowska et al., 2017). Similar to directional sound localization, it is substantially less accurate than visual distance perception, and in fact, even more imprecise. Auditory distance estimates are often subject to considerable error, typically underestimating distant sound sources and overestimating nearby sources, particularly those within one meter (Zahorik et al., 2005). It has been suggested that the underestimation of distant sound sources may reflect an evolutionary mechanism providing an 'extra margin of safety' by enhancing preparedness for potential threats. This hypothesis is further supported by findings indicating that approaching objects are processed with perceptual priority (Ghazanfar et al., 2002). The estimation of auditory distance relies on integrating various acoustic cues, including intensity, the direct-to-reverberant energy ratio (DRR), spectral characteristics, and binaural cues. However, visual input also significantly influences auditory distance perception, often

overriding or modulating auditory cues, particularly in multisensory environments (Zahorik, 2001; Zahorik et al., 2005).

Typically, the auditory percept is integrated with input from other sensory modalities, including the visual, vestibular, and somatosensory systems (Blauert, 1997; Kadunce et al., 2001). As outlined above, head movements are also crucial, and to a lesser extent, bone conduction and tactile feedback of sound waves. On a neurobiological basis, this is reflected by bidirectional projections between the inferior colliculus (the primary auditory system) and the superior colliculus, where multimodal input is integrated (Kadunce et al., 2001). These pathways may be relevant for spatial attention and spatial learning.

Visual dominance over auditory processing is a well-documented phenomenon. To begin with, the visual capture effect, also known as the ventriloquist effect, describes the localization of an auditory event toward a plausible visual target, as long as the spatial or temporal disparity does not exceed certain thresholds of implausibility (Jack & Thurlow, 1973; Mershon et al., 1980). This may result in perceptual shifts of up to 30° in azimuthal integration (Slutsky & Recanzone, 2001). An even more pronounced manifestation of this dominance is the proximity image effect, where visual input can entirely override auditory distance cues. For instance, a sound may be perceived as emanating from a visible loudspeaker in the front row, regardless of its actual origin (Mershon et al., 1980). In summary, auditory localization relies on mechanisms such as binaural processing, is comparatively spatially imprecise, and is often strongly influenced by visual input and context.

2.3 Audio renderings

Audio renderings serve as the acoustic counterparts to visualizations. Just as interactive head-tracked VR devices allow users to explore virtual environments with their visual senses, audio renderings enable users to perceive and engage with these environments through sound. More technically, “auralization is the technique of creation and reproduction of sound on the basis of computer data” (Vorländer, 2008). Although Vorländer’s definition originates in the field of spatial audio, it can be broadened to include a broader range of applications. Within the scope of this thesis, audio renderings are defined as simulations intended to evoke realistic auditory experiences. When referring to audio renderings, this thesis refers either to binaural auralizations, which aim to create immersive, three-dimensional hearing impressions, or TTS synthesis, which converts written text into speech. Together, these techniques play a crucial

role in simulating convincing virtual environments, particularly those focused on social interaction.

2.3.1 Binaural auralizations

Auralization refers to techniques that make a room audible. They use psychoacoustic knowledge of spatial perception, outlined above, as well as fundamentals of acoustics, including wave propagation, reflection, and absorption. These principles are fundamental to understanding sound behavior in enclosed spaces and are therefore essential for acoustic engineering, particularly for simulating sound fields and modeling wave propagation using differential equations (Vorländer, 2008). In particular, these techniques simulate how a virtual sound source would generate sound pressure at the left and right ears, depending on the positions of both the source and the receiver. Also, the acoustic properties of the to-be-simulated scene, such as room dimensions, absorption characteristics, and source directivity, are accounted for in the rendering (Vorländer, 2008). There are various techniques for simulating acoustic behavior in enclosed spaces, including geometrical acoustics methods such as ray tracing and wave-based approaches. A central element in these simulations is the generation of room impulse responses (RIRs), which characterize how sound propagates from a source to a receiver within a given environment.

Virtual sound sources can be auralized using either loudspeakers or headphones. For psychological research on virtual social interactions, headphones are considered the more practical choice in this work. They integrate seamlessly with HMDs, and most contemporary HMDs already come equipped with built-in headphones that can be used to present auralizations. To produce a realistic spatial hearing impression over headphones, HRTFs are incorporated based on the incidence of the direct sound or reflections (T. Wendt et al., 2014). This results in a so-called binaural room impulse response (BRIR) and is simulated for several receiver head positions with varying spatial resolution (Lindau & Weinzierl, 2009). The BRIRs are further convolved with an anechoic audio recording. By incorporating real-time head-tracking of the listeners' head position and dynamically selecting the corresponding BRIR for convolution (Jaeger et al., 2017), it becomes possible to maintain the perception of a static virtual sound source even during head movements. Conveniently, the real-time head-tracking capabilities of modern HMDs can be used for this purpose. So-called head-tracked binaural auralizations (HTBAs) enhance the interactivity and realism of the auditory experience (Blau et al., 2021; Stärz et al., 2025).

Ideally, the perceptual illusion of a physical sound source can be created using HTBAs (Brandenburg et al., 2020). In both natural hearing and auralizations, perception does not directly mirror physical stimuli. Instead, it represents a mental construct formed from sensory input. According to Helmholtz's theory of perception, sensations serve as indicators of external stimuli rather than reproducing them (Blauert & Braasch, 2020). Auditory illusions arise when perception diverges from physical reality due to misleading input. They reflect unconscious inferences that usually support efficient perception but can lead to errors when their assumptions are invalid. An intended perceptual illusion in the context of auralizations is the externalization of a sound presented via headphones. Externalization is the phenomenon in which a sound source is perceived as located in the surrounding environment rather than originating from inside the head, which is typically the case with standard headphone playback without binaural auralizations (Best et al., 2020). An externalized virtual sound can therefore be considered more realistic than an internalized one. Another critical quality criterion is plausibility, which refers to whether the virtual sound source meets the listener's expectations of the physical one (Lindau & Weinzierl, 2012).

Alternative evaluation metrics include perceptual attributes such as authentic timbre, distortion, and tone richness (Brinkmann et al., 2017; Weinzierl et al., 2018). The impact on the emotional state (e.g., measured via subjective or physiological measures) and overall experience quality in a VR scene that includes HTBAs can also serve as an evaluation criterion. Measures such as presence, discussed later, provide insight into the effectiveness of these audio systems (Nicol et al., 2014). However, consistency of HTBA ratings is sometimes low, even among listeners with high expertise in virtual acoustics (Stärz et al., 2025). Spatial accuracy offers a more objective measure. For instance, the time needed to localize a virtual sound source and the overall localization precision in azimuth, elevation, and distance can indicate the performance of HTBAs (Jenny & Reuter, 2020; Nicol et al., 2014). Virtual sound sources that can be localized with similar accuracy as physical ones can be regarded as highly realistic. The significant impact of visual input on sound source localization has been previously demonstrated. This relationship remains highly relevant in the context of audiovisual virtual environments. Research has shown that visual distances and object sizes appear compressed when viewed through a head-mounted display (Buck et al., 2018) or on large-screen displays, such as cave automatic virtual environments (CAVE) systems (Ghinea et al., 2018). Consequently, studies investigating sound source localization or distance estimation of virtual sound sources in virtual reality should account for the potential

interaction between the visually compressed scene and the corresponding acoustic environment.

Auralizations can be employed in architectural acoustics, as they allow listeners to perceive how a space would sound before its physical construction or modification. Beyond architecture, auralization techniques are also valuable in fields such as vehicle design and noise control, where they can support the evaluation and optimization of acoustic environments (Vorländer, 2008). Finally, another important application of binaural auralizations, and the primary focus of this work, is their use in VR, where they support immersive and perceptually accurate auditory experiences within simulated environments (Stärz et al., 2022).

2.3.2 Text-to-speech synthesis (TTS)

TTS refers to techniques that make written text audible, as if spoken by a human voice. This process relies on databases of recorded speech to generate new utterances (Khanam et al., 2022). In recent years, the field has shifted from traditional machine learning methods to more advanced deep learning approaches within the broader domain of AI (Kaur & Singh, 2023). The overarching goal is to generate speech that is both intelligible and natural-sounding. TTS synthesis typically involves two main stages: first, the input text is analyzed and transformed into a phonetic representation; second, this representation is used to generate speech waveforms. Evaluation of TTS systems involves both objective and subjective metrics. Objective assessments include algorithmic comparisons and signal-based evaluations. In contrast, subjective evaluations often rely on Mean Opinion Scores, in which human listeners rate speech quality on a scale from poor to excellent (Cooper et al., 2024). Other methods use phoneme-distance metrics to analyze phoneme-level errors (Pirklbauer et al., 2023) and data-driven approaches to evaluate synthetic speech (Williams et al., 2020). Typically, the more closely a TTS system resembles natural human speech, the higher its overall quality is considered.

A study with over 1,000 participants found several TTS systems to be non-inferior to human speech in comprehension and pleasantness (Cambre et al., 2020). Furthermore, it was suggested that voice quality should be evaluated based on application-specific criteria rather than a single metric. Advancements in TTS have even enabled the generation of synthetic voices that are realistic enough to deceive both human listeners and automated systems (Wenger et al., 2021). These voices can be tailored to reflect specific characteristics, such as age and gender, or to replicate the voice of a particular individual (Kaur & Singh, 2023).

Although the latter raises serious concerns regarding misuse and identity theft, TTS renderings offer substantial benefits, as evidenced by their widespread integration into everyday applications, including navigation systems (e.g., TomTom, Google Maps), digital assistants (e.g., Siri, Alexa), e-learning platforms, customer service interfaces, YouTube videos, and conversational AI tools like ChatGPT. Also, for VR research, TTS provides advantages over pre-recorded human speech. Beyond reducing costs and efforts, it enables greater flexibility and standardization in the speech output of virtual agents, which is particularly valuable in research settings such as clinical studies, where consistent and adaptive interaction is essential. Automated TTS responses reduce variability introduced by human speakers, including differences in accent, intonation, and delivery style. Although emotional voice synthesis has been an active research area for many years, current methods still fall noticeably short of the quality achieved by human voices, even after decades of development (Schuller & Schuller, 2018; Zhou et al., 2022).

2.4 Virtual social interactions

Audio renderings provide the technological basis for creating acoustically realistic virtual scenes. With its origins in the gaming industry, VR nowadays has widespread applications across several scientific fields. It also plays a vital role in psychological research due to its advantages, such as high controllability, standardization, and the ability to create ecologically valid environments (Kothgassner & Felnhofer, 2020; Kyrlitsias & Michael-Grigoriou, 2022). These strengths have led to its frequent use in experimental settings. The high relevance of VR is reflected in the programs of major German psychology conferences, including Psychologie und Gehirn (2022) and the Deutsche Psychotherapie Kongress (2024), where VR-based studies are regularly presented. An essential application of VR in psychological research is virtual social interactions. One of the major advantages is that the social interaction can be kept constant across participants (Pan & Hamilton, 2018). Manipulations can be accomplished unobtrusively, since the scene is rich in stimuli and therefore subtle changes often remain unnoticed at an explicit level. Furthermore, manipulations can be conducted efficiently that would be impossible in real life, such as changing only particular features of a person while others remain untouched (Bombari et al., 2015).

Also, ethical concerns can be mitigated by testing virtual social interactions, e.g., in the well-known Milgram obedience experiment, in which the effects of observing electroshocks from one person to another are investigated (Peña et al., 2024; Slater et al.,

2006). By using virtual agents rather than real humans as recipients of shocks, no physical suffering is inflicted.

Therefore, virtual social interactions offer more standardization and flexibility than the investigation of real social interactions. At the same time, the ecological validity of virtual social interactions can be assumed to be substantially higher than that of investigations that typically involve high standardization, e.g., using vignettes and pictures or videos displayed on a non-immersive screen (Kothgassner & Felnhofer, 2020). Several well-known findings in social psychology have been successfully replicated using social virtual interactions, for a review see (Bombari et al., 2015).

2.4.1 Applications related to social fear

Virtual social interactions can be used in research on the basic principles of social behavior and socio-cognitive processing, as well as in applied psychological research, such as social skills training (see Figure 2). During the coronavirus pandemic, virtual teaching became a highly relevant issue, though it had been explored earlier. This includes asynchronous online platforms and VR classrooms (Dai et al., 2024; Gillies & Pan, 2018; Liou & Chang, 2018). The ecological validity of VR allows manipulation of learning environments, making it a valuable tool in educational research. A key finding is that VR classrooms enhance learner motivation (Cabero-Almenara et al., 2019; Liou & Chang, 2018). Virtual social interactions in psychological research encompass a broad spectrum of applications, ranging from basic research, such as the replication of experiments like the Milgram study, to applied approaches including leadership training and therapeutic interventions (Alcañiz et al., 2024; Bombari et al., 2015; Gillies & Pan, 2018; McCall & Blascovich, 2009). This section will focus on applications in clinical psychology, with particular emphasis on social anxiety.

To begin with, virtual social interactions can be used in research on the fundamental principles, such as the genesis or the maintenance of social anxiety. Social anxiety is the fear of being negatively evaluated or being rejected in social or performative situations (American Psychiatric Association, 2013). Patients suffering from this mental disorder avoid social situations or endure them with increased distress, which leads to an impairment of daily functioning and quality of life. Typical situations that trigger social anxiety are public speaking, meetings with unknown people, or, more generally, being the center of attention. It was shown that social fear can be successfully conditioned in virtual interactions (Reichenberger et al., 2017). Also, the conditioned fear could be successfully extinguished in

VR, indicating the significance of virtual social interactions in both the investigation and the treatment of social anxiety (Reichenberger et al., 2020).

Figure 2: Public speaking training in front of a supportive (left) and unsupportive (right) virtual audience. Reproduced from Kroczek & Mühlberger, 2023, licensed under CC BY 4.0.

Furthermore, VR can be used to practice appropriate behavior in social interactions, especially when they are demanding such as self-assertive behavior. Social skills training in VR can elicit physiological arousal and enable the learning of new behavior (Reichenberger et al., 2022). At the same time, physiological parameters and gaze behavior can be monitored smoothly, which would not be as easy in real-life interactions. Monitoring gaze, e.g., avoidance of eye contact, can be used to enhance the new behavior (Schmidt-Peter et al., 2025). VR training for social skills has been shown to perform better than alternative training approaches (Howard & Gutworth, 2020).

Building on these advantages of VR for social skills training, its potential extends beyond behavioral learning to the systematic investigation of stress responses in controlled environments. The Trier Social Stress Test (TSST) is the laboratory gold standard for investigating stress, especially in the neuroendocrinological domain (Kirschbaum et al., 2008). Using this standardized tool, researchers can systematically examine the neurobiological substrates of social anxiety disorder and explore its links to comorbid mental and somatic conditions. (Grace et al., 2022). The virtual version of the TSST (VR-TSST) has proved to be an appropriate alternative to the time-consuming and high-organizational-effort standard laboratory procedure involving at least six persons (Goodman et al., 2017). The VR-TSST is highly standardized and logistically more efficient. Also, equal and even higher subjective stress ratings were found in a study comparing an in-vivo-TSST to a VR-TSST (Shiban et al., 2016). Regarding neuroendocrine stress responses, the evidence is ambiguous, ranging from equal responses to reduced reactions in the VR-TSST (Bahr et al., 2021; Shiban et al., 2016; Zimmer et al., 2019). It has been suggested that lower levels of social presence and, therefore, reduced experience of social-evaluative threat in the VR version may contribute to the observed differences in stress responses. Social presence, an essential metric

in social virtual interactions, will be outlined in the subsequent section (2.4.2). Previous research has identified the social-evaluative component as the primary trigger of acute stress during the TSST (Allen et al., 2017; Frisch et al., 2015), which might be reduced when social presence is low. Therefore, enhancing social presence in VR represents a key objective. Subsequent sections will explore current findings on the determinants of social presence, including the role of acoustic realism.

Finally, an important application of virtual social interactions is their use in treating social anxiety disorders. With almost 4.5% point prevalence in a European population, its curation is highly relevant to the mental health care system (Ohayon & Schatzberg, 2010). In the German evidence-based recommendations for the diagnosis and treatment of social anxiety disorder, it is stated that the first-choice treatment is cognitive behavioral therapy, including exposure therapy (Bandelow et al., 2014). This implies that the patients actively seek experiences that are usually avoided or feared with the guidance of therapists. Virtual reality exposure therapy is a time-efficient and effective alternative to in vivo exposure therapy (Wechsler et al., 2019). However, in the treatment of social anxiety disorders, VR-exposure was found to be slightly less effective than in vivo exposures. The authors again suggest that virtual agents in VR scenes elicit less social presence and reduced social evaluative threat. In general, the effectiveness of VR applications has been shown to increase with immersion (Wiebe et al., 2022). Therefore, a major goal is to explore ways to increase immersion and social presence in virtual social interactions.

2.4.2 Immersion, realism, and presence

Immersion and presence are key metrics for evaluating the quality or effectiveness of virtual experiences. Immersion refers primarily to the technological aspects. Presence, in contrast, captures the psychological dimension, specifically the users' feelings, perceptions, and behaviors within the virtual environment (Slater & Wilbur, 1997). A construct named “perceptual immersion” describes “the degree to which a virtual environment submerges the perceptual system of the user” (Biocca & Levy, 2013), indicating the importance of multimodal stimulation. In the field of virtual acoustics, the term plausibility commonly refers to the perceived credibility of a simulation (Lindau & Weinzierl, 2012). Within VR research, a concept with a seemingly similar meaning is referred to as realism, or often visual realism. Although the terminology differs across domains, both constructs aim to capture the degree to which a virtual experience aligns with users’ expectations based on real-world perception (Chalmers & Ferko, 2008; Newman et al., 2022). Realism is also a construct in the

investigation of TTS, particularly in efforts to make TTS indistinguishable from recordings of real human voices. In both virtual acoustics and TTS research, specific simulations have achieved levels of realism that closely approximate real-world conditions (e.g. Brinkmann et al., 2017; Wenger et al., 2021). However, in the domain of visual VR, a comparable degree of realism has yet to be attained and, at least from the perspective of this work, remains beyond conceivable limits. In mixed reality settings, simulated environments remain distinguishable from real-world scenes. To date, there appears to be no published evidence demonstrating that users cannot reliably distinguish between real and simulated visual input. In contrast, for HTBAs and TTS, the indistinguishability of real and rendered stimuli has been shown (e.g. Brinkmann et al., 2017; Wenger et al., 2021). Realism and immersion are therefore closely interrelated concepts, with the latter describing the extent to which a technology can deliver an "illusion of reality" (Slater & Wilbur, 1997).

Immersion, as defined by Slater and Wilbur (1997), can be measured objectively. It increases with the system's ability to exclude physical reality, provide an extensive scene, deliver vivid sensory input (e.g., a photorealistic visual scene), and ensure consistent multimodal feedback. Interactivity with the scene is essential, e.g., with head movements resulting in corresponding changes, such as a shifted visual scene and spatially stable sound-source positions. Accurate tracking, matching of sensory input, and an egocentric perspective therefore support immersion.

While immersion describes the technological capabilities of a VR system, presence refers to a subjective state of consciousness, the "sense of being in the virtual environment" (Slater & Wilbur, 1997). Presence encompasses users' responses, emotions, and behaviors within VR. In a later definition, Lee (2014) abstains from using terms such as illusion due to their normative connotations and emphasizes that perception is always a distorted version of human sensation, and even more so of the physical scene. Presence is thus defined as "a psychological state in which virtual [..] objects are experienced as actual objects in either sensory or non-sensory ways." (Lee, 2004). Perceptual realism, described as "life-like creation of the physical world by providing rich sensory stimuli" (Lee, 2004), contributes to the degree of presence in VR. Physical, social, and self-presence are distinguished, with social presence defined as "[…] experiencing the representation of other humans who are connected by technology." Social realism refers to the perception of virtual people as being comparable to real people. This applies to both avatars and artificial agents, implying "experiencing artificial objects manifesting humanness" (Lee, 2004).

These concepts highlight the multifaceted nature of metrics used to assess the quality of virtual experiences. Social presence, in particular, is a broad construct. Biocca et al. conceptualize it in analogy to spatial or physical presence, which concerns the sense of being located in a virtual space. They define social presence as "the sense of being together" (Biocca et al., 2003), emphasizing that the representation of sentient others involves more than spatial positioning. Social presence can arise not only through interaction with humans but also with computers, including AI (Rogers et al., 2022). The degree of social presence is suggested to depend on three central factors: first, the extent to which an entity is perceived as sentient; second, the capacity of a communication medium to transmit social cues; and third, the consistency of an entity's behavior (Schott et al., 2025; Skarbez et al., 2018). In the literature, a distinction is sometimes drawn between social presence, which involves interaction with an agent or avatar, and co-presence, which merely requires awareness of another being within VR (Skarbez et al., 2018). Social presence has been identified as a positive predictor of communication outcomes, including enjoyment and social influence (C. S. Oh et al., 2018). Immersion has again been found to enhance social presence, although users' personality traits and the specific context of the VR application also play a significant role (C. S. Oh et al., 2018). In the present thesis, two presence-related metrics will be distinguished and applied as follows: *physical presence*, referring to the sense of being in the VR scene, and *social presence*, referring to the feeling of being with another entity in VR.

Presence is typically assessed using standardized questionnaires such as the Multimodal Presence Scale (MPS) or the I-Group Presence Questionnaire (Berkman & Çatak, 2021; Makransky et al., 2017). Each instrument captures distinct aspects of its underlying concept of presence. However, the various definitions and conceptualizations of presence questionnaires, which have been criticized for often being vague and imprecise, impede the interpretation and comparison of results (Grassini & Laumann, 2020). An item content analysis of 38 presence questionnaires concluded that there was a lack of consistent conceptualization, suitable items for operationalization, and alignment with the underlying concepts (Nannipieri, 2022). The author therefore suggests using single items or item pairs that closely align with the respective concept. The underlying motivation for measuring presence is often to assess other constructs, such as the user experience or to validate an interaction technique (Xiao et al., 2025). These aspects may be evaluated through more specific items (e.g., perceptual realism) or via emotional responses, engagement levels, or objective criteria. Increasingly, physiological correlates of presence are being investigated. These include measurements of brain activity via electroencephalography, brain imaging

techniques such as functional magnetic resonance imaging, and heart rate, skin conductance, or skin temperature (Grassini & Laumann, 2020). However, physiological findings related to presence are often inconsistent and lack replication, which may once again reflect the concept's vague and insufficient definition (Grassini & Laumann, 2020).

Recently, implicit measures of presence, or more precisely, the degree of deception by VR, have been proposed. For example, sitting on a virtual chair without verifying its physical presence or relying on memory from VR scenes to locate an object in the corresponding real-world scene has been interpreted as confusion of reality and VR (Wiesing et al., 2025). These behaviors are considered implicit indicators of presence. Another implicit approach to evaluating the quality of HTBAs in VR involves reorienting attention through auditory stimuli. The automatic movement of the head towards a sound source is known as the turn-to-reflex (Vorländer, 2008). To evaluate social virtual interactions, the impact of a virtual counterpart on participants' attitudes or the social-evaluative threat induced by the interaction can serve as an indirect measure of physical or social presence (Wiesing et al., 2025; Zimmer et al., 2019). In summary, various concepts and measurement approaches exist to evaluate audiovisual virtual interactions. Due to differing conceptualizations, interpreting findings remains challenging. However, certain factors influencing the quality of VR experiences have been investigated. The following section focuses on the impact of auditory factors.

2.4.3 The effect of auditory immersion on social presence

Given the importance of spatial hearing for socio-cognitive processes, this thesis investigates how auditory immersion affects virtual social interactions. As outlined in the beginning, psychological and VR research tends to emphasize visual aspects (Hutmacher, 2019). Immersive qualities are known to be positive predictors of presence. However, most research on social presence focuses on the influence of the visual appearance of the virtual interaction partners (C. S. Oh et al., 2018). In general, social realism has been shown to enhance social presence, for example, through realistic gaze behavior of agents (Schott et al., 2025) or the implementation of bidirectional gaze mechanisms (Andrist et al., 2017). Auditory immersion remains an understudied dimension of social presence (C. S. Oh et al., 2018). In an early study on telecommunication systems outside the context of VR, mediated communication, compared with face-to-face interaction, was evaluated among businessmen (Christie, 1974). While the implementation of a visual channel was considered most important, auditory immersion also played a significant role. When each communication partner was represented by a separate loudspeaker, reflecting spatialized sound, the medium was perceived as more

useful. This suggests a need for rapid speaker identification in complex virtual social settings as well. In a more recent study on videoconferencing, spatial audio was found to enhance social presence and facilitate turn-taking during a group survival task (Nowak et al., 2023). Improved task performance may result from increased interactivity, a stronger sense of shared space, and greater ease of understanding. These findings suggest once more that accurate speaker localization positively influences mediated social interactions. A study on video-game enjoyment manipulated visual and auditory immersion (Skalski & Whitbred, 2010). Subcomponents of presence, including engagement, perceived realism, and social realism, were investigated. Interestingly, high-definition displays did not outperform standard-definition displays. In contrast, the surround sound system was rated higher than two-channel audio across all variables except social realism (Skalski & Whitbred, 2010). It was suggested that social realism depends more on content than on formal factors such as audio quality. The ego-shooter's violent setting may have been too distant from everyday life to foster social realism. When it comes to the implementation of binaural auralizations specifically, the body of evidence becomes even more limited. Two studies addressed its effect on multi-party communication in audiovisual VR settings. First, the VR setting was compared to real-world interaction. Communication in VR is inferior in terms of social presence, with no observable effect of spatial audio on either social presence or user behavior (Immohr et al., 2023). In a collaborative VR task, binaural auralizations were preferred over diotic audio (Immohr, Rendle, Lammert, et al., 2024). However, this preference did not translate into measurable differences in social presence, plausibility, or communication behavior, which may have been due to the task-irrelevance of spatial audio in that context.

So far, only a limited number of studies have explored the impact of sound spatialization on mediated interactions involving emotional content. An audio-only study investigated the effects of mono, stereo, and binaural sound on social presence and emotion recognition when listening to a recording of an emotional multi-speaker conversation (Dicke et al., 2010). Binaural sound led to higher presence ratings than mono, especially when listeners were positioned among the speakers rather than separated from them, highlighting the role of sound-source localization. However, emotional understanding and involvement were unaffected by binaural audio, suggesting that externalization of sound may have a limited impact on the perception of emotional content in speech. In contrast, spatialized sound was found to enhance emotional reactions, presence, and emotional realism when participants listened to a piece of music designed to induce negative affect (Västfjäll, 2003). In summary, emotional involvement, social presence, and realism are key metrics in social virtual

interactions. There are several indications that plausible and externalized HTBAs may positively influence these dimensions, although further research is needed to substantiate this assumption.

3 Research objectives

The preceding sections provided an overview of research on socio-cognitive processing. Central concepts related to the mental processes involved in social understanding and interaction were introduced, with particular emphasis on the role of binaural hearing. This forms the basis for understanding how humans perceive and respond to virtual agents. Furthermore, the fundamentals of audio renderings were presented, including binaural auralizations and TTS synthesis. Their applications and evaluation frameworks were outlined, offering a basis for assessing these techniques in VR.

VR has emerged as a valuable tool for investigating social interactions in psychological research and applied contexts. Specific applications related to social anxiety, a prevalent mental health condition, were examined. The literature underscores the importance of enhancing social presence in VR. This is particularly relevant for scenarios involving social-evaluative threat, where the effectiveness of VR appears to lag behind that of non-social applications.

Metrics relevant to virtual social interactions were introduced, with a particular focus on auditory aspects. The influence of auditory immersion on virtual interactions was reviewed, revealing a gap in the literature concerning the effects of audio renderings in socially relevant and demanding VR scenarios. This thesis aims to systematically investigate the degree of immersion and realism that can be achieved through audio renderings in social VR, and to examine how these renderings affect socio-cognitive processing, presence, and affective responses. Figure 3 provides an overview of the main research objectives and hypothesized relationships between the topics and metrics of this dissertation.

The first primary objective is to identify appropriate behavioral and subjective evaluation methods for assessing the degree of correspondence between perceptions of real and rendered scenes. Realism and immersion are critical factors influencing the effectiveness of VR (Agrawal et al., 2020; Newman et al., 2022; C. S. Oh et al., 2018; Slater & Wilbur, 1997). In addition to ratings and questionnaires that measure presence, realism, and perceived audio quality, a more objective approach involves analyzing sound source localization accuracy in the rendered auditory scenes. As mentioned above, previous research has

highlighted the phenomenon of visual dominance over auditory input in distance perception and sound source localization (Gardner, 1968; Jack & Thurlow, 1973). This effect may be particularly prominent in VR, where visual distances are perceived as compressed (Buck et al., 2018; Kytö et al., 2015). Furthermore, VR enables the use of a range of measurement methods, such as eye tracking and the virtual positioning of objects (Grumiaux et al., 2022; Huisman et al., 2021; Schleicher et al., 2010). A suitable method for acoustic VR research remains to be determined. These methods may also interact with VR-specific characteristics, such as the use of HMDs (Majdak et al., 2010; Maruhn et al., 2019; Poirier-Quinot & Lawless, 2023). To address these questions, the study will examine how the display of virtual rooms with varying visual dimensions influences sound source localization and how this interacts with different suitable measurement techniques. As the methodology itself is of primary interest, the auditory scene will remain consistent, while only the visual environment will be manipulated. Employing physical sound sources ensures the highest possible level of auditory realism, thereby preventing the acoustic scene from being influenced by potential limitations or artifacts associated with rendering techniques.

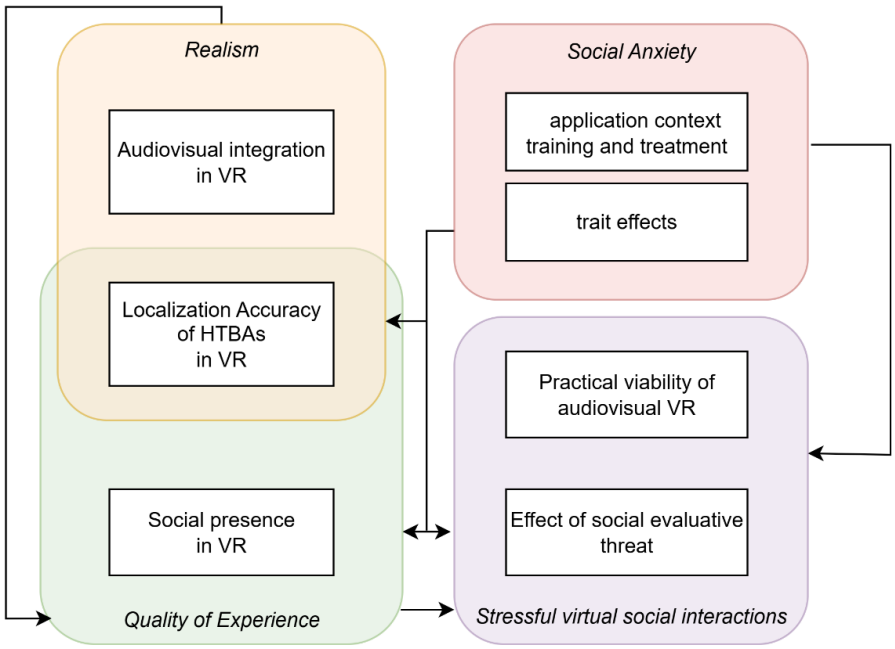

Figure 3: Conceptual map of core research objectives and hypothesized links between topics and metrics. VR = Virtual reality; HTBAs = head-tracked binaural auralizations. Audiovisual integration effects and localization accuracy of HTBAs in VR are investigated as indicators for realism, which is as a key contributor to quality of experience. Social presence is the most important metric of quality of experience in this project and is expected to be affected by realism, social anxiety, and social evaluative threat. Social anxiety is also expected to affect auditory localization and forms the broader context for this project, e.g., for treatments and training in VR.

A further central objective of this work is to select an audio rendering technique that combines high auditory realism with practical applicability in psychological VR research. This investigation builds upon previous work conducted by the acousticians and project partners in Oldenburg, who developed the technology for plausible and externalized HTBAs (Blau et al., 2021; Stärz et al., 2025; T. Wendt et al., 2014). Auditory immersion will be investigated by comparing various HTBAs with physical sound sources and an anchor audio condition. Key variables include source localization accuracy as an indicator of realism, social presence, and subjective assessment of audio quality. The findings will provide a foundation for selecting appropriate techniques in subsequent studies.

The next phase of this work involves implementing a scenario to investigate social virtual interactions under socially stressful conditions. To this end, the VR-TSST will be employed. This well-established paradigm reliably induces social-evaluative threat and enables the examination of stress responses on neuroendocrinological, psychophysiological, and subjective levels (Bahr et al., 2021; Goodman et al., 2017; Gunnar et al., 2021; Kirschbaum et al., 2008; Zimmer et al., 2019). The VR-TSST will be used to systematically assess the impact of different audio rendering methods on affective responses and perceived social presence, and to examine how these variables interact with participants' levels of social anxiety. Previous research has demonstrated that the TSST can elicit threat in healthy individuals, making it a suitable tool for studying socially stressful situations in non-clinical populations (Frisch et al., 2015; Grace et al., 2022). These insights may contribute to the development of VR interventions for social anxiety disorder (Grillon et al., 2006).

As a first step, the study will compare TTS synthesis with recordings of natural human speech. TTS has been shown to achieve a high level of realism and, in some cases, is indistinguishable from human voices (Khanam et al., 2022; Wenger et al., 2021). However, it remains to be demonstrated whether the use of TTS in scenarios designed to elicit social-evaluative threat yields stress-response and social-presence levels comparable to those achieved with human voice recordings. The integration of TTS into psychological VR research offers promising advantages, particularly in terms of standardization (Pan & Hamilton, 2018). Moreover, it may yield valuable insights into socio-cognitive processing within virtual environments.

The final goal of this thesis is to provide insights into the influence of plausible and externalized perceived HTBAs on affective responses in socially stressful situations, and to identify contributing factors such as social presence and attention. To achieve this, varying degrees of audio realism are examined in relation to their effects on these variables within VR

scenarios designed to induce either high or low levels of social-evaluative threat. Social anxiety will be analyzed as a predictor, and differential effects are expected.

Overall, this dissertation project aims to investigate the potential of advanced technology to increase auditory immersion in applications relevant to social and clinical psychology. While the project also aims to gain fundamental knowledge about the interaction between virtual acoustic and visual environments, the broader context was defined by social interaction. Emphasis was placed on the implications for the assessment and treatment of social anxiety.

The following research questions were addressed:

1. Does the VR-specific visual scene influence auditory perception and presence, and what are appropriate measures for investigating socio-cognitive processing in audiovisual VR?
2. How closely do advanced HTBAs approximate physical sound sources in terms of localizability, presence, and perceived audio quality?
3. Are stressful virtual situations using TTS stimuli able to evoke effects on stress, presence, and anxiety that are comparable to those elicited by human voice recordings?
4. Do highly plausible HTBAs lead to increased levels of presence, stress, and attention compared to non-spatial audio, and is this effect more pronounced in low- vs. high-stress virtual scenarios?

4 Study 1: Impact of Visual Virtual Scene and Localization Task on Auditory Distance Perception in Virtual Reality

Sarah Roßkopf, Andreas Mühlberger, Felix Stärz, Matthias Blau, Steven van de Par, and Leon O. H. Kroczek

The following article was accepted on 13 January 2025 and published online on 14 March 2025 in IEEE Transactions on Visualization and Computer Graphics following peer review. The official citation that should be used in referencing this material is: Roßkopf, S., Mühlberger, A., Stärz, F., Blau, M., Van de Par, S., & Kroczek, L. O. H. (2025). Impact of Visual Virtual Scene and Localization Task on Auditory Distance Perception in Virtual Reality. *IEEE Transactions on Visualization and Computers Graphics. 5*, 31. 10.1109/TVCG.2025.3549855.

4.1 Abstract

Investigating auditory perception and cognition in realistic, controlled environments is made possible by virtual reality (VR). However, when visual information is presented, sound localization results from multimodal integration. Additionally, using head-mounted displays leads to a distortion of visual egocentric distances. With two different paradigms, we investigated the extent to which different visual scenes influence auditory distance perception, and secondary presence and realism. To be more precise, different room models were displayed via HMD while participants had to localize sounds emanating from real loudspeakers. In the first paradigm, we manipulated whether a room was congruent or incongruent to the physical room. In a second paradigm, we manipulated room visibility - displaying either an audiovisual congruent room or a scene containing almost no spatial information- and localization task. Participants indicated distances either by placing a virtual loudspeaker, walking, or verbal report. While audiovisual room incongruence had a detrimental effect on distance perception, no main effect of room visibility was found but an interaction with the task. Overestimation of distances was higher using the placement task in

the non-spatial scene. The results suggest an effect of visual scene on auditory perception in VR implying a need for consideration e.g. in virtual acoustics research.

4.2 Introduction

Creating immersive virtual scenes with a high level of presence is crucial in many different research areas. In a recent systematic review on multisensory VR (Melo et al., 2020), it was summarized that if being perceived coherently and not in conflict with the virtual or real world, multisensory stimuli improve the quality of VR experience. However, it may be challenging to create such consistent virtual environments because of the specific visuospatial perception in VR. Compared to the real-world equivalent, a virtual scene appears minimized to users. When a head-mounted display (HMD) is used as VR device, egocentric visual distances are compressed to on average 74-82 % of the modelled size (Kelly, 2023; Renner et al., 2013). For visual depth perception, many different features are used, such as accommodation, convergence, or binocular disparity, which seem to be altered when looking within an HMD (Creem-Regehr et al., 2023). From the technical side, lower weight, a higher field of view, and a higher resolution of the HMD were found to reduce the visual distance compression effect (Kelly, 2023). But also body-based experience with the virtual scene for example provided by a self-avatar can help to calibrate the visual virtual distances (Creem-Regehr et al., 2023).

Over the last thirty years, evidence was gained that the implementation of acoustic stimuli increases presence (Agrawal et al., 2020; Hendrix & Barfield, 1995). Recent advances in virtual acoustics succeeded in synthesizing spatialized three dimensional sound that is highly realistic and natural (Blau et al., 2021; Brandenburg et al., 2020; Neidhardt et al., 2022). The physical aspects of a sound field in a to-be-synthesized room are reproduced with varying accuracy. Hearing impressions can be synthesized that are indistinguishable from real sound sources, indicating a high level of immersion (Stärz et al., 2022). Virtual acoustics have a broad field of application, e.g. helping to evaluate the room acoustic of concert halls or office cubes even before they were instructed. They have the potential to enhance the training of professional musicians or children with auditory processing and perception disorder, and virtual exposure therapy in the context of mental disorders (Neidhardt et al., 2024). Also the gaming industry and metaverse researchers are interested in implementing spatial audio, knowing that consistency between sound information and their spatialization increases presence in an auditory virtual environment (Larsson et al., 2008).

However, audiovisual room incongruence was shown to decrease quality features in virtual acoustic research (Werner et al., 2016). To be more precise, when the audible room impression conflicted with the visible room impression, the auralizations lost conviction. This is in line with the above-mentioned findings that multisensory stimuli enhance virtual experiences if there are no intervening discrepancies. Audiovisual scene incongruencies, were also found to lower the accuracy of sound source distance estimation (Gil-Carvajal et al., 2016). Even higher rates of front-back confusion were found. In the here described experiments, virtual sounds were used whereas the visible scene was a physical room, meaning that no visual virtual reality was displayed. Evidence is scarce on audiovisual integration when both acoustic and visual scene is virtual.

When investigating multimodal integration in virtual environments, the well-documented visual dominance in audio-visual sound source localization must be considered. The so-called visual capture effect describes the strong influence of vision on auditory perception. An acoustically presented source is localized towards a plausible visual target, as long as the spatial disparity does not exceed about 30° (Jack & Thurlow, 1973). In an augmented reality experiment, even greater spatial disparities were found to be tolerable, indicating higher visual capture effects in VR (Kytö et al., 2015). Also the perceived sound source distance is influenced by visual cues (Zahorik et al., 2005). Audiovisual stimuli that are presented together are integrated into one percept. In a laboratory study, the visual capture effect was able to drastically reduce perceived sound distances, e.g. from a real distance of 9m to a perceived distance of 1m (Gardner, 1968). On the other hand, audiovisual stimuli were also found to increase the accuracy of the comparatively imprecise auditory localization (compared to visual localization, for a review: Zahorik et al., 2005).

Notably, also manipulating acoustic features influenced visual distance perception (Huang et al., 2021). Higher reverberation times are associated with farther distances. Using this effect, the perceived depth of a virtual environment using an HMD could be increased. To add, by simultaneously presenting a spatially incongruent far-distanced sound, the estimated visual distance of an object could be increased (D. Liu & Rau, 2020). To sum up, auditory cues could affect visual depth perception in VR.

Few studies investigated sound distance perception in VR. Wearing an HMD itself was found to impair localization accuracy of pink noise played from real loudspeakers (Ahrens et al., 2019) and virtual sound sources (Ambisonics, Huisman et al., 2021). Visual information about possible source locations helped to compensate for these detrimental effects. Besides, positive effects of training (Steadman et al., 2019) and vibrotactile feedback (Mirzaei et al.,

2021) on sound source localization in VR were found. Yet, evidence on effects of the visual scene on sound distance and therefore depth perception in VR is scarce.

The above-described findings on audiovisual integration and the visual distance compression in VR indicate that auditory distance perception may be altered when using an HMD. The "minimized" visual scene may influence perceived distances of real sound sources (e.g. of loudspeakers) in the physical room because visual information was found to calibrate the auditory space (Kolarik et al., 2013). Indeed, in two VR studies using an HMD, overestimation of auditory distances of loudspeakers were found (Kroczek et al., 2024; Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2024). Participants indicated the perceived sound position on average 1.20 m and respectively 1.40 m further away than the real loudspeaker was (highest physical distance was 4.80m). When gaze behavior was used to measure perceived sound source positions, lower overestimates were found, indicating an influence of the measurement method (Roßkopf et al., 2024). A study using non-immersive visual virtual environment, to be more precise, a computer monitor displaying loudspeakers at different distances found no effect of this visual distance manipulation on distance perception of virtual sound sources, but a lowering of perceptual fluctuation (Song & Ma, 2022). The extent to which the display of an immersive scene in HMDs influences sound distance perception has yet to be clarified.

Our research is motivated by three main questions:

1. Is sound distance perception influenced by the display of a congruent vs. incongruent immersive virtual scene?
2. To what extent does the display of a visually compressed scene, as shown for HMDs, influence sound distance perception?
3. How does the interplay of audiovisual scene and task effect accuracy of localization and subjective experience in VR.

We investigate these questions with two sequentially conducted experiments. In the first experiment, the focus is on the investigation of audiovisual room incongruence. Findings indicated a substantial sound distance overestimation, especially in the incongruence condition, highlighting the influence of visual information. This motivated a second experiment, where a more general effect of visual scene was investigated by manipulating availability of visual information. Additionally, the second experiment was designed to investigate the influence of the measurement method itself. The way in which estimated distances are indicated may also interfere with the special features of perception in VR. Another important goal is to investigate whether subjective experiences in VR in terms of

quality of experience (e.g. presence) and adverse effects are influenced by audiovisual room incongruences, room visibility and tasks. All hypotheses and analyses were preregistered on osf.io (osf.io/nj8k2, osf.io/vjcer), where also all raw data and analyses scripts are publicly accessible.

4.3 Experiment 1

The goal of experiment 1 was to investigate the effects of displaying an audiovisual congruent vs. incongruent room on auditory distance perception and subjective experience in VR. To manipulate room (in)congruency, we created two differently sized visual models of the room in which the experiment took place. Sound stimuli were played from three loudspeakers in the physical room yielding the advantage of a "real" acoustic scene whereas the visual scene can be manipulated independently. Based on previous findings of a detrimental effect of a divergent physical room when reproducing an acoustically virtualized scene (Werner et al., 2016), we derived the following hypotheses:

H1: Individuals will be significantly worse in sound source localization in an incongruent in comparison to in a congruent room in VR.

H2: Presence and realism are enhanced in a congruent room.

4.3.1 Methods

4.3.1.1 Sample

Our sample (N = 30) consisted of 27 female and 3 male participants aged between 18 and 32 years (M = 22.9, SD = 3.55). The majority of participants were students (80%) and all had a high school diploma.

The sample size was based on an a-priori power analysis using G*Power 3.1 (Faul et al., 2009) indicating a sample size of 27 to be sufficient to detect a effect, d = 0.5, with *alpha* set at .05 and 1 – *beta* = .80 for a paired sample t-test. This effect size (medium) was chosen according to effects of audiovisual room divergence on the quality feature "externalization" for which odds ratios between 1.8 and 5.1 (for significant results) were found. We increased the sample size to 30 to compensate for dropouts and exclusions.

Healthy adult individuals with self-reported unimpaired hearing, normal or corrected to normal vision, and German speaking experience of minimum 5 years were included in the study. All participants gave written informed consent. The study was in line with the Declaration of Helsinki and approved by the ethical review committee of the University of Regensburg (Ref-No.: 20-1804-101).

4.3.1.2 Materials

Audiovisual room manipulation

Audiovisual room congruence was manipulated using a within-subjects design. Half of the participants started in the congruent, the other half in the incongruent room.

The experiment took place in a seminar room of the University of Regensburg (room size: 6.8m x 4.8m x 3.3m). The room consists of four concrete walls, an acoustically optimized ceiling, and a carpet on concrete floor (see00). Reverberation time in the room was measured as 0.80 s (T20). For the visual virtual room, we created two photorealistic models of the seminar room with the Unreal Game Engine (v 4.26, Epic Inc.) and Blender (v 2.79). The first model corresponded to the real dimensions of the seminar room while the second model represented an enlarged room version (11m x 7m x 4.1m), so that the divergent room had almost three times the volume than the congruent room (see 0). The initial viewpoint towards the screen was the same for both room models, but the divergent one was enlarged to the right, back and the ceiling. In addition, we took care that the distance between participants, loudspeakers, and the front wall was equal in both room models to rule out the possible confounder that the front wall would restrict the responses more in the congruent room.

Technical setup

During the experiment, sounds were played from three active monitor speakers (GENELEC 8030C, Genelec, Finland) via an audio interface (RME Fireface UCX, RME Audio, Germany) controlled by a custom Python script. Three loudspeakers were positioned in a row frontal to the participants (0° azimuth) at a distance of about 1.5 m, 2.5 m and 3.5 m (see Figure 4). All speakers were placed on tripods with the height of the acoustic center at 1.20 m, therefore substantial lower than the height of the ears of the standing participants. This enabled perception of direct sound from all three speakers and prevented occlusion.

Figure 4: A: Setup of loudspeakers in the experimental room. B: Placement task from participants' view. C: Congruent audiovisual room. D: Incongruent audiovisual room.

As sound stimuli recordings of human speech were used in line with the prior studies finding sound distance overestimation in VR (Kroczek et al., 2024; Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2024). Two different sets of stimuli were used, first, semantically neutral nouns produced by male speakers either with angry or neutral voice (Quadflieg et al.,

2008), second, female voice stimuli gained from a German language learning program (Funk et al., 2013). All stimuli were root mean square (RMS)-normalized to -24dB (full scale) to ensure an appropriate loudness for speech. A total of 50 different stimuli were used repeatedly for each loudspeaker position and each room condition, resulting in 300 trials.

The virtual environment was presented via an HMD (Vive Pro Eye, HTC) and using the Unreal Game Engine (v 4.27, Epic Inc.). A male and a female virtual agent were created using MakeHuman (v 1.2) and Blender (v 2.79) for the placing paradigm (see Figure 4). The use of agents has been considered as coherent to the speech stimuli. In a previous study, no influence of object type (loudspeaker vs. agent) was found (Kroczek et al., 2024). The virtual agents were animated sitting on a chair with slight breathing movements. In a seated position, the height of agents' mouths corresponded to the acoustic centre of the loudspeaker.

4.3.1.3 Measurements and Data Processing

To measure sound source localization, participants had to place a virtual agent at the position in the virtual room, where they assumed the sound source. Via the HTC vive motion controller, the agents could be placed continuously on the floor of the virtual room (see Figure 4). More precisely, the x- and y- axis, but not the z-axis coordinates of the position of the virtual agents could be altered. The agents only appeared after the sound playback was completed and the participants actively started the placement. Therefore, the placement of the agents should not have led to a visual capture effect. During the experiment, the position of participants, virtual agents (position after placement), and the real loudspeaker positions were tracked. Deviations between estimated and real sound source position (euclidian distance vector in cm) were calculated as the primary outcome variable for sound localization error. Localization error was averaged per room condition (150 trials each), loudspeaker position, and participant for statistical analyses. Additionally to this primary outcome variable, we computed following supplementary indicators for sound source localization accuracy (non-preregistered analyses): We analyzed the amount of trials (in %), in which the agent was placed outside the walls of the visual room (Trials outside Room) and classified them as invalid. We further classified data as invalid if front-back confusions occurred. Further we computed an indicator for distance deviance by subtracting the physical distance from the estimated distance between the participant and sound source. Consequently, a distance deviance of 0 equals perfect distance estimation, whereas a positive value indicates an overestimation of distance. As an indicator for azimuthal localization accuracy, the angle deviance in degrees was computed using the dot product of the vector between participant and real sound source and the vector between participant and estimated sound source. As indicator

for systematic angular deviance on the horizontal plane (to left or right side), the azimuth error was calculated. For the absolute distance error (in cm) as an indicator of the overall accuracy of distance estimation (regardless of whether the distance was overestimated or underestimated) the mean of absolute distance deviance was calculated (see Table 1). To measure subjective experience of virtual reality and audio scene, two 9-point Likert scaled ratings were implemented within the scene. The first item concerned the experience of presence (translated from German “In the virtual environment, I had a sense of really being in the seminar room.”). The second item concerned the perceived realism (translated from German “The current hearing impression was similar to the impression in a real seminar room.”). Rating data were averaged per room condition and participant for statistical analyses.

4.3.1.4 Procedure

After given written informed consent and filling in demographic questionnaires, participants were introduced to the controller and the HMD. To prevent a view on loudspeaker positions, participants entered the seminar room “blindfolded” by wearing the HMD and were guided to the starting position by the experimenter and virtual footprints. Then, the respective visual room model determined via randomization was displayed. At the beginning of the VR experiment, participants were asked to read the instructions to ensure unimpaired sight. All instructions were presented on a whiteboard surface on the opposite wall. As first task, participants were instructed to look around the virtual room for one minute to ensure that they were aware of the dimension of the actual room. Then, a two-part practice run began. In the first part of the exercise, participants learned to handle the controller and to place the virtual agents at requested positions. Practice trials were repeated as long as the deviation between placed agent and target exceeded 20 cm. In the second part, the subjects learned to operate the rating scales implemented within the VR scenario.

At the beginning of each experimental trial, the head orientation was examined. If the head orientation exceeded an angle of 10° (roll, pitch, or yaw), participants were instructed by a red coloured text to look straight forward. When accomplished, the sound stimulus was played back. Then, participants conducted the localization task. By pressing the trigger button on the motion controller, a virtual agent appeared at the targeted position and could be moved anywhere on the floor. When moved on the x axis, the agent was rotated, so that the mouth was always directed towards the participant. There was no time limit. After confirming the position with another button press, the agent disappeared, and the next trial began. Ratings were obtained in 60 (out of the 300) trials in a pseudorandom selection (20% catch trials).

After the first half of the experiment (150 trials), participants were escorted to an outer room and the HMD was removed for a brief, self-paced break. Then, the HMD was remount and participants re-entered the seminar room, reached the same starting position and the other visual room model was displayed for the second half of the experiment. Again, 150 trials were completed, afterwards, participants filled in post-questionnaires in the outer room.

4.3.1.5 Statistical Analyses

Preregistered analyses were conducted to test detrimental effects of audiovisual room incongruence. Two-tailed paired t-tests were conducted to compare differences in dependence of room models each for the variable localization error and the rating variables presence and realism. We furthermore analyzed whether a general bias towards overestimation of sound distances can be found. Therefore, one-samples t-tests comparing the difference means to zero will be computed. Regarding a possible overestimation of distance, repeated measures analysis of variance (ANOVA) were analyzed to find biases between the three real distances. An alpha of $p < .05$ was determined. Normality assumption was tested using Shapiro Wilk tests. No Deviation from normality was observed for any of hypothesis tests' outcome variables, neither for localization accuracy, $W = 0.982$, $p = .878$, nor for distance deviance, $W = 0.976$, $p = .703$, nor for presence, $W = 0.938$, $p = .082$, or realism rating, $W = 0.938$, $p = .082$. If necessary, the results were corrected for sphericity using the Greenhouse-Geisser correction.

4.3.2 Results

4.3.2.1 Sound Source Localization Accuracy

Figure 5 provides an overview of real and estimated source positions from the placement task for each congruency condition. It can be seen that front-back confusions occurred in both conditions in less than 1% of trials. Overestimation of distances occurred frequently. In Table 1, several sound source localization parameters are shown, and whether they differed significantly between the audiovisual room conditions (directed paired *t*-tests). Per participant and room condition 150 trials were analyzed.

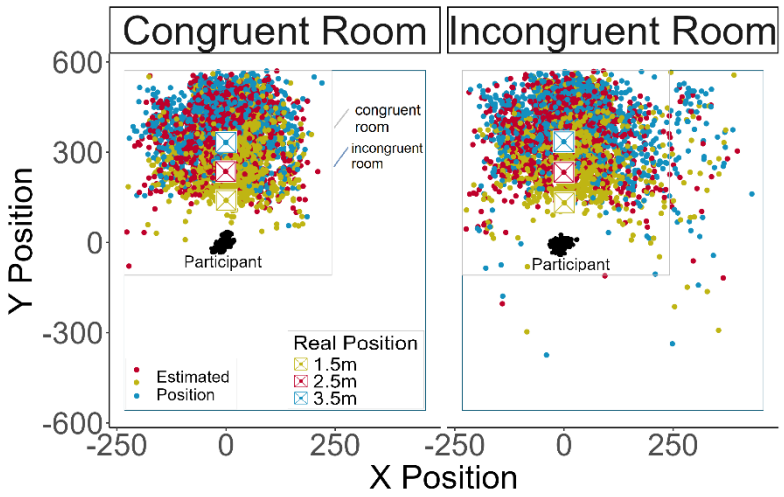

Figure 5: Tracked positions in the virtual room. Note that participants' forward orientation is towards the positive y-coordinate.

A paired directed t-test revealed a detrimental effect of audiovisual room incongruence on localization error $t(29) = 2.63$, $p = .007$, $d = 0.48$. The distance vector between the real and the estimated position of the sound source was higher in the incongruent audiovisual room condition in comparison to the congruent audiovisual room condition (see Table 1).

Table 1: Localization Accuracy Parameters for Room Condition

Variable	congruent audiovisual room		incongruent audiovisual room			
	M	SD	M	SD	p	d
Trials outside Room (in %)	0.4	1.0	0.7	1.1	.183	0.17
Front-Back Confusions (in %)	0.2	0.4	0.7	1.5	.016	0.41
Invalid Trials (in %)	0.5	1.0	1.1	1.8	.033	0.35
Localization Error (in cm) [a]	152.3	81.4	164.6	86.0	.007	0.48
Angle Deviance (in °, 0-90) [a]	15.9	8.6	17.4	10.1	.001	0.59
Azimuth Error (in °,-180-180)	1.2	13.4	3.1	16.4	.018	0.15
Distance Deviance (in cm) [a]	127.2	99.7	132.0	104.6	.215	0.15
Abs. Distance Error (in cm) [a]	135.2	99.7	142.0	104.6	.102	0.24

[a] For these analyses invalid trials were excluded.
Bold values indicate statistical differences of directed *t*-tests ($p < .05$)

To test whether there is a general bias towards the overestimation of sound source distances, we conducted a one-sample t-test comparing the mean distance deviance (in cm) to zero. The general deviance between real and estimated sound distance (over all experimental conditions and source positions) was 129.6 cm (SD = 102.2). This value is significantly different from 0, $t(29) = 13.75$, $p < .001$, indicating that sound distances were overestimated in general.

A repeated measures ANOVA revealed that there is a significant difference in the overestimation bias between the three real distances, $F(1.1, 31.8) = 66.6$, $p < .001$, $\eta_p^2 = .70$. Furthermore, a significant interaction effect between loudspeaker position and audiovisual room divergence was found, $F(2, 58) = 7.98$, $p < .001$, $\eta_p^2 = 0.22$. Post hoc t-tests revealed that overestimation of distances was significantly stronger for incongruent compared to congruent

room condition for the nearest sound source, $t(29) = 2.35$, $p = .026$, $d = 0.43$, while no differences were found for the middle and farthest sound source. Note that overestimation was highest at the nearest compared to middle and farthest sound source in both room conditions ($M_{near} = 172$, $SD_{near} = 103.5$; $M_{middle} = 139$, $SD_{middle} = 92$, $M_{far} = 75.8$, $SD_{far} = 80$).

While no interaction effect of loudspeaker position and audiovisual divergence on angle deviance was found, one was found for azimuthal error (as indicator for a systematic deviance to either the left or right side), $F(1.62, 47.10) = 4.24$, $p = .027$, $\eta_p^2 = 0.13$. Here, only for the farthest source position, a significant difference between rooms was found, $t(29) = 2.92$, $p = .007$, $d = 0.53$, indicating a stronger tendency of positioning to the right.

To test possible practice effects, linear mixed effect models with the trial number (indicating the number of previous trials) as fixed factors and a random intercept for the participants were be conducted. The trial number had a significant effect on localization accuracy, $F(1,8969) = 174.92$, $p < .001$. Per previously absolved trial, the overall localization error was reduced on average by 0.13 cm.

4.3.2.2 Subjective Experience in VR

Figure 6 provides an overview of ratings of presence and spatial audio quality in the different room conditions. Presence was not rated significantly higher in the audiovisual congruent room than in the audiovisual incongruent room condition, $d = 0.1$, $p > .05$.

The ratings of spatial audio impression did also not differ significantly between the audiovisual congruent room and the audiovisual incongruent room condition, $d = 0.05$, $p > .05$.

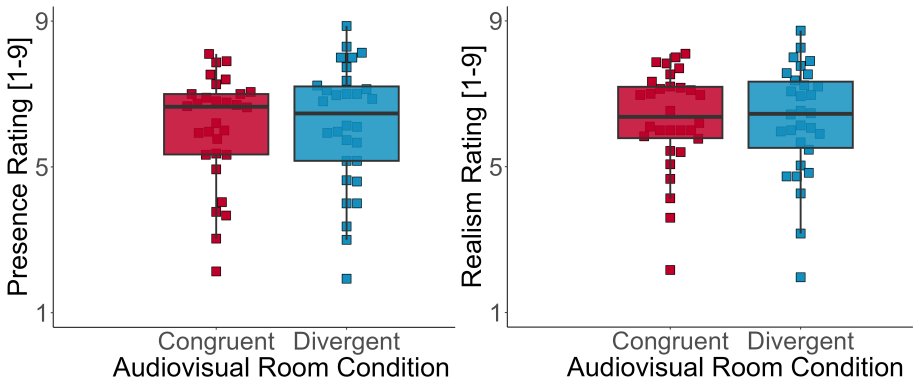

Figure 6: Ratings of Presence (left) and Realism (right).

4.3.3 Discussion of Experiment 1

A detrimental effect of audiovisual room divergence on sound source localization was found as hypothesized. Contrarily, no effect on presence or realism was found. Within the VR

scenario, participants placed a virtual agent at the perceived sound source position. The sounds emanated from real loudspeakers in the room. When the displayed virtual room had the same dimensions as the real room, higher localization accuracy was found in comparison to when the displayed virtual room was divergent. Especially regarding the azimuthal localization, lower accuracy was found for the audiovisual incongruent room condition.

Distances of loudspeakers were generally overestimated, particularly more in the incongruent room. Overall, we found a mean overestimation of distances of about 1.30 m (real mean distance of 2.5 m). Contrasting the current findings, in psychoacoustic studies conducted in real environments, a general underestimation of sound distances is described (for a review: Zahorik et al., 2005). One possible explanation for the overestimation of real sounds' distances in the current study may be the multimodal integration of the auditory scene and the visual scene that is compressed because it is presented using an HMD (Renner et al., 2013). The well described ventriloquist effect - meaning that acoustic source estimations are shifted towards the visual object may have occurred (Gardner, 1968; Jack & Thurlow, 1973), or the visual capture effect - meaning that spatially separate audiovisual stimuli are fused to one location (L. Chen & Vroomen, 2013). The proximity image effect describes particularly strong visual capture. Regardless of the physical distance only the first loudspeaker in a straight line is perceived as a sound source even when sound pressure levels largely vary (Gardner, 1968). Since the visual scene of experiment 1 offered no plausible (visible) source positions in the middle of the room from which the sounds emanated, which is an artificial setting, visual capture effects exceeding the mainly found space (about +/-15°) could have occurred. Potentially, the localization was shifted towards the walls or objects positioned near the walls (tables, chairs). Alternatively, the method which was used to indicate perceived sound distances may have contributed to the current finding of strong overestimation. In a previous study, higher sound distance overestimation was found when the placement task was used in comparison to an eye-tracking task (Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2024). It must also be said, that the human auditory system is comparably poor at sound distance perception (compared to auditory azimuthal localization and compared to visual distance estimation, Durlach et al., 1993). Therefore, sound distance perception in virtual reality might be especially susceptible to the peculiarities of the visual scenes used.

Interestingly, participants improved their localization accuracy in the course of the experiment while no feedback was given. This may be an indicator for an adaptation to the virtual environment. Potentially, besides visual and haptic feedback, also acoustic cues may help to calibrate the perceived space in VR (Altenhoff et al., 2012).

To sum up, this experiment found a detrimental effect of audiovisual incongruence on the localization of real sound sources in VR (higher rates of front-back confusions, lower distance estimation accuracy) confirming an influence of visual virtual scene on auditory distance perception. No effect of audiovisual incongruence on presence and subjective realism was found. Last, a general overestimation of sound source distances was found. The extent to which the auditory distance perception is influenced by visual compression using HMDs is yet unknown. Therefore, a second experiment was conducted to shed a light on the influence of room visibility on auditory distance perception in VR.

4.4 Experiment 2

In experiment 1, not only an influence of audiovisual room convergence on localization accuracy was found, but also a general overestimation of auditory distances (in both room conditions). This overestimation could originate from VR specific audiovisual integration effects where the visual scene dominates auditory perception, similar to e.g. the ventriloquist effect (Jack & Thurlow, 1973) or the proximity-image effect (Mershon et al., 1980). As outlined above, visual spaces and objects are perceived as about a fourth smaller than the corresponding physical ones using an HMD (Renner et al., 2013). Alternatively, the found overestimation may result from the used measurement method. When a perceived auditory distance is indicated within a visual scene, e.g. by placing a virtual agent at the speakers' position, the distance must be transformed. Since visual objects are perceived as closer in VR, participants must then place e.g. the agent at a more distant position to reach the visual distance, which equals the acoustically perceived distance.

The used measurement method affects distance estimations, both in conventional laboratory studies and in VR studies. A higher overestimation of auditory distances was found for a placement task in comparison to an eye-tracking task in a VR study (Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2023). In a conventional laboratory study, higher accuracy was found when participants indicated the perceived position with a visual marker compared to verbal reports (Etchemendy et al., 2018). Concerning visual egocentric distance estimations, verbal answers were found to be more accurate than visually guided (walking up to a covert object, Maruhn et al., 2019). Hence, the method with which the perceived distance is measured has to be taken into account as well as a possible interplay with the peculiarities of visual perception in VR using HMDs.

Experiment 2 was designed to disentangle the impact of the visually compressed scene and the used measurement method on auditory distance perception in VR. Therefore, a fully

visible and an almost blind room version, and three different localization tasks are investigated. If the overestimation of distances found in experiment 1 is mainly driven by the HMDs' visual compression effect, less distortion may be found in the blind condition assuming a higher focus on auditory cues. Furthermore, an interplay of visual compression effect and measurement method especially for placement task is assumed. Based on the findings that visual egocentric distances are compressed in VR (Buck et al., 2021; Renner et al., 2013) we derived following hypotheses:

H1: Auditory distances are overestimated when a fully visible virtual room is displayed via HMD in comparison to a virtual room with very reduced spatial cues (blind condition).

H2: We expect largest sound distance overestimation for the placement task requiring a high degree of audiovisual interaction (Etchemendy et al., 2018).

H3: Task difficulty is rated highest for verbal estimation (being the most indirect method, Etchemendy et al., 2018), and lowest for the placement task whereas the rating will not depend on room visibility.

H4: Adverse effects will be highest in the walking condition in the blind room.

4.4.1 Methods

4.4.1.1 Sample

The sample (N = 30) consisted of 28 female and of 2 male participants aged between 18 and 30 years (M = 21.3, SD = 3.4). The majority of participants were students (90%). Concerning hearing experience, 73.3 % of the participants reported having learned playing a musical instrument and having played for on average 8.5 years.

According to a power analysis with G*Power 3.1 (Faul et al., 2009), a sample size of at least N = 27 is required for the planned statistical calculations of an ANOVA with repeated measures with alpha = .05 and 1 – beta = .80 in order to find an effect of size d = 0.5. This size is based on Experiment 1, where a mean effect size of d = 0.48 was found. The planned sample size was increased to 30 to compensate for possible dropouts or outliers. All participants gave written informed consent. The study was in line with the Declaration of Helsinki and approved by the local ethics committee (University of Regensburg, see Experiment 1).

4.4.1.2 Materials and Design

The same software, loudspeakers, and physical room, as described for Experiment 1 were used. The three loudspeakers were positioned in a row in front and about 70 cm left of the

participants. This left-sided position was chosen due to the walking task. The loudspeakers were at distances of 1.68 m, 2.65 m, and 3.82 m to the participants. Only female speech stimuli, single words and up to five-word sentences gained from German language learning program (Funk et al., 2013), were used. We used the HMD (Vive Pro, HTC) in a wireless configuration to enable unhindered movement around the room. The virtual room models were spatially aligned to the physical room, via an in-house-developed two-point calibration technique using custom-made mounts for the HTC motion controller (Kroczek et al., 2023).

The study had a within-subjects design with two factors: Room visibility with two levels (visible room vs. blind condition), and measurement method with three levels (placement task vs. walking task vs. verbal report). The experimental conditions were presented block wise, resulting in six different blocks (visible-placement, visible-walking, visible-verbal, blind-placement, blind-walking, blind-verbal). Each block consisted of 30 trials (ten different items respectively per three different source positions). The order in which participants absolved the experimental blocks was counterbalanced via a quasi-randomization list created with R (Development Core, 2019). While half of participants started with the visible room condition, the other half started with the blind condition. The order of position x stimulus combinations within each block and the order in which the measurement methods were absolved was randomized (always the same within a participant for each room condition).

4.4.1.3 Visual Room Manipulation

Two different VR room models were used. For the “visible room” condition, we used the above-described congruent room model (see room manipulation Experiment 1). For the “blind” condition we reduced spatial cues of this room to a minimum. Therefore, the walls, the ceiling, and the floor were black-colored, and all light sources were eliminated. Due to safety reasons, enlightened lines on the floor were provided (both room models) to indicate the possible range of motion for the walking task (see Figure 7). When participants crossed the boundary lines, red warnings were triggered instructing the participants to step back into the allowed range. Besides the lines, the only visible cues were the instruction screens (white writing on black ground) and the slightly lit loudspeakers in the placement task.

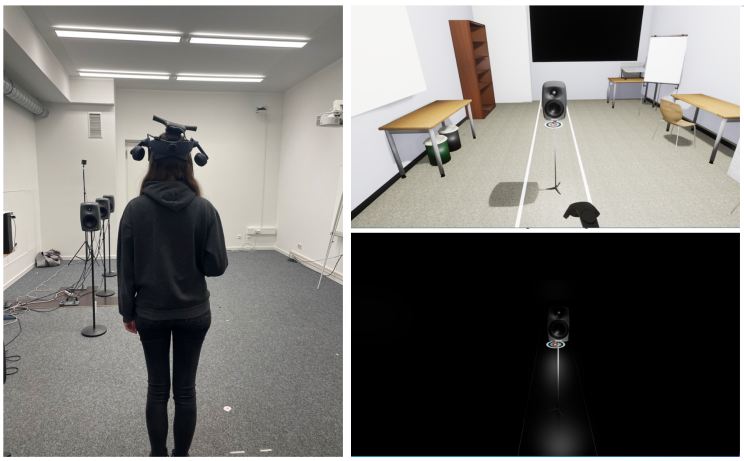

Figure 7: On the left: participant absolving the placement task in the experimental room. On the right: participants' perspective in VR from the placement task in the visible room (top) and the blind condition (bottom).

4.4.1.4 Measurement Method

Three different methods to measure perceived sound source distance were used. Participants were instructed to ignore the angle of the sound source and only indicate its distance. Immediately before each task, the measurement method was practiced using visual target objects. During the experiment, the position of participants, the virtual visual loudspeaker (position after placement), and the physical loudspeaker positions were tracked. In each task, the estimate was given after the sound had been played to the end. For analysis of the distance error, the physical distance of the loudspeaker to the participant during playback of audio (only y axis) was subtracted from the estimated distance.

Placement Task

In the placement task, participants were instructed to place a virtual visual loudspeaker at the perceived distance of a sound source via the HTC motion controller (see Figure 7). The loudspeaker only appeared after the sound playback was completed to prevent visual capture effects. The difference between the virtual loudspeaker and the participant was logged as estimated distance. For practicing the placement task, the virtual visual loudspeaker had to be placed on the visual target object.

Walking Task

In the walking task, participants are instructed to walk to the position at which the distance of the sound source was perceived and then confirm this position with a button press via the HTC motion controller. The difference from participants finally confirmed position and the starting position of the participant during audio playback was logged as estimated distance. In practice trials, participants had to walk up to the visual target object and then press a button.

Verbal Task

In the verbal report task, participants were instructed to speak aloud their estimation of sound source distance. The experimental instructor recorded and converted the participants' answers to cm where required. For practicing the verbal task, participants had to speak aloud at which (egocentric) distance they perceived the visual target object. This data was also used for indicator of visual compression (see Results).

4.4.1.5 Rating of Experience in VR

For measuring the quality of experience in and adverse effects of VR, continuous rating scales were implemented within the scene. Each scale consisted of a rating item (see Table 2), two verbal anchors each for the 0 (not at all) and the 100 (very much) poles, and a slider, which could be moved continuously via the HTC motion controller.

Table 2: Ratings

N°	Rating of	Question
1	Difficulty	How difficult did you find it to specify the distance in the previous task?
2	Certainty	How certain were you in your assessments?
3	Presence	How much did you just experience virtual reality as if you were really "there"?
4	Discomfort	How much general discomfort/discomfort did you feel during the previous task?
5	Vertigo	How much dizziness or loss of balance did you experience during the previous task?
6	Nausea	How much nausea or stomach discomfort did you experience during the previous task?

The rating item to measure presence based on spatial presence question from the multimodal presence scale (Makransky et al., 2017). The items to measure adverse effects were derived from the simulator sickness questionnaire (Kennedy et al., 1993).

4.4.1.6 Procedure

After given written informed consent and filling in demographic questionnaires, participants were introduced to the controller and the HMD was put on. Participants entered the seminar room "blindfolded" and were guided to the starting position by the experimenter and virtual footprints. Then, the VR presentation started with an open-spaced virtual area (open sky platform without any room information), where general instructions about the controller, ratings, and tasks are given and practiced. Then, according to the randomization list, the first experimental block was started. Each experimental block began with three practice trials and ended with the ratings. After a short break, the next experimental block started, and this procedure was repeated until all experimental conditions were absolved. At the end, participants were led out of the room "blindfolded" and the HMD was taken off. Finally, a follow-up questionnaire assessed possible hypotheses and interferences and physical presence via the corresponding subscales of the MPS (Makransky et al., 2017).

4.4.1.7 Statistical Analyses

The distance estimations were averaged over all trials for each participant and each experimental condition. During the examination of one participant, a technical error occurred, and an experimental block had to be interrupted and repeated. The data of the incomplete block was fully replaced with the repetition. During the walking task, by all appearances erroneous trials occurred, in which implausible starting positions were logged. Therefore, starting positions deviating more than three times the standard deviation from the mean were replaced with the mean starting position coordinates. Furthermore, trials in which front-back confusions occurred or trials in which the loudspeaker was placed outside the room, were excluded (0.002% across all participants).

As first step, repeated-measures ANOVAs were conducted. Significant interaction effects of visual room condition and measurement method were investigated using post-hoc *t*-tests. To analyze, whether there is a bias towards overestimation, one-samples t-tests comparing the distance difference means to zero were computed. When more than one *t*-test are conducted, Holm procedure (Holm, 1979) to correct for multiple comparisons was applied. Normality assumption was tested using Shapiro Wilk tests. No Deviation from normality was found for distance error (W = 0.951, p = .182), nor for the rating of task difficulty (W = 0.961, p = 0.332), presence (W = 0.953, p = .207), or uncertainty (W = 0.983, p = .893). Deviation from normality was only found for the rating data concerning adverse effects (discomfort, W = 0.900, p = .008; vertigo, W = 0.818, p < .001; nausea, W = 0.642, p < .001). As a consequence, nonparametric tests were used for analysis of adverse effects. Interaction effects were evaluated by comparing differences of visible and blind room condition between measurement conditions via Friedman rank sum tests followed by pairwise Wilcoxon signed-rank tests. If necessary, the results were corrected for sphericity using the Greenhouse-Geisser correction.

4.4.2 Results

4.4.2.1 Auditory Distance Perception

A repeated-measures ANOVA revealed a significant interaction effect of room condition and measurement condition, $F(1.47, 42.72) = 3.83$, p = .042, η_p^2 = 0.12. As can be seen in Figure 8 distance error in terms of overestimation is significantly higher when the placement task was used in comparison to walking and verbal report and that this effect is even increased in the blind room condition. The difference between the visible and the blind room condition is

higher in the placement task in comparison to the walking task (t[29] = 4.35, p < .001, d = 0.79) but not in comparison to the verbal task (t[29] = 1.11, p = .277, d = 0.20).

Overall, a main effect of measurement method was found, F(1.34, 38.96) = 6.70, p = .008, η_p^2 = 0.19. Post-hoc tests showed higher overestimation of distances during the placement task than during walking task, t[29] = 5.7, p < .001, d = 1.04, and than during the verbal task, t[29] = 2.84, p = .016, d = 0.52. As outlined before, this effect was modulated by the used room condition.

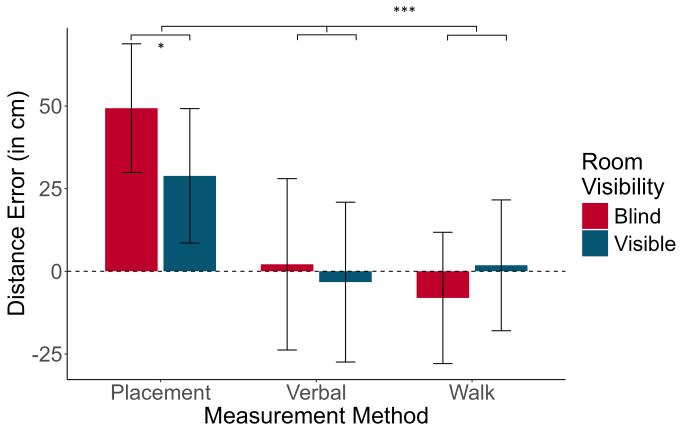

Figure 8: Mean difference between physical and estimated distance. Error bars indicate the standard error.

No general overestimation of source distances was found, t(29) = 0.87, p = .389, d = 0.10. Though, an effect of speaker position on distance overestimation was found, F(1.12, 32.48) = 66.06, p < .001, η_p^2 = 0.69. The distance of the two nearer loudspeakers (168cm, 265cm) were overestimated whereas the distance of the farthest loudspeaker (382cm) was underestimated (see Figure 8). This effect was neither modulated by room visibility nor by measurement method. The distance to the nearest loudspeaker was significantly more overestimated than the distance to the farthest (t[29] = 7.93, p < .001, d = 1.45). The distance to the middle loudspeaker was overestimated significantly higher than to the farthest, t(29) = 16, p < .001, d = 2.92, whereas between the near and the middle distance estimation, no difference was found, p > .05.

4.4.2.2 Subjective Experience in VR

After each experimental block, participants rated their experience during the previous task via six rating items (see Table 2). Figure 9 provides an overview of mean ratings per attribute, room visibility and measurement method. To begin with, a main effect of measurement method on difficulty of task was found, F(2, 58) = 27.96, p < .001, η_p^2 = 0.49. Confirming our hypotheses, difficulty of verbal estimation was rated significantly higher than difficulty of placement task, t(29) = 6.29, p < .001, d = 1.15, and of walking task, t(29) = 5.63, p < .001,

$d = 1.03$, whereas these last two measurement methods did not differ significantly ($d = 0.13$), $p > 05$. In contrast to our hypotheses, a main effect of room visibility on task difficulty was found, $F(1, 29) = 5.32$, $p = .028$, $\eta_p^2 = 0.16$. Averaged over all measurement conditions, difficulty was rated higher for the blind condition than for the visible room condition, $t(29) = 2.31$, $p = .028$.

The last hypothesis concerned adverse effects (lower row in Figure 9). We expected adverse effects to be more pronounced in the walking task and in the blind room condition. No interaction effect of room visibility and measurement method was found on discomfort and nausea, whereas an interaction effect was found on vertigo, $\chi 2\ (2) = 8.57$, $p = .014$. The difference between vertigo in the visible and the blind condition depended on the measurement method. Post hoc comparisons revealed that only for the blind room condition, higher vertigo was found after walking ($W = 183$, $p = .020$) and after placement task ($W = 281$, $p = .001$) in comparison to after the verbal task.

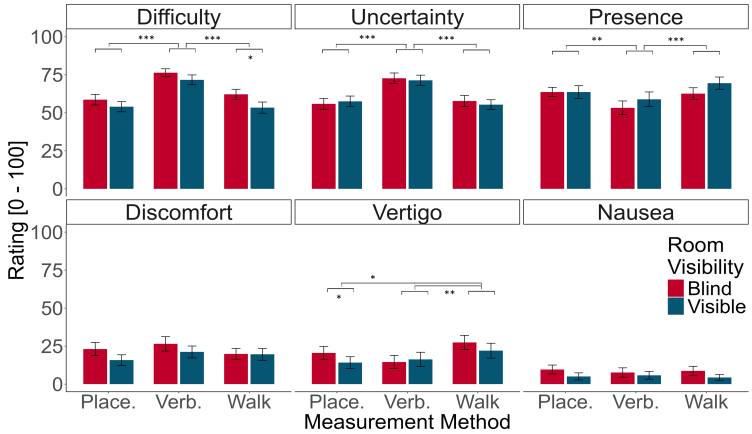

Figure 9: Mean ratings of subjective experience items. Error bars indicate the standard error.

4.4.2.3 Visual distance compression

We furthermore investigated whether visual compression occurred (Renner et al., 2013). Therefore, the answers from the practice trial of the verbal task were analyzed ($N = 20$) where participants had to indicate the estimated egocentric distance to visual targets at three different real distances (see Table 3). Overall, visual distances were underestimated by an average of 41.43cm ($SD = 50.10$ – real mean distance: 260cm). The ratio between estimated and real distance equals about 84 %. A significant underestimation of visual distance was therefore found, $t(19) = 4.0$, $p < .001$, $d = 0.89$. More overestimation occurred in the blind room compared to the visible condition, $t(19) = 2.17$, $p = .043$, $d = 0.48$.

Table 3: Visual Distance Compression

	Visible Room		Blind Room	
Real Distance in cm	Estimated Distance in cm			
	M	SD	M	SD
175	157	54.5	135	34.1
276	238	60.4	206	53.5
329	68.6	72.5	277	72.5
Ratio	89%		79%	

4.4.3 Discussion of Experiment 2

In the second experiment, we tested auditory distance perception with and without visual scene. The HMD either displayed a replication of the physical room in which participants were present or a "blind" version, with minimal spatial cues. The measurement method used was found to influence the indicated sound source distance in VR and this effect was modulated by room visibility. In addition, effects on subjective experience (presence, and vertigo) in VR were found. In experiment 2, participants had to indicate the perceived distance of real sound sources via three different methods: by placing a loudspeaker icon, walking or verbal report.

To begin with, the effect of the different measurement methods depended on the room model which was displayed. Higher accuracy of sound distance perception was expected in the blind condition due to enhanced focus on auditory cues since studies showed that closing the eyes led to higher auditory attention (Wöstmann et al., 2020) and brain activity in auditory cortex regions (Marx et al., 2004). Dominance of the visual system over other senses was found to be reduced by eye closure (Brodoehl et al., 2015). Though, in this study, the reduction of visual stimuli affected - instead of a benefit – a boost of overestimation of auditory distance using the placement task. Indeed, we found higher visual distance compression in the blind condition. Potentially, the "blind" scene did not offer options to "calibrate" distances and sizes of objects (such as widgets or loudspeaker icons). The visible room included everyday life stimuli such as chairs and tables which may have helped participants adapting to VR. Humans were found to use objects' familiar size to estimate distances to them in laboratory paradigms, especially when spatial cues were reduced e.g. by monocular view (Higashiyama, 1984). Also in VR, adding spatial cues such as texture or furniture reduced the error of object size estimations (Hornsey & Hibbard, 2021).

Furthermore, a main effect of measurement method on distance error was found. As expected, overestimation of source distances was significantly higher for the placement task in comparison to walking and verbal report. Via walking task, participants indicated

numerically lower distances in the blind condition, potentially due to more cautious movements.

Not only effects on sound distance estimations but also on subjective experience in VR were found. Walking and placement tasks resulted in higher presence ratings compared to the verbal task. During the verbal task, the virtual scene was comparable steady only varying acoustically (audio playback) whereas during the walking task, participants e.g. moved through the VR room triggering new sounds via button press, and during the placement task, participants additionally manipulated the visual scene (loudspeaker) by handling the controller. Therefore, higher levels of sense of agency can be assumed for walking and placement in comparison to verbal task. Sense of agency means the feeling of being in control (Jeunet et al., 2018) and was found positively affect presence (Jicol et al., 2021). This concept is therefore suggested as underlying mechanism for the presence differences found between the measurement methods.

During the walking task, in comparison to verbal task not only higher levels of presence but also of vertigo were reported. Participants moved through the virtual scene by physically walking therefore visuo-vestibular incoherence – which was found to induce cybersickness (Kemeny et al., 2020) - was theoretically avoided. Nonetheless, no surveillance was done on whether the resolution of the HMD, in this case, a frame rate of up to 90 Hz for the HTC Vive pro eye, remained steady throughout the course of the task. Since participants continuously walked for about 20 min through the room during the walking task, variations in refresh rate are likely to result in adverse effects. Latency jitter was found to increase cybersickness (Stauffert et al., 2018) as well as higher navigation speed (Kwok et al., 2018).

4.5 General Discussion

4.5.1 Study findings and implications

In two experiments, we investigated the influence of different visual scenes displayed via an HMD on the perceived distance of real sound sources in the physical room. While audiovisual room incongruence had a detrimental effect on sound localization accuracy, no main effect of room visibility was found. Since we found reduced localization accuracy when an audiovisually incongruent scene was presented, but not in general in the absence of visual information, this can be taken as an indicator of multimodal integration effects such as visual capture. The enlarged room dimensions in the incongruent model may have influenced the

processing of the room acoustical auditory cues, such as reverberation and frequency, on which auditory distance perception is known to rely (Zahorik et al., 2005).

This has implications for psychoacoustic experiments in mixed-reality environments. Audiovisual VR offers the major advantage of systematically manipulating the visual and the auditory input separately. This can be helpful in perception-related research, such as investigating distance perception in VR. Also, the findings may be helpful to enhance effectiveness of VR simulations. Especially in the field of virtual acoustics where often acoustic and visual scene are rendered (Neidhardt et al., 2024) the effect of vision on auditive perception must be considered. When transfer-plausibility (production of convincing virtual acoustics, Wirler et al., 2020) is targeted, as for example for architectural applications, visual room models should reflect the real rooms' properties and acoustic features. To add, it should be taken care that the position of individuals in the visual and the physical room as well as the relative distances to sound sources matches (Kroczek et al., 2023). Accurate coherence of audiovisual scene is probably particularly important in application fields such as auditory localization training or screening in VR, e.g. for children with spatial processing disorder (Ramírez et al., 2024). The possibility of enlarging the virtual room by the factor by which the used HMD compresses the scene should be considered and evaluated.

The increased visual distance compression in the blind condition is an unintended but helpful finding (exploratorily). Distance overestimation was also increased in the blind condition (measured with the placement task) supporting the idea of visual distance compression as the main contributor of the auditory distance overestimation effect. Therefore, the displayed scene and possible compression effects should be considered in audiovisual VR.

Also, the method with which sound distance perception is measured as well as the interplay with the VR scene must be considered. Distances indicated via the placement of a virtual agent or object were greater than the physical distances - no matter which room model was displayed. The placement task inquires a high amount of cross modal – audiovisual integration (Etchemendy et al., 2018) potentially making it susceptible towards the visual compression effect when using an HMD. Cross modal integration is also needed in the walking task, but here, further cues from body movement and proprioception could have compensated for visual compression. For experiment 1, nonetheless, the placement task was irreplaceable. The walking task would not have been possible due to the physical walls and the verbal task lacks multi-dimensionality. It should also be noted that the audiovisual room divergence effect found was the result of a within-subject manipulation, and therefore independent of the placement task contributing to greater overestimation.

Concerning future studies on sound source localization in VR, one may ask whether the placement task is suitable, given the increased distance estimations. However, the subjective data is in favor of the placement task, being rated as (numerically) least difficult, not increasing adverse effects but resulting in comparably high presence. While distance overestimation was highest, standard deviations were not increased. When minding the systematic overestimation, the placement task can be seen as agreeable method to measure sound distance perception in VR and may be especially helpful to measure within-subjects effects. The walking method is evenly suitable regarding task difficulty, certainty, and presence. Though, adverse effects are increased. Also, the logistic effort for the walking task is the highest since unobstructed movement through a room is needed. This restricts the walking task to real rooms and loudspeakers placed laterally to or above the participants. Possible solutions for these restriction are virtual-only movements e.g. controlled via joystick and virtual sound sources, e.g. binaural auralizations (for a review: Brandenburg et al., 2020). Last, the verbal task resulted in the most accurate distance estimations (in terms of average distance error closest to zero), but also in numerically high variance between participants. Furthermore, the verbal task resulted in the worse subjective experience in VR (based on presence, difficulty, and certainty ratings). Also anecdotally, the participants gave poor feedback on the verbal task, labelling it as unpopular task. To sum up, depending on the goal of an application, measurement methods should be chosen, and the peculiarities of VR should be considered.

Overall, the distances to the two nearer loudspeakers (< 2.70m) were overestimated, but not to the farthest loudspeaker (~ 3.80m). In the psychoacoustic literature, a compressive power function between perceived and physical distance is described - – only close sound sources (< 1.90 m approximately) are overestimated (for a review: Zahorik et al., 2005). A possible conclusion could be that the compressive power function could also apply to virtual environments but may be “shifted”. This shift may be explained by the peculiarities of visual virtual scene, e.g. the compression of visual distances. Future research should aim at describing a power function for distance estimations in VR based on a meta-analyses as used by Zahorik et al. (2005) for real environments when an adequate number of studies on distance estimation in VR can be found.

Besides effects of visual scenes on sound distance perception, subjective experience in VR was investigated. Contrary to our expectations, no effect of audiovisual room incongruency was found on ratings of presence and realism. Human perception is prone to the construction of stable and consistent scenes. Participants were found to not ignore changes in

virtual room size even when consistent motion and stereo cues were presented (Glennerster et al., 2006). Potentially, also in this study, human perception involuntarily overcame the audiovisual incongruencies. Besides, breaks in plausibility were found to be independent from presence (Brübach et al., 2022). While the implementation of spatialized audio in VR improved presence, perceived realism was not affected (Hendrix & Barfield, 1995). The authors suggested that the visual scene dominated the realism ratings more than the acoustic scene. All here used visual room models can be seen as visually equally realistic. Maybe this affected the ratings of VR experience more than the audiovisual incongruence. One may conclude that while audiovisual room incongruence should be avoided for experiments focusing on auditory localization, it seems unproblematic concerning for user studies focusing on experiencing in VR.

4.5.2 Limitations and future studies

Methodological limitations may have contributed to our findings, that the closer sound sources were more overestimated than the farer ones. The plausible range for distances concerning the far loudspeaker is more limited than for the closer ones due to the visual back wall, although care was taken that the distance between the far loudspeaker and the back wall is at least the distance between all loudspeakers. Future studies using larger spaced rooms could help to overcome this limitation.

A further methodological limitation concerns the representiveness of the sample. The participants of both experiments were majoritarian students, german motherlanguage speakers, and female due to the structure of students at our faculty and therefore potential participants. The sample should be more balanced in the future to increase generalizability.

4.6 Conclusion

VR scenes can profit from the implementation of realistic auralizations by higher presence. However, VR-specific audiovisual integration may occur due to incongruencies of visual and auditory scene, and visual compression effects when using an HMD. With two experiments, an influence of visual scene on auditory distance perception of physical sound sources was shown. Accuracy was lower when an audiovisual incongruent room was displayed via HMD. Furthermore, the visual scene and the method to indicate estimated sound source distances interacted. Higher distance overestimation was found for a placement task, especially when a room model with reduced spatial cues was used. Presence was increased for tasks (placement and walking) for which a higher sense-of-agency (compared to verbal task) can be assumed.

Therefore, the choice of an appropriate measurement method for sound distance perception should depend on specific goals and needs. To sum up, the interaction of visual and acoustic scene in VR must be considered. Open questions concern the perceptual mechanisms of audiovisual integration in VR and influences on quality of experience.

4.7 Supplemental materials

All supplemental materials are available on OSF (at osf.io/nj8k2, osf.io/vjcer, osf.io/n98xp, osf.io/2c9ry), under a CC BY Attribution 4.0 International license. In particular, they include (1) the preregistration of all hypotheses and statistical analyses, (2) csv files for raw data (log files from Unreal), (3) xlsx files for anonymized demographic data, (4) R files for statistical analyses and creation of figures. Data gained in the experiments were partly presented at the congresses of German Acousticians (DAGA) at Hannover 2022 and Hamburg 2023 and published (non-peer-reviewed) within the conference proceedings (Roßkopf et al., 2023; Roßkopf et al., 2024).

5 Study 2: The impact of binaural auralizations on sound source localization and social presence in audiovisual virtual reality: converging evidence from placement and eye-tracking paradigms.

Sarah Roßkopf, Leon O. H. Kroczek, Felix Stärz, Matthias Blau, Steven van de Par, and Andreas Mühlberger

The following article was accepted on 19 September 2024 and published online on 16 December 2024 in Acta Acustica following peer review. The official citation that should be used in referencing this material is: Roßkopf, S., Kroczek, L. O. H., Stärz, F., Blau, M., Van de Par, S., & Mühlberger, A. (2024). The impact of binaural auralizations on sound source localization and social presence in audiovisual virtual reality: converging evidence from placement and eye-tracking paradigms. *Acta Acustica*, *8*, 72. https://doi.org/10.1051/aacus/2024064.

5.1 Abstract

Virtual Reality (VR) enables the presentation of realistic audio-visual environments by combining head-tracked binaural auralizations with visual scenes. Whether these auralizations improve social presence in VR and enable sound source localization comparable to that of real sound sources is yet unclear. Therefore, we implemented two sound source localization paradigms (speech stimuli) in a virtual seminar room. First, we measured localization continuously using a placement task. Second, we measured gaze as a naturalistic behavior. Forty-nine participants compared three auralizations based on measured binaural room impulse responses (BRIRs), simulated BRIRs, both with generic and individual head-related impulse responses (HRIRs), with loudspeakers and an anchor (gaming audio engine). In both paradigms, no differences were found between binaural rendering and loudspeaker trials concerning ratings of social presence and subjective realism. However, sound source localization accuracy of binaurally rendered sound sources was inferior to loudspeakers. Binaural auralizations based on generic simulations were equivalent to renderings based on individualized simulations in terms of localization accuracy but inferior in terms of social

presence. Since social presence and subjective realism are strongly correlated, the implementation of plausible binaural auralizations is suggested for VR settings where high levels of (social) presence are relevant (e.g., multiuser interaction, VR exposure therapy).

5.2 Introduction

There is a wide range of applications for virtual reality (VR) scenarios in many different scientific fields. For instance, there is broad evidence for the usefulness of VR in psychotherapy (see reviews: Carl et al., 2019; Wechsler et al., 2019). What unites many disciplines that deal with VR is the interest in creating a realistic and convincing virtual environment. An important concept besides realism is presence. Presence is defined as the subjective experience of "being there" – in the virtual scene (Slater, 2018). There are several internal and external factors on presence (for a review: Diemer et al., 2015). Internal factors are among other psychological variables, such as mood or state. In general, higher levels of arousal contribute to presence (Diemer et al., 2015). In addition, external factors such as features of the VR, content, and tasks contribute to presence (Felton & Jackson, 2022). In the field of acoustics, research is conducted on various methods to synthesize acoustic virtual environments (Brandenburg et al., 2020). These methods, further called auralizations, aim at creating a realistic spatial auditory impression (for a review: Neidhardt et al., 2022). High degrees of realism of head-tracked binaural auralizations regarding acoustical properties such as reverberance or source distance could be confirmed (Blau et al., 2021). Furthermore, in an audiovisual virtual seminar room scenario, participants were not able to reliably distinguish between real sound sources (loudspeakers) and binaural auralizations (Stärz et al., 2022). In this study, we will use these head-tracked binaural auralizations and will refer to them as (plausible) binaural auralizations in the following.

Importantly, studies that include spatial audio and improved audio quality (i.e. a higher number of audio channels) have reported increased levels of presence (Poeschl et al., 2013; Skalski & Whitbred, 2010). Furthermore, task-relevance of the three-dimensionality of audio influence presence (Bormann, 2005). Presence is even suggested to serve as a measurement variable for audio quality in VR since ratings are sensitive to changes in bass or sound pressure level (Freeman & Lessiter, 2001). A psychologically highly relevant and specific aspect of presence is social presence, which means the feeling of being together with another person in the virtual environment (Biocca et al., 2003). It was found, that the higher realism of virtual agents enhances social presence (for a review: C. S. Oh et al., 2018). Taken together, this suggests that a positive effect of spatial audio and therefore acoustic realism on social

presence can be assumed. The extent to which plausible binaural auralizations are creating (social) presence and how they influence the perceived acoustic realism in audiovisual VR is an open question. An increase in social presence would be especially valuable for applications in the field of virtual exposure therapy for example social anxiety disorder (for an overview: Emmelkamp et al., 2020).

A further goal of this study was to examine the extent of realism of plausible binaural auralizations in terms of sound source localization accuracy in VR. Close-to-real acoustic properties and indistinguishability of plausible binaural auralizations from real sound sources have already been shown (Blau et al., 2021; Stärz et al., 2022). Especially for natural acoustic stimuli such as speech or music in reverberant acoustic environments such as seminar rooms authentic simulations can be achieved (Brinkmann et al., 2017). An open question is whether the plausible binaural auralizations also allow close-to-real sound source localization in an audiovisual virtual seminar room scenario and whether this affects the subjective experience. Since visual cues are known to influence sound source localization, known as the ventriloquist effect (see L. Chen & Vroomen, 2013), effects of visual VR have to be considered. A further open question is whether individualized binaural auralizations are needed or whether binaural auralizations based on a default specification are equivalent in this context. Evidence on the need for individual head-related impulse responses (HRIRs) in the rendering of spatial audios is inconsistent (see Guezenoc & Seguier, 2018). While in earlier research in particular, sound source localization performance was found to be worse when using generic HRIRs (e.g. Møller et al., 1996), later research indicates an equivalency of individualized and generic HRIRs especially for lifelike stimuli such as speech (Begault et al., 2001; Blau et al., 2021). However, it was also found that even small deviations in sound source localization induced by non-individualized in comparison to individualized HRIRs binaural auralizations of free-field drone noise impaired perceived realism in VR (Jenny & Reuter, 2020). There is also broad evidence that the effect of individualized vs. non-individualized HRIRs is differentially depending on individual characteristics such as hearing experience (Prud'homme & Lavandier, 2020) or particularly deviating head shapes such as in children (Braren & Fels, 2021).

When comparing sound source localization accuracy, not only differences between binaural auralizations have to be taken into account. Also, the peculiarities of audiovisual virtual reality affect sound source localization. Not only the perception of visual objects and rooms is compressed to less than three-quarter of the original size using a head mounted display (HMD) (Buck et al., 2018; Renner et al., 2013) but wearing the HMD itself decreased

sound source localization accuracy of pink or white noise played by loudspeakers (Ahrens et al., 2019; Poirier-Quinot & Lawless, 2023). Providing visual information about possible source positions could in part compensate for the negative effects of the HMD (Ahrens et al., 2019; Huisman et al., 2021). It is also well documented that when an audiovisual scene is being integrated, visual cues dominate the auditory ones (ventriloquist effct, Jack & Thurlow, 1973). Furthermore, we could previously show that visual virtual room information altered distance perception (Roßkopf, Kroczek, Stärz, Blau, Van de Par, et al., 2023). When an audiovisual congruent room was displayed, participants were more accurate in terms of distance estimations than when an oversized virtual room was displayed. Also, the selection of a suitable measurement tool for sound source localization in VR is an important issue. Auditory distance perception is often examined based on verbal reports (of explicit scales or unitless implicit scales) or motoric tasks such as walking (for a review: Zahorik et al., 2005). The influence of the method itself to measure egocentric distances as well as the interaction with the virtual environment must be taken into account. When participants had to estimate egocentric distances (to visual targets) in VR, verbal answers were found to be more accurate than visually guided walking, where participants indicated distances by walking up to turned off visual targets (Maruhn et al., 2019). Visual, haptic, and locomotive feedback can improve egocentric distance perception in VR (Adams et al., 2022), but influences also depend on the technologies used (Buck et al., 2018). In real-life scenario, the movement of visual markers to adjust estimated sound source positions – further called placement paradigm - was found to provide more accurate estimations than verbal reports (Etchemendy et al., 2018). Furthermore, the placement paradigm enables continuous measurement of distance and azimuthal sound source perception. The placement task was therefore used in several studies and was found to be a sensitive tool for investigating differences in distance perception (Kroczek et al., 2022; Roßkopf et al., 2023).

One further advantage of the use of VR is that relatively complex scenes can be created and that high flexibility for different measurement tools and strategies is provided. By tracking gaze behavior, a comparable naturalistic task for sound source localization is provided. Humans tend to look at people who are speaking. This gaze behavior is modulated by audiovisual speech integration (Foulsham & Sanderson, 2013). Speaker-directed gaze orientation is not only part of multimodal social attention but also influences auditory perception. Acoustic cues are derived more accurately when presented frontally or slightly lateral to the head (Middlebrooks & Onsan, 2012). Additionally, shifting the gaze toward a sound source enhances cue discrimination even when the head is not moved (Maddox et al.,

2014). Gaze behavior can also be used as a measure of sound source localization (Schleicher et al., 2010). Eye-tracking paradigms provide a naturalistic and implicit tool for measuring attentional resources and can be seen as a task that has a high relevance for real-world situations (validity) and is at the same time as standardized as possible (Roßkopf et al., 2024).

This study aimed at investigating sound source localization with two different paradigms. First, a placement paradigm was used to precisely and continuously measure sound source localization accuracy. Second, an eye-tracking paradigm was used to gain evidence of the usability of this unobtrusive and naturalistic measurement method and on the localizability of different Audio Conditions in a more complex seminar room scene including virtual agents. The data of 25 participants on the eye-tracking paradigm has already been published in the proceedings of the Forum Acusticum 2023 at Torino (Roßkopf et al., 2024). Since we regard externalization of auralizations as a prerequisite for more or less accurate sound source localization, we also investigated in-head localization or externalization of Audio Condition.

We compared the plausible binaural auralizations to real audio sources (loudspeakers) and an anchor (state-of-the-art three-dimensional[3D]-audio-sound implemented within the VR engine – Steam Audio v 4.1.4, Valve Corporation, 2022, Bellevue, WA, USA). The anchor was selected due to its practical relevance and external validity as a common and easy-to-implement audio presentation technique in research where audio effects are typically not the main target of investigation (e.g. VR exposure therapy). Notably, while the steam audio engine can be seen as such a practical technique, it also incorporates important features such as room geometry, surface material and head tracking. Based on these thoughts, the steam audio engine is an ideal comparison to test whether more sophisticated yet technically complex auralizations (such as stimuli rendered with RAZR) can yield improvements in VR experience (i.e. with respect to social presence and subjective realism) which justifies the increased technical complexity. We, therefore, investigated whether the accuracy of sound source localization of plausible binaural head-tracked binaural auralizations equals that of real sound sources and is superior to the state-of-the-art game-engine anchor. Furthermore, we were interested in whether similar effects for subjective experience in terms of social presence and perceived spatial audio quality can be achieved and whether these dimensions are correlated.

We hypothesized that all plausible binaural auralization methods are equivalent to real audio sources regarding sound source localization. This was investigated in the placement paradigm with respect to distance and azimuthal localization. In the eye-tracking paradigm,

sound source localization accuracy was investigated in terms of how often participants directed their gaze toward the virtual agent at the sound source position. If one comparison results in significant differences, the hypothesis must be declined. Next, we hypothesized that the binaural auralizations simulated with RAZR (Wendt et al., 2014) which are based on generic head-related impulse responses (HRIRs) are equivalent to simulated BRIRs (RAZR) using individual HRIRs (simIndivHRIRs) and also to binaural auralizations based on measured BRIRs (measHATS). We therefore hypothesize that the binaural auralizations are all equivalent regarding sound source localization. Third, we expected similar effects (equivalency of binaural auralizations and real audio source; equivalency of binaural auralizations) for subjective experience measured via two rating items, first concerning social presence, and second concerning subjective realism. Since the anchor was found to be perceived less frequently externalized in pilot tests when it was played among the other audio conditions, we expect inferior sound source localization of the anchor condition. Lastly, we hypothesize superiority of the binaural auralization and the loudspeakers over the anchor concerning social presence and subjective realism, since they were precisely tailored to each other and the experimental room. All hypotheses were preregistered (https://osf.io/9yqf7; https://osf.io/2q4y3).

5.3 Methods

The goal of the study was to compare three different plausible binaural auralization with loudspeakers and an anchor in audiovisual virtual environments. In our investigations, two different measurement paradigms were implemented in a virtual seminar room (see Figure 12 and Figure 13). Main outcome variables were sound source localization accuracy and subjective experience in VR (ratings of social presence and realism).

5.3.1 Sample

Healthy adult individuals with self-reported unimpaired hearing, normal or corrected to normal vision, and German-speaking experience of a minimum of 5 years were included in the study. Our sample ($N = 49$) consisted of 38 female and 11 male participants aged between 19 and 46 ($M = 23.2$, $SD = 4.6$). The majority of participants were students ($n = 44$). All participants gave written informed consent. The study was in line with the Declaration of Helsinki and approved by the local ethics committee (University of Regensburg).

5.3.2 Room and visual virtual setup

The experiment took place in a seminar room of the University of Regensburg (room size: 10.6m x 7.1m x 3.3m, reverberation time: 0.91s). The room consists of four concrete walls, one of which is equipped with a large mirrored window, an acoustically optimized ceiling, and a carpet on concrete floor (see Figure 12 and Figure 13). For the visual virtual room, we created a photorealistic model of the seminar room with the Unreal Game Engine (v 4.27, Epic Inc.) and Blender (v 2.79, Blender) using textures based on high-resolution photographs (as already described in the previous study: Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2023). The visual virtual environment was presented via an HMD (Vive Pro Eye, HTC). This device was also used for the measurement of gaze by eye-tracking. For audiovisual virtual reality, an inaudible workstation with passive cooling was used (Silentmaxx PC Kenko S-770i). The visual representation of the virtual room (in terms of HMD position and direction) was matched to the real room via an in-house-developed two-point calibration technique using custom-made mounts for the HTC motion controller (Kroczek et al., 2023). Since the headphones were mounted to the HMD this also implied partial calibration of the headphone position (only pitch, yaw and roll data). Data on our calibration technique was collected and a very high accordance of real and virtually visible positions could be affirmed (see Kroczek et al., 2023). Virtual agents were created using MakeHuman (v 1.2) and Blender (v 2.79).

5.3.3 Auditory setup

5.3.3.1 Audio Conditions

We compared five different audio presentation modes (for an overview see Table 4). The first Audio Condition involved loudspeakers in the room as real sound sources, which provided the best possible comparison condition. Next, we used head-tracked binaural auralizations based on three different BRIR sets, for which high plausibility was found (Stärz et al., 2022).

More precisely, the second Audio Condition, referred to as measHATS, was a binaural auralization based on BRIR sets that were measured in the real room using a commercial head-and-torso-simulator (HATS; Kemar type 45BB, GRAS Sound and Vibration A/S, Holte, Denmark). Therefore, the source directivity of the loudspeakers (also used as comparison condition) were implicitly included. The head-above-torso orientation of the HATS was varied between -90 and 90° in 5° steps, resulting in 37 azimuthal orientations. The spatial resolution is based on the proposed minimum BRIR grid resolution by Lindau et al. (2008). It was shown that for the majority of listeners (> 95%) the here used spatial resolution of BRIRs is sufficient to create a plausible binaural simulation for natural stimuli such as music in

reverberant spaces (Lindau & Weinzierl, 2009). MEMS microphones (TDK type ICS-40619, TDK InvenSense, San Jose, CA, USA) inserted into the ear canals of the HATS using PIRATE earplugs (Denk et al., 2019) were used for the measurements. BRIRs were measured using multiple exponential sweep stimuli (for further details see Blau et al., 2021). The elevation angle was fixed at 0°, implying that the sound source remained static even when participants raised or lowered their head.

The third and fourth Audio Condition were binaural auralizations based on BRIR sets which were simulated using RAZR (v 0.962b, T. Wendt et al., 2014). The simulated reverberation time T20 was fitted to previously measured monaural impulse responses of the seminar room. Additionally, the source directivity of the loudspeakers (same as for measHATS and comparison condition loudspeaker) was included in the simulation via a database with directivities measured by ourselves (see Blau et al., 2021). The simulated room impulse responses were combined with measured head-related impulse responses (HRIRs). The measurement system for the HRIRs (see Figure 10) is a replication of the setup constructed and used at Jade Hochschule Oldenburg, for further details see (Blau et al., 2021). Simulated BRIRs were obtained for 37 azimuthal head-above torso orientations (-90° to 90° in 5° steps) and nine elevation angles (-30° to 30° in 7.5° steps).

Table 4: Overview of investigated Audio Conditions. For plausible binaural auralizations detailed information on used binaural room impulse resonse (BRIR) sets, used head-related impulse response (HRIR) set and headphone equalization (HPEQ), spatial resolution, and frequency independent direct-to-reverberant energy ratio (DRRs) in dB of BRIRs are given.

	Name of Audio Condition	Acoustic Specifications						
Plausible binaural auralizations		BRIR	Loudspeaker Directivity Genelec 8030b	HRIR	HPEQ	Spatial Resolution of BRIRs elevation	azimuth	DRR
	measHATS	measured in real room	implicitly included	Kemar HATS	individual	-	-90° to 90° in 5° steps	1.65
	simIndivHRIRs	simulated using RAZR	Measured	individual	individual	-30° - 30° in 7.5° steps	-90° to 90° in 5° steps	-0.54
	simHATS	simulated using RAZR	Measured	Kemar HATS	individual	-30° - 30° in 7.5° steps	-90° to 90° in 5° steps	0.39
Comparison conditions		Description						
	Loudspeaker	Loudspeakers as real sound sources in the room (Genelec 8030b)						
	Anchor	head-tracked binaural 3D auralizations created by Steam Audio v 4.1.4 (Valve Corporation)						

In Audio Condition number three, further referred to as simIndivHRIRs, individually measured HRIRs were used for the rendering of BRIRs. In Audio Condition number four, subsequently referred to as simHATS, generic HRIRs were used, measured with the above-described HATS. The simulated ear height of the plausible auralizations depended on the experimental task. In the placement task, the simulated and real ear height was set to 1.30m since participants were seated in an auditorium (using a height-adjustable chair). During the eye-tracking paradigm, the simulated ear height was set to 1.60m, since participants took up the lecturer position in front of the auditorium. No adjustment of participants' real ear height was made here, as the natural standing position of participants in front of an auditorium was targeted.

Last, the fifth audio mode, subsequently called anchor, consisted of head-tracked binaural 3D auralizations created by a state-of-the-art audio engine (Steam Audio v4.1.4, Valve Corporation, Bellevue, WA, USA) implemented in the Unreal Engine. Real-time ray tracing was used for modeling physics-based reverb. We used the above described virtual room as static geometry, and specified the room acoustic via predefined acoustic material properties (e.g., carpet for the floor). Frequency-dependent occlusion and sound propagation via nearest-neighbor option were chosen. The volume attentuation was adapted on a one-individual perceptive level towards the loudspeaker condition, to avoid salient loudness differences.

Figure 10: Auditory measurement system of individual and generic (HATS) head-related impulse responses (HRIRs).

5.3.3.2 Technical setup

We compared binaural auralizations to real sound sources in the room. Two-way active loudspeakers (Genelec 8030b, Genelec Oy, Isalmi, Finnland) were used as real sound sources

in the room. All other Audio Conditions were presented using a headphone amplifier (Lake People G103P, Lake People Electronic GmbH, Konstanz, Germany) and extra-aural headphones (AKG K1000, AKG Acoustics GmbH Vienna, Austria), which were mounted to the HMD with custom-made 3D printed supports (Stärz et al., 2023). Compared to circumaural headphones, the spectral influence of the extra-aural headphones on the sound field produced by a real loudspeaker is smaller (Schneiderwind et al., 2020) especially for speech stimuli (Lladó et al., 2022) but it cannot be completely excluded. Likewise, the presence of the HMD may introduce subtle spectral colorations in comparison to not wearing an HMD. Nonetheless, no differences regarding plausibility could be found for binaural auralizations based on measurements with or without HMD in a prior study (Stärz et al., 2024). For playback on loudspeakers and headphones, an external audio interface (RME Fireface, UC, Audio AG Haimhausen, Germany) was used.

5.3.3.3 Audio Stimuli

The stimuli consisted of dry recordings of female speech and were derived from a German learning program (studio21 A1 und A2, Cornelsen Verlag, Funk et al., 2013). The stimuli were loudness normalized (based the integrated loudness function from the Matlab "Audio ToolboxTM") following EBUR 128 (MathWorks, 2022) and Hann windowed (10 ms) to avoid cutting artifacts. For the placement paradigm, German greetings (e.g. "Hallo") were used. For the eye-tracking paradigm, typical language course statements from one word (e.g. "station") to five-word sentences (e.g. "What is it called in German?"). The order of stimulus presentation was pseudo-randomized via randomization lists. For the presentation of stimuli, we created five different randomization lists per paradigm, each beginning with a different audio mode (lists were counterbalanced across participants). Three blocks per list were created within which combinations of stimuli, speaker position, and Audio Condition were repeated equally often. The stimuli were pseudo-randomized within the three blocks with the following constraints: not more than three repetitions of the same Audio Condition, same position, and same utterance.

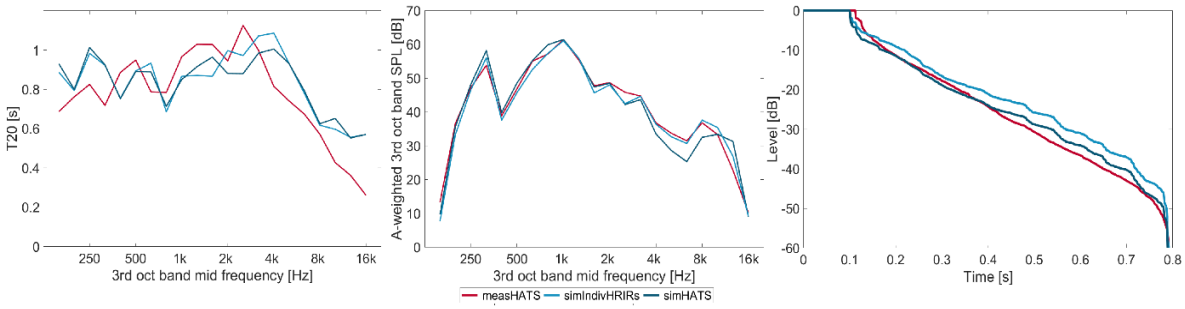

Figure 11: Objective frequency dependent data derived from BRIRs (frontal-close speaker see Figure 12). Left: reverberation time (T20), middle: A-weighted third-octave band sound pressure levels after convolving the speech stimulus used in the listening test with the BRIRs, right: energy decay curves.

5.3.4 Design

We manipulated the Audio Condition with 5 levels: real sound source, measured HATS BRIRs, simulated BRIRs based on individual HRIRs, simulated BRIRS based on HATS, and anchor. We further varied the position (angle, distance) from which the sounds were played (loudspeakers) or simulated. With two different tasks, we investigated the influence of binaural auralizations on sound source localization, social presence, and subjective realism in virtual reality.

5.3.5 Procedure

The experimental procedure comprised two appointments. On the first appointment, participants gave written informed consent and filled in questionnaires on demographic data, their hearing experience, and their degree of social anxiety with the social phobia inventory (SPIN, Sosic et al., 2008). Furthermore, the individual measurement of participants' head-related impulse responses (HRIRs) and an individualized headphone equalization (HPEQ) was conducted.

The second appointment consisted of two parts, first the placement paradigm, second the eye-tracking paradigm. For the general preparation of both experimental parts, the current status of hearing impairment (self-report) was collected and impaired participants (e.g. current otolaryngologicical symptoms) were excluded. Next, participants were informed about and prepared for the case of motion sickness. Further, they reported about their affective state via the positive and negative affect schedule (PANAS, Krohne et al., 1996). Next, the HMD was fitted to participants by adjusting the position to the one used during the individual HPEQ measurement, via recorded strap positions (individually adjusted length of side strap and length of Velcro fastener on crown). Participants entered the seminar room "blindfolded" (by wearing the HMD) and were guided to the starting position by the experimenter and virtual

footprints. This procedure allowed participants to remain unaware of the positions of the loudspeakers.

Afterward, the first experimental part consisting of the placement task was conducted (see Methods placement task). Presence during the placement paradigm was measured via the multimodal presence scale (MPS, Makransky et al., 2017). After completing the placement task, participants took a break outside the experimental room and VR. Then, participants accomplished the second experimental part, the eye-tracking paradigm, after which again, presence was measured via MPS. Additionally, post-assessment questionnaires on both experimental parts and mood (PANAS) were conducted. Neither before, during nor after the experiment did the participants see the room or the source positions since they always wore the HMD when being there.

5.3.6 Sound source localization measurement

5.3.6.1 Placement Paradigm

Specific setup

Participants had to place a virtual agent at the position in the virtual room, where they assumed the sound source. The placement was accomplished by pointing the HTC Vive motion controller towards a selected position at which then, the agent appeared and which had to be confirmed by another button press. The agents could be placed continuously on the floor of the virtual room without any prior restrictions. More precisely, the x- and y-axis, but not the z-axis coordinates of the position of the virtual agents could be altered to indicate sound source location. The agent only appeared when a specific position has been selected (is hidden when the sound is played) to prevent a possible bias of visual information on auditory localization (ventriloquist effect, Jack & Thurlow, 1973). If a sound was in-head localized, participants were instructed to place the agent close to their own chair (radius of 80 cm) resulting in the agent disappearing only a sphere being left.

The gaze direction of agents was shifted during positioning so that agents always kept facing toward the participants. The sound sources also faced the participants. The loudspeakers in the room and the simulated sound sources were at a height of 1.15, corresponding to the height of the mouth of the virtual agent. There were 4 different source positions; speaker 1: 2.80m, 0°; speaker 2: 4.80m, 0°; speaker 3: 2.45m, 45°; speaker 4: 2,15m, 90° (see Figure 12). During the experiment, the position of participants, virtual agents (position after placement), and the sound source positions were tracked. Participants were seated in the auditorium of a seminar room. Speech stimuli, that were supposed to evoke an

orientation reaction (German Greetings) were played from the source positions. At the end of this experimental part, a brief assessment of perceived elevation of sound sources was conducted to gain preliminary insights whether Audio Conditions were perceived elevated. Therefore, participants had to adjust the height of a loudspeaker icon to the height at which they perceived a sound source. For all Audio Conditions, the same source position (speaker 1) and sound stimuli were used (one trial per condition).

The procedure of the placement paradigm

Participants were seated on a height-adjustable chair placed in the auditorium of the room. The height of participants' ears was adjusted to 1.30m. Then, the HMD displayed an exact replication of the seminar room. The participant was handed the controller with which the rest of the experiment could be conducted. Furthermore, participants were instructed to move their head only on a horizontal plane of -90 to 90°.

To learn the handling of the controller and the placement task practice trials were conducted where participants had to, first, place an agent on a visual target, second, place the agent at the position in the seminar room, where they assumed the sound source, third, indicate in-head localization of sounds, and last, use the interface for ratings. The practice trials were repeated until participants succeeded in placing the agent on a target with no more than 20cm deviance and when predefined rating buttons were selected.

Following the practice trials the main task was started. It consisted of three blocks with 100 trials in total. After each block, participants were instructed to take a break. Each trial had the same procedure. Before the sound presentation, participants had to orient towards a fixation cross on the frontal wall. If the rotation of the HMD deviated more than 10° from the fixation cross, a red text was displayed, instructing participants, to "Please look straight forward." If verified, the sound was played at the predefined (via randomization list) location. Then, participants placed a female agent at the location in the room, where they assumed the sound source. Participants were able to change the position as often as they liked and there was no time limit for the task. Finally, participants had to confirm the chosen position with another button press and the next trial started after a delay of 3 seconds. After a fifth of trials, (in total 20 trials specified via a randomization list), a rating followed the localization task. Each rendering at each position had to be rated once concerning the (social) presence and the subjective realism. The rating interface was implemented in the script and was handled via the controller. The first item of the rating was the social presence rating: "Ich habe den Eindruck, dass die Begrüßung gerade von einer anwesenden Person stammen könnte." which translates as "I had the impression that the greeting a moment ago could have come from a present

person.” The second item of the rating was: “Der Klang war so wie in einem Seminarraum.” which translates as “The sound was like being in a seminar room.” We used the second item as indicator for subjective realism of Audio Conditions. The ratings consisted of the statement and a 9-point Likert scale beneath it. The furthest left button was labeled “stimme nicht zu” [I disagree”], and the furthest right button was labeled “stimme zu” [I agree]. We previously validated in a pilot study with five participants whether laypeople naïve towards room acoustic research can comprehend and answer these items. We tested six different room acoustic quality and realism items adapted from the Room Acoustical Quality Inventory (RAQI, Weinzierl et al., 2018) and three social presence items adapted from the Multimodal Presence Scale (MPS, Makransky et al., 2017) and chose the items which were the easiest to answer.

After all 100 trials had been presented, participants accomplished the task of indicating the perceived elevation of sound sources. Then, participants were guided out of the experimental room and took off the HMD. Last, participants completed a questionnaire on presence, the MPS, and a post-experiment questionnaire.

Figure 12: Setup of placement paradigm. On the left: position of participant and loudspeakers in the room. In the middle: Placement Task (Agent had to be placed at perceived sound source). On the right: source and listener positions.

5.3.6.2 Eye-tracking paradigm

Specific setup

Participants were instructed that they would participate in a virtual language learning course. Participants were positioned in front of the auditorium at the lecturer position with a virtual visual notebook in front of them. All instructions, rating scales, and vocabulary stimuli were presented on the screen of the virtual notebook. The auditorium of the seminar room was filled with 16 virtual female agents. The virtual agents were animated sitting on a chair and breathing. They were positioned to fill the whole auditorium, see Figure 13. In the notebook, different written words were displayed. Participants were instructed that the words would be read aloud by one of the agents. To measure sound source localization, participants were

instructed to look at the location in the room where they assumed the sound source. Gaze behavior was recorded and analyzed during the task. If participants did not externalize a sound, which means that in-head localization of a sound occurred, they were instructed to direct their gaze toward a blue button on the keyboard of the virtual notebook.

Figure 13: Setup of eye-tracking-paradigm. On the left, the position of participants and loudspeakers (in block 1) in the room are shown. In the middle, a visual virtual room with the eye-tracking task for sound source localization is depicted (participants had to look at the perceived sound source). On the right: source and listener positions.

The 120 trials of the eye-tracking paradigm were conducted in two blocks with varying source positions. The position of in total eight agents exactly matched the (virtual) loudspeaker position in the room (four per block). The real and virtual loudspeakers were directed forward (parallel to the side walls) and accordingly, all virtual agents directed their gaze straight forward. The position of the agents' mouths was at 1.15m, which corresponded to the height of the acoustic center of the loudspeakers. All agents wore a face mask, to avoid influences of visual cues on localization (ventriloquist effect, Jack & Thurlow, 1973). Loudspeakers were placed at distances from 2.70m to 6.80m, and at azimuthal angles of 2° to 27° (see Figure 13).

At the end of this experimental part, perceived loudness was assessed to gain preliminary insights on possible differences between the Audio Conditions. Therefore, participants were asked about the perceived loudness of a first sound in comparison to a second sound (0 = much quieter, 1 = a bit quieter, 3 = neutral, 4 = a bit louder, 5 = much louder). The comparison (only one trial per Audio Condition) was always against the loudspeaker condition, but the order was alternated.

Procedure

This experimental part started with the five-point eye tracking calibration procedure provided by and presented with the Vive Pro Eye. The producers claim an accuracy of 0.5 – 1.1° for this eye-tracking system (Vive.com). We furthermore validated the accuracy of eye-tracking for each participant. Therefore, 24 visual target icons were presented at the positions where the agents were placed in the main experiment and participants were instructed to look towards the bull's eye of each target and then press a button. Gaze acuity was measured in

degree for angle deviance and cm for distance deviance. After this, several practice trials were conducted to ensure understanding and manageability of tasks. Handling of ratings, the eye-tracking task, and what to do when sound was in-head localized were practiced. The instruction was to look towards the spot where participants assumed the sound source.

After the practice trials, the first 60 trials (first block) were run. These were followed by a break (about 5 min) during which loudspeakers were rearranged and during which participants had a seat next to the previous position within the experimental room still wearing the HMD. Possible auditory cues during the rearrangement of the loudspeakers were masked by brown noise played via headphones. The loudspeakers were rearranged to increase the overall number of source positions despite a limitation by number of speakers and the interface. Then, the next 60 trials were conducted. All trials started with the visual display of the vocabulary item (word or short sentence) in the notebook (see Figure 13). The orientation of the participant towards the notebook during the sound onset was controlled. If the rotation of the HMD exceeded 10°, a red text was displayed, instructing participants, to "Please look towards the screen." If verified, the sound was played back at the designated location. Head movements and gaze toward the source were encouraged as soon as the sound was played. The gaze behavior was recorded and analyzed for three seconds. If no valid fixation on an object in the scene (no adjustment of gaze direction or fixation of the wall) was found, the trial was repeated. If the task was completed, the visual display of the vocabulary item disappeared and after an inter-trial interval of three seconds, the next trial started.

After a sixth of the trials, the rating scales were presented in VR and had to be completed. As in the first task, each rendering at each position had to be rated concerning (social) presence and subjective realism. The rating interface was displayed on the virtual notebook, and was handled via the controller. The first item of the rating was the social presence rating: "Ich habe das Gefühl, dass gerade eine anwesende Person zu mir gesprochen hat." which translates as "I have the feeling that a person present has just spoken to me." The second item of the rating was the subjective realism rating, which was the same as in the placement task.

At the end of the localization task, the perceived loudness of the auralizations each in comparison to the loudspeaker was assessed. After the VR experiment, participants were guided to the anteroom, and again questionnaires on the experiment (difficulty of task, hypotheses, etc.) and experience in VR (MPS) had to be answered. Then, the final assessment of mood was completed via the PANAS questionnaire.

5.3.7 Outcome variables

All analyses for hypotheses tests were preregistered (https://osf.io/9yqf7; https://osf.io/2q4y3).

5.3.7.1 Sound source localization accuracy

Placement paradigm

Deviations between estimated and real angle (in degree) and distance (in cm) were calculated as primary outcome variables for sound source localization accuracy. Per Audio Condition and participant, 20 trials were averaged. The angle deviance was computed using the dot product of the vector between participant and real sound source and the vector between participant and estimated sound source position, converted to degrees for better interpretability (all analyses scripts are accessible in a public repository (https://osf.io/9yqf7; https://osf.io/2q4y3). The distance deviance (xy plane only) was computed as follows: the Euclidian distance between the participant and the estimated sound source position is subtracted from the Euclidian distance between the participant and the real sound source. As a consequence, a distance deviance of 0 equals perfect distance estimation, and a positive distance deviance indicates an overestimation of distance. Following our preregistered hypothesis rationale, an equivalency of Audio Conditions concerning localization accuracy will only confirmed, if equivalency can be shown for both dependent variables.

Eye-tracking paradigm

As the primary outcome variable for sound source localization accuracy, the rate of fixations on correct agents per participant and Audio Condition was calculated. Only the first fixation was analyzed following a pre-registered analysis plan. Gaze behavior was analyzed offline using a custom Matlab script (v R2022a, The MathWorks, Inc., Natick, MA, USA) which categorized the gaze as fixation or saccade behavior. Fixations were defined using both velocity (< 75°/s) and gaze duration (> 200 ms) criteria [59]. Additionally to correct fixations, angle deviance as well as distance deviance of fixated and real source position were calculated as further indicators for sound source localization accuracy (not preregistered).

Supplementary indicators

Additionally to these primary outcome variables (angle deviance and distance deviance), we computed following supplementary indicators to classify data concerning sound source localization (non-preregistered analyses): We analyzed the amount of trials (in %), in which the agent was placed outside the walls of the visual room (Trials outside Room) and classified them as invalid. We further classified data as invalid, if the sound was not perceived

externalized or if front-back confusions occurred. As a global indicator for localization accuracy, we computed "overall localization error" (in cm), which is calculated by the Euclidian distance between the estimated and real position of the sound source. As an indicator for systematic angular deviance on the horizontal plane (to left of right side), the azimuthal error was calculated, which ranges from -180° to + 180°. For the absolute distance error (in cm) as an indicator of the overall accuracy of distance estimation (regardless of whether the distance was overestimated or underestimated) the mean of absolute distance deviance was calculated.

5.3.7.2 Subjective experience

Ratings of social presence and subjective realism were analyzed to investigate subjective experience during audio-visual presentation. The mean rating value per participant and Audio Condition was computed as the dependent variable.

5.3.8 Statistical analyses

Statistical analyses were conducted in the R environment (v 4.3.2, Development Core, 2019) using the packages lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), emmeans (Searle et al., 1980), and BayesFactor (Bayes-package, Morey & Rouder, 2014). Generalized mixed-effect models of binomial family were computed for categorical data (externalization, front-back confusion, correct fixations), linear mixed-effect models were computed for continuous data (angle deviance in °, distance deviance in cm). All models included fixed effects for Audio Conditions. Likelihood ratio tests were conducted to affirm the random effects structure. The final models included random slopes for Audio Conditions by subjects. We used the Bobyqa option of the lmerTest package to optimize models. To test the specified models for the significant effect of the fixed effects, *F*-values for continuous data and $\chi 2$ values for categorical data are computed with an analysis of variances. Post-hoc simple contrast comparisons were conducted to test for differences following the preregistered hypotheses (https://osf.io/9yqf7; https://osf.io/2q4y3). Alpha was set to 5%, tests were corrected for multiple comparisons using the Holm method (Holm, 1979).

Since we hypothesized not only differences between but also equivalency of particular Audio Conditions, we conducted tests for equivalency. Following our preregistered analysis plan, if post hoc comparisons revealed no significant difference, Bayes factors were computed to test the probability of an equivalency of Audio Conditions. For higher clarity and better interpretability, Bayes factors for paired *t*-tests of the respective comparison of Audio Conditions were computed. We used non-informed prior distributions. We defined Bayes

Factors greater than three as confirmative following the suggestions of Wagenmakers (Wagenmakers, 2007). All anonymized data, as well as analysis scripts, are accessible in a public repository (https://osf.io/9yqf7; https://osf.io/2q4y3).

5.4 Results

5.4.1 Placement paradigm

Figure 14 provides an overview of real and estimated source positions from the placement task for each Audio Conditions except the anchor. It can be seen, that front-back confusions occurred in all Audio Conditions, but more often in binaural auralization trials. All in all, the placement patterns appear to be comparable. Overestimation of distances occurred frequently. In Table 5, all outcome variables are reported. Note that per Audio Condition and Participant, 20 trials were analyzed. Thus, for outcome variables that are reported in %, values < 5% implicate that on average in less than one trial per participant the relevant characteristic of the outcome variable (e.g., internalization) occurred.

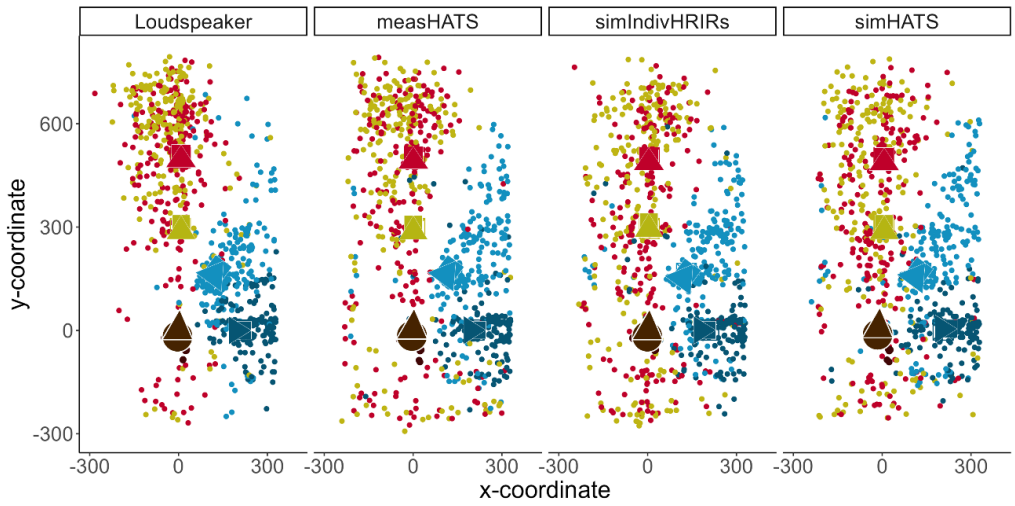

Figure 14: The positions of real sound positions (loudspeaker icons), participants (dark brown, schematic head for frontal direction heading nose), and the estimated sound positions (color depends on the real sound position of that trial) are shown as a function of their y- and x-coordinate in the virtual room. Note that participants' forward orientation is towards the positive x-coordinate.

5.4.1.1 Externalization rate

As can be seen in Table 5, in a vast majority of trials the sounds were not in-head localized (meaning high externalization rates) in all Audio Conditions except the Anchor. A mixed-effect logit model revealed a significant main effect Audio Condition, $\chi 2\ (4) = 119.9$, $p < .001$. Post-hoc comparisons revealed significant differences of Loudspeaker against Anchor ($\beta = -9.11$, $Z = -7.58$, $p < .001$), measHATS against Anchor ($\beta = -11.47$, $Z = -6.85$, $p < .001$),

simIndivHRIRs against Anchor ($\beta = -8.84$, $Z = -8.84$, $p < .001$), and simHATS against Anchor ($\beta = -7.96$, $Z = -9.22$, $p < .001$). No significant differences of Loudspeaker against all binaural auralizations could be found as well as no significant differences of comparisons between the different plausible binaural auralizations.

Table 5: Outcome variables from placement task for audio condition

Outcome Variable	Loudspeaker		measHATS		simIndivHRIRs		simHATS		Anchor	
	M	SD	M	SD	M	SD	M	SD	M	SD
Trials outside Room (in %)	2.9	14.1	2.7	12.5	1.9	10.8	1.9	8.2	1.7	5.6
In-head localization Trials (in %)	0.8	2.4	3.1	13.5	3.6	11.7	4.1	11.6	78.4	31.6
Front-Back Confusions (in %)	4.4	10.0	11.0	17.5	10.7	16.7	12.4	16.8	25.3 [a]	22.4
Invalid Trials (in %)	7.0	15.9	12.9	21.3	14.7	23.5	16.2	21.8	73.3	28.6
Overall localization error (in cm) [b]	185.1	122.6	207.7	125.1	196.9	127.7	200.0	121.0	212.4[a]	111.6
Angle Deviance (in °, 0°- 90°) [c]	11.1	12.7	13.8	17.4	14.4	17.8	15.2	19.3	33.2 [a]	27.3 [a]
Azimuthal Error (in °, -180° - 180°) [b]	-0.6	16.8	-0.03	22.2	-1.4	22.8	1.5	24.5	-4.7[a]	42.8
Distance Deviance (in cm) [b]	117.1	170.9	126.1	175.9	91.0	185.8	95.7	179.0	-142.0[a]	150.5
Absolute Distance Error (in cm) [b]	163.8	170.9	177.0	175.9	164.7	185.8	165.5	179.0	164.9 [a]	150.5

[a] These values cannot be interpreted consistently due to insufficient valid trials. [b] For these analyses invalid trials were excluded.

5.4.1.2 Front-back confusions

Since in more than three out of four trials of the anchor stimuli were not perceived externalized, we did not include this condition in the statistical analysis for differences concerning front-back confusions. A mixed-effect logit model revealed a significant main effect of Audio Condition, $\chi 2$ (3) = 13.92, p = .003. The only significant difference of post-hoc comparisons between Audio Conditions concerning front-back confusions could be found for loudspeakers against simHATS ($\beta = 2.69$, $Z = -2.74$, $p = .036$).

5.4.1.3 Sound source localization accuracy

Trials in which in-head localization or front-back confusion occurred were not included in this data analysis since they cannot be interpreted consistently in terms of localization accuracy (referred to as invalid trials). For this reason, too, the anchor condition was not included in this data analysis due to too many invalid trials. There were 25 participants (more than half of N*)* with not a single valid trial for the anchor condition, and in all participants ($N = 49$) more than half of the trials for the anchor condition were invalid.

Angle deviance

Data from one further participant had to be excluded due to only invalid trials for the measHATS condition. A linear mixed-effect model revealed no significant main effect of Audio Condition, $F(3,65.54) = 2.72$, $p = .051$. Therefore, Bayes factors are computed to gain evidence for the hypothesis, that the binaural auralizations are equivalent to real sound sources and that all plausible binaural auralizations are equivalent concerning azimuthal localization. The Bayes factor for the hypothesis, that the plausible binaural auralizations are

equivalent to loudspeakers is 0.74, which can be interpreted as anecdotal evidence for non-equivalency (Jeffreys, 1998). The Bayes factor for the hypothesis, that simHATS is equivalent to simIndivHRIRs is 4.72, which can be interpreted as moderate evidence for equivalency of these two conditions. The Bayes factor for the hypothesis, that measHATS and both simulated binaural auralizations are equivalent is 6.17, which can be interpreted as moderate evidence for equivalency.

Distance deviance

The same data as for angle deviance was included for the analysis of distance deviance. A linear mixed-effect model revealed a significant main effect of Audio Condition, $F(3,65.16) = 8.28$, $p < .001$, on distance deviance. Post-hoc comparisons revealed no significant differences in Loudspeaker against measHATS, but significantly higher distance deviance of the loudspeaker against simIndivHRIRs ($\beta = 34.08$, $t = 3.46$, $p = .005$), and against simHATS ($\beta = 25.50$, $t = 2.51$, $p = .046$). Furtherly, significant higher distance deviances are found for measHATS in comparisons against simIndivHRIRs ($\beta = 37.79$, $t = 4.34$, $p = .001$) and simHATS ($\beta = 29.21$, $t = 3.67$, $p = .001$). There was no significant difference between simIndivHRIRs and simHATS. The Bayes factor for the hypothesis, that simIndivHRIRs and simHATS are equivalent is 3.46 (can be interpreted as moderate evidence for equivalency).

In prior studies, we constantly found an overestimation of source distances when using visual VR (Kroczek et al., 2022; Roßkopf et al., 2023). Therefore, we compared the mean distance error against zero. With a mean average of 75.82 cm ($SD = 113.41$) over all Audio Conditions except Anchor, there is strong evidence for an overestimation of sound distances, $t(195) = 9.36$, $p < .001$.

5.4.1.4 Subjective experience

Social presence

In Figure 15 the influence of Audio Condition on social presence and subjective realism is displayed. Ratings of all participants were included for analysis of rating data ($N = 49$). A linear mixed-effect model revealed a significant main effect of Audio Condition, $F(4,84.43) = 21.05$, $p < .001$, on social presence. In anchor trials, participants rated their social presence lower ($M = 3.3$, $SD = 2.4$) than in loudspeaker trials ($M = 5.9$, $SD = 2.1$, $\beta = -2.57$), measHATS trials ($M = 5.5$, $SD = 2.4$, $\beta = -2.16$) and both simulated BRIR trials (simIndivHRIRs: $\beta = -2.60$, simHATS: $\beta = -1.96$), all ps $< .001$. One further significant difference could be found. Social presence was rated higher for simIndivHRIRs ($M = 5.9$, $SD = 2.3$) than for simHATS ($M = 5.3$, $SD = 2.2$, $\beta = -0.49$, $t = -2.87$, $p = 0.036$). The Bayes

factor for the hypothesis that participants perceive equivalent social presence during loudspeaker and binaural auralizations trials is 0.61, which can be interpreted as anecdotal evidence for non-equivalency. The Bayes factor for the hypothesis that social presence is equivalent for measHATS and both simulated binaural auralizations is 4.63.

Subjective realism

Overall, the relationship between Audio Condition and subjective realism is similar to the relationship between Audio Condition and social presence. A linear mixed-effect model revealed a significant main effect of Audio Condition, $F(4,91.94) = 15.67$, $p < .001$, on subjective realism. Participants rated the realism of the anchor lower ($M = 3.6$, $SD = 2.2$) then the realism of loudspeakers ($M = 5.5$, $SD = 2.0$, $\beta = -1.93$, $t = -7.58$, $p < .001$), of measHATS ($M = 5.4$, $SD = 2.2$, $\beta = -1.86$, $t = -6.8$, $p < .001$), of simIndivHRIRs ($M = 5.7$, $SD = 2.1$, $\beta = -2.10$, $t = -7.40$, $p < .001$) and lower than simHATS ($M = 5.2$, $SD = 2.0$, $\beta = -1.60$, $t = -6.85$, $p < .001$). No other significant differences could be found in any other comparison between Audio Conditions. The Bayes factor for an equivalency of the subjective realism of loudspeakers and plausible binaural auralizations is 5.19. The Bayes factor for an equivalency of subjective realism of binaural auralizations based on measured BRIRs and simulated BRIRs is 6.43. For the hypothesis that simHATS and simIndivHRIRs are equivalent regarding subjective realism, a Bayes factor of 0.08 was found. This can be interpreted as anecdotal evidence for a non-equivalency.

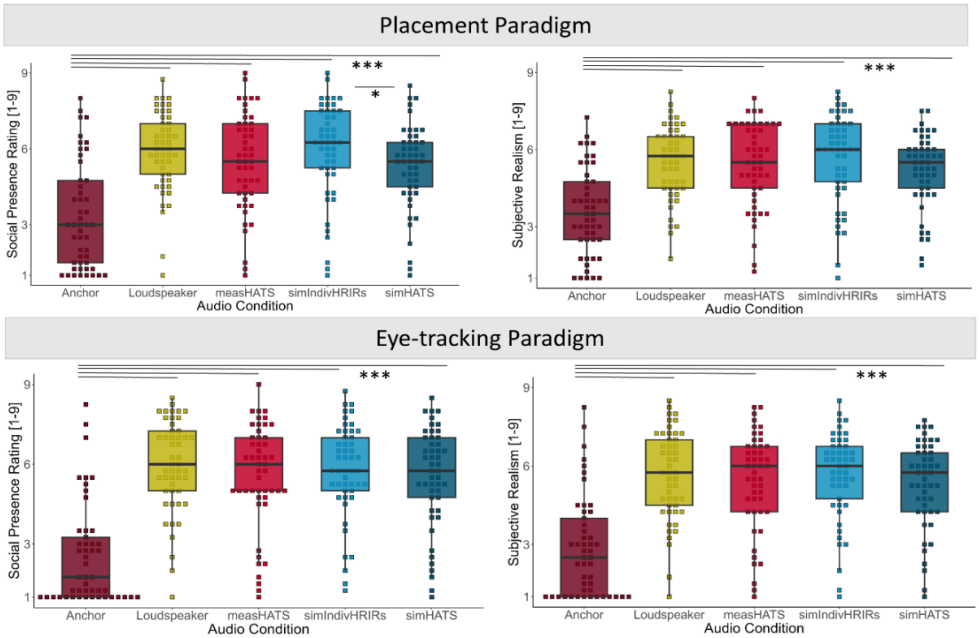

Figure 15: Subjective Experience. On the left: Social presence ratings ["I had the impression that the greeting a moment ago could have come from a present person." (placement paradigm) or "I have the feeling that a person present has just spoken to me." (eye-tracking paradigm); 1 = "I disagree", 9 = "I agree"] as a function of Audio Condition. On the right: Realism ratings ["The sound was like being in a seminar room." 1 = "I disagree", 9 = "I agree"] as a function of Audio Condition and separate for each paradigm.

5.4.2 Eye-tracking paradigm

5.4.2.1 Eye-tracking accuracy

We could confirm a high accuracy of the eye-tracking data. Per participant the individual deviance of gaze from 24 targets was measured. The median angular deviance was 0.93° (first row: 0.78°, second row: 0.98°, third row: 1.02, fourth row: 0.96°). The median distance deviance was 5.66 cm (first row: 3.96 cm, second row: 4.94, third row: 6.60, fourth row: 7.71).

5.4.2.2 Externalization rate

Again, in a vast majority of trials (in %) sounds were perceived externalized for loudspeaker condition (M = 96.3, SD = 10.1), measHATS condition (M = 90.4, SD = 17.8), simIndivHRIRs (M = 90.8, SD = 18.4), and simHATS (M = 90.6, SD = 16.9), whereas Anchor condition stimuli were mainly not perceived externalized (M = 18.9, SD = 22.7). A mixed-effect logit model revealed a significant main effect of Audio Condition, $\chi 2\,(4) = 93.73$, $p < .001$, on externalization. Post-hoc comparisons revealed significant differences of loudspeaker against Anchor ($\beta = -10.79$, $Z = -4.14$, $p < .001$), measHATS against Anchor ($\beta = -11.82$, $Z = -6.17$, $p < .001$), simIndivHRIRs against Anchor ($\beta = -11.28$, $Z = -6.45$, $p < .001$), and simHATS against Anchor ($\beta = -13.0$, $Z = -5.32$, $p < .001$). No significant differences of loudspeaker against all binaural auralizations could be found and no significant differences of comparisons between the different plausible binaural auralizations.

5.4.2.3 Sound source localization accuracy – rate of correct fixations

Trials in which a sound was not externally perceived were not included in this data analysis. In Figure 16 the rate of trials, in which participants' first fixation was towards the correct agent (agent placed at speaker position) is shown respectively for each Audio Condition. Since again lots of trials for the Anchor condition are invalid (M = 61.56, SD = 25.48), the Anchor condition was excluded for inference statistics. A mixed-effect logit model revealed a significant main effect of Audio Condition, $\chi 2\,(3) = 22.77$, $p < .001$, on correct fixation. Post-hoc comparison revealed significant differences of loudspeaker against measHATS ($\beta = 0.49$, $Z = 3.63$, $p = .001$), against simIndivHRIRs ($\beta = 0.65$, $Z = 5.06$, $p < .001$), and against simHATs ($\beta = 0.56$, $Z = 4.27$, $p < .001$). There were no significant differences between the plausible binaural auralizations. For the equivalency hypothesis of simIndivHRIRs and simHATS concerning fixations on correct agents, a Bayes factor of 2.63 could be found (anecdotal evidence for equivalency). A similar degree of evidence could be found for the

equivalency hypothesis of measured and simulated Binaural auralizations (Bayes factor = 1.45).

5.4.2.4 Subjective experience

Social presence

In Figure 9, data on subjective experience during the eye-tracking paradigm of 49 participants are shown. A linear mixed-effect model again revealed a significant main effect of Audio Condition on social presence, $F(4,87.9) = 33.44$, $p < .001$. Participants rated social presence lower for anchor trials ($M = 2.6$, $SD = 2.1$) than for loudspeaker trials ($M = 5.9$, $SD = 2.2$, $\beta = -3.36$, $t = -11.10$, $p < .001$), than for measHATS trials ($M = 5.6$, $SD = 2.3$, $\beta = -2.99$, $t = -8.69$, $p < .001$), simIndivHRIRs trials ($M = 5.8$, $SD = 2.1$, $\beta = -3.21$, $t = -10.07$, $p < .001$) and lower than for simHATS trials ($M = 5.5$, $SD = 2.3$, $\beta = -2.94$, $t = -8.77$, $p < .001$). No other significant differences could be found in any other comparison between Audio Conditions. The Bayes factor for an equivalency of social presence during loudspeaker and plausible binaural auralizations trials is 1.51; for equivalency of measured and simulated binaural auralization trials the Bayes factor is 5.00, and for an equivalency of simIndivHRIRs and simHATS we found a Bayes factor of 0.45.

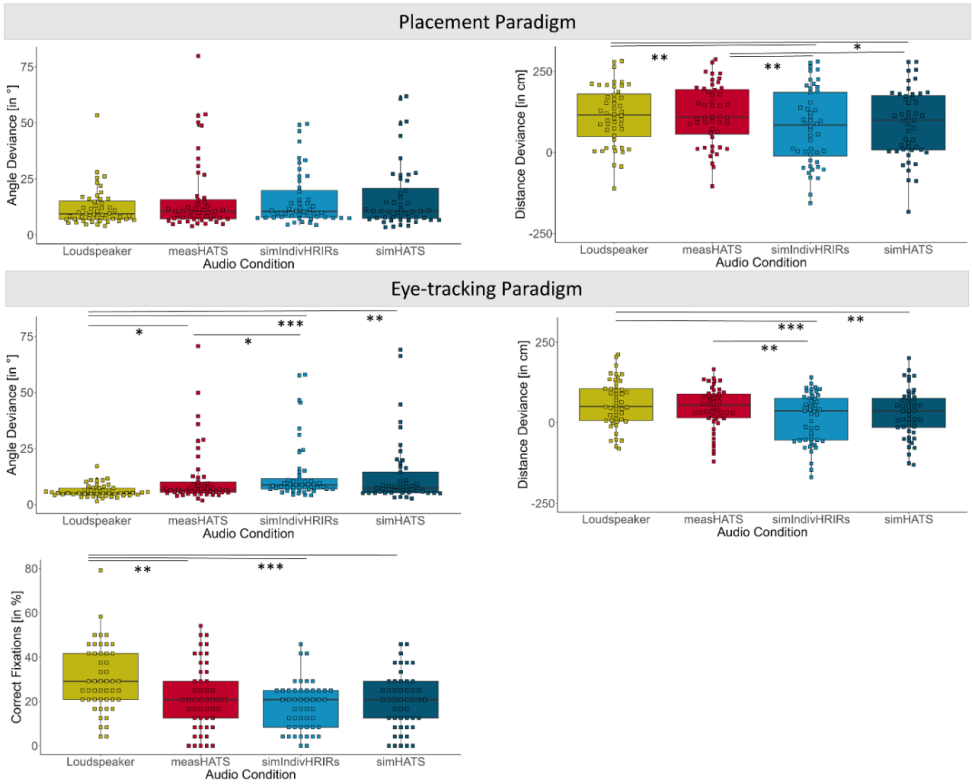

Figure 16: Sound Source Localization Accuracy. On the left: Angle Deviation in Degree between real and estimated sound position as a function of Audio Condition. On the right: Distance Deviance in cm between real and estimated sound position as a function of Audio Condition and separate for each paradigm. For the eye-tracking paradigm, one further figure shows the rate of correct fixations (in %).

Subjective realism

Similar results could be found for the rating of realism. A linear mixed-effect model revealed a significant main effect of Audio Condition on subjective realism, $F(4,71.3) = 29.91$, $p < .001$. Participants rated the subjective realism of the anchor lower ($M = 2.9$, $SD = 2.2$) than the subjective realism of loudspeakers ($M = 5.6$, $SD = 2.2$, $\beta = -2.77$, $t = -10.09$, $p < .001$), of measHATS ($M = 5.4$, $SD = 2.2$, $\beta = -2.55$, $t = -7.86$, $p < .001$), of simIndivHRIRs ($M = 5.6$, $SD = 2.1$, $\beta = -2.77$, $t = -9.77$, $p < .001$) and lower than simHATS ($M = 5.3$, $SD = 2.1$, $\beta = -2.43$, $t = -7.99$, $p < .001$). No other significant differences could be found in any other comparison between Audio Conditions. Following Bayes factors were found for the hypothesis of equivalency of loudspeakers and plausible binaural auralizations: Bayes factor = 3.48, for measured and simulated Binaural auralizations: Bayes factor = 5.93, for simHATS and simIndivHRIRs: Bayes factor = 0.25.

5.4.3 Comparison of paradigms

Figure 17 shows the angle and distance deviance between the estimated and real sound source position as a function of Audio Condition and used sound source localization measurement paradigm. For the calculation of deviances between the estimated distance or angle and the real distance or angle, further processing of eye-tracking data was carried out. An outlier analysis was done within each found fixation and samples were excluded which were more than two standard deviations away from mean distance to participant. We then took the sample in which the gaze hit was the closest to the participants' position for further comparison. We used the same calculation methods as described for the placement paradigm.

We compared the results gained via the two different paradigms to measure sound source localization. The comparison must be interpreted cautiously since the continuous data gained from the eye-tracking paradigm is unequivocal to the placement data. First, externalization rates were numerically higher (4.6 %) in the placement paradigm. Front-back confusions could not be measured in the eye-tracking paradigm, since trials were repeated until no more fixation on the wall occurred (wall behind participants).

Distance deviance seem to be numerically higher in the placement paradigm (about 61 cm higher), as well as angle deviance (about 4.9° more). On an individual participant level, an accordance between sound source localization accuracy in both paradigms was found, almost the same correlations were found for both outcome variables, see 0. Angle localization accuracy in the first paradigm was correlated with angle deviance in the second paradigm, $r(47) = .42$, $p = .003$. Within the loudspeaker condition, a correlation of $r(47) = .51$, $p < .001$

was found, for measHATS a correlation of $r(47) = .37$, $p = .010$, for simIndivHRIRs $r(47) = .37$, $p = .010$, and for simHATS $r(47) = .39$, $p = .006$. Also the accuracy from both paradigms concerning distance estimation was correlated, $r(47) = .42$, $p = .003$. For loudspeaker trials no significant correlation was found, $r(47) = .18$, $p = .229$, whereas for measHATS a correlation of $r(47) = .38$, $p = .008$, for simIndivHRIRs $r(47) = .48$, $p < .001$, and for simHATS a correlation of $r(49) = .43$, $p = .002$, was found.

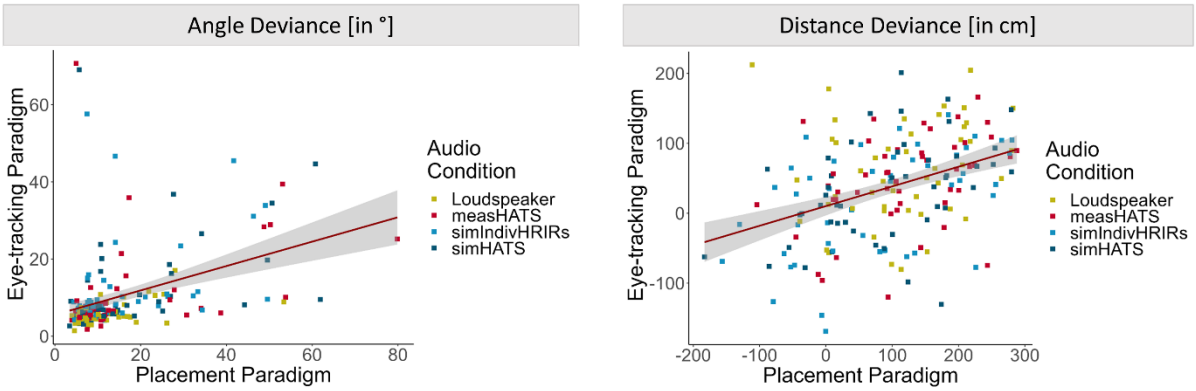

Figure 17: Correlation between placement paradigm (x-axis) and eye-tracking paradigm (y-axis) concerning Sound Source Localization Accuracy. On the left: Angle Deviation in Degree. On the right: Distance Deviance in cm.

5.4.4 Exploratory analyses

5.4.4.1 Social anxiety, negative affect, and egocentric source distance

Participants' level of social anxiety was assessed. We analyzed whether social anxiety influences the perception of sound source distance. We used "social stimuli" – speech – which are fear-related in persons with social anxiety. The background of this analysis is that fear is known to have a significant influence on human perception, and also on distance perception (Gagnon et al., 2013). However, we found no significant correlation between social anxiety, measured as the sum of the SPIN questionnaire, and perceived egocentric distance of sound sources. Furthermore, we analyzed whether positive or negative affect influenced perceived sound source distance, and again found no significant correlations.

5.4.4.2 Social presence and subjective realism

In the placement paradigm, the ratings of a sound concerning social presence and realism were highly correlated, $r(47) = .76$, $p < .001$. For loudspeaker trials, a correlation of $r(47) = .53$ was found, for measHATS trials $r(47) = .79$, for simIndivHRIRs $r(47) = .85$, for simHATS $r(47) = .72$, $p < .001$, and for the anchor trials a correlation of $r(47) = .74$ was found, all $ps < .001$. Also during the eye-tracking paradigm, these two attributes were highly

correlated, $r(47) = 0.87$, $p < .001$. In this task, for loudspeaker trials, a correlation of $r(47) = .74$, for measHATS trials $r(47) = .84$, for simIndivHRIRs trials $r(47) = .88$, for simHATS $r(47) = .83$, and for anchor trials $r(47) = .75$ was found, all ps $< .001$.

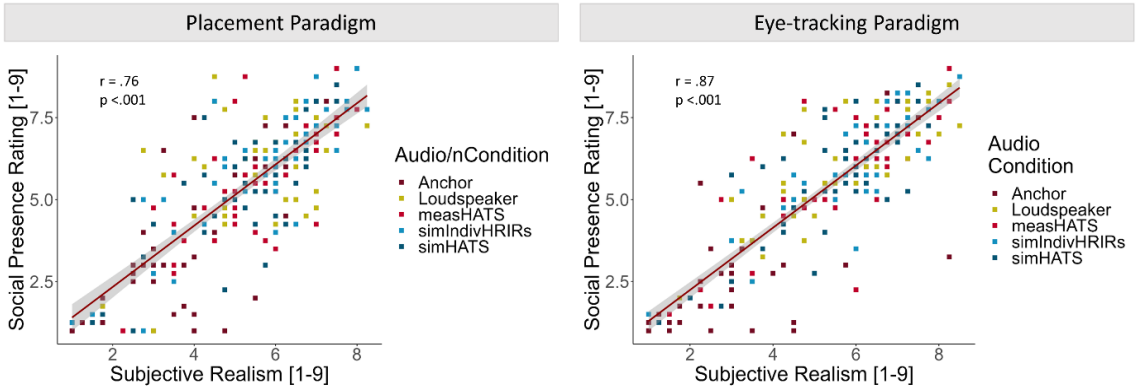

Figure 18: Correlation between social presence and realism rating during placement paradigm (on the left) and eye-tracking paradigm (on the right).

5.4.4.3 *Learning effects*

Referring to prior studies, we exploratory investigated whether the accuracy of sound source localization increases over time (placement paradigm). Again, anchor trials and invalid trials were excluded from analysis. A linear mixed model with trial number as fixed-effect and subject as random effect revealed a significant effect of the number of prior trials on distance deviance, $F(1, 3473.4) = 11.50$, $p < .001$. With an intercept of 108.58 and a $\beta = -0.32$, the distance deviance decreased over time. No effect of prior trials could be found on angle deviance.

5.4.4.4 *Additional descriptive analyses*

In the supplementary material (https://acta-acustica.edpsciences.org/articles/aacus/olm/2024/-01/aacus240025/aacus240025-1-olm.pdf), data can be found on follow-up questionnaires (e.g., difficulty of sound source localization, estimated amount of loudspeaker trials and loudspeakers). Furthermore, data on the elevation and loudness perception of Audio Conditions as well as measurement of sound pressure levels is provided supplementary. Shortly, there are perceptional differences between Audio Conditions, both for elevation and loudness perception. Last, data on hearing experience of the participants and its relationship with sound source localization accuracy can be found in the supplementary material. Shortly, only years of practice playing a musical instrument were significantly negatively correlated with distance deviance.

5.5 Discussion

5.5.1 Placement paradigm

In the placement paradigm, participants positioned an agent at the estimated sound source. Sound source localization accuracy and subjective experience of plausible binaural auralizations were compared to loudspeakers and an anchor condition. In the context of this experiment, for successful localization it is needed that a sound is perceived externalized. Therefore, we first analyzed the rate of externalized trials per Audio Condition. Indicating plausibility of the used binaural auralizations, high rates were found. Furthermore, there was no significant difference between loudspeakers and binaural auralizations in rates of externalization.

In the anchor condition, in over three out of four trials the sound was perceived inside the head (not externalized). This is unexpected since we used a state-of-the-art audio engine (Steam Audio v4.1.4, Valve Corporation, Bellevue, WA, USA). One possible explanation for this finding is the lack of headphone equalization (only) in the anchor condition. We used extra-aural headphones mounted to the HMD. This allowed a direct comparison to the loudspeaker without listeners being aware whether a headphone rendering was played or the real loudspeaker. For plausible binaural auralizations, we adjusted BRIRs regarding the headphone HMD position in terms of individually measured and computed individual headphone equalizations. This equalization could not be provided for the anchor, since the audio plugin is implemented in the gaming engine. Hence, headphone equalization was found to be crucial for e.g. distance estimation, since an absence of equalization resulted in significantly more in-head localizations (Sunder et al., 2014). An alternative explanation for the surprisingly infrequent externalization of the anchor could be due to the experimental setup. We presented high-quality renderings beside the anchor and thus the contrast of the anchor to the simulations based on BRIRs which were precisely tailored toward the real room may have influenced the perception of the anchor. The BRIRs included room simulation (T. Wendt et al., 2014), high quality HRIRs, and source directivity. The reverberation times of all conditions except the anchor were very similar (see Figure 11) underlining again the sharp contrast to the anchor. It was shown that among several BRIR conditions using different late reverberations only for the most extreme modification (when no late reverberation cues were presented), significantly lower externalization rates were found (Schneiderwind et al., 2023). The anchor uses ray tracing, which is normally used only for early reflections which could also explain low externalization rates (Cuevas-Rodríguez et al., 2019). Further experiments on

context and contrast effects on externalization are being planned. However, due to the low rates of externalized trials, the anchor audio condition was excluded from further analyses on sound source localization. To sum up, the anchor differs from the other very similar conditions both in input parameters and on an output level in externalization and subjective experience. No conclusions can be drawn on the low acoustic performance of the anchor since this question was no goal of this study. The results can, nonetheless, be taken as a hint that subjective realism and social presence may be enhanced when spatial audio is perceived externalized.

Another fundamental characteristic of accurate sound source localization is the non-occurrence of front-back confusion. Here, a significant effect of Audio Condition was found. However, only for simHATS stimuli, higher rates of front-back confusion were found in comparison to the loudspeaker condition. For further analyses, only trials in which sounds were perceived externalized and no front-back-confusion occurred, were analyzed. It has to be mentioned that comparably high rates of front-back confusion and also confusion in the real loudspeaker condition were found. A possible explanation for this could be that the stimuli were in parts rather short (on average 1.5 s) and that our participants, naïve to room acoustics, did not systematically use head-movements to localize sound source positions.

We were then interested in whether sound source localization of plausible binaural auralizations and loudspeakers were equivalent. Therefore, we analyzed the angle deviance between the real (or simulated) sound source and the estimated sound source as indicator for azimuthal localization. Concerning azimuthal localization, performance was not significantly affected by Audio Condition. We further gained evidence that all plausible binaural auralizations (measHATS, simIndivHRIRs, and simHATS) are equivalent concerning azimuthal localization via Bayes Factors. Nonetheless, we could not find support for an equivalency of real loudspeakers and binaural auralizations concerning azimuthal localization (Bayes Factor < 3).

We then analyzed whether sound distance perception differed between loudspeakers and plausible binaural auralizations. Unexpectedly, the distance deviance was significantly higher in the loudspeaker condition in comparison to binaural auralizations based on simulated BRIRs (both simIndivHRIRs and simHATS). The same was found for the measHATS condition. No difference between measHATS and loudspeakers was found. Furthermore, we gained evidence for the equivalency of simIndivHRIRs and simHATS concerning distance perception. Regarding the good correspondence in perceived azimuth/direction between different binaural auralizations compared to significant differences

found in the perceived distance, this finding can be explained considering that perceived azimuth direction is dominated by the direct sound field component based on the precedence effect (Blauert, 1997). For distance perception, on the other hand, the source directivity, which is incorporated in the room simulations made for some auralizations, will influence the direct-to-reverberation ratio (DRR), which in turn could influence perceived distance. Indeed, there were variations in the DRR (see Table 1 for frequency-independent DRRs and Figure 2 for energy-decay-curves) between the binaural auralizations. A detailed investigation of the interplay of DRRs and visual virtual scene and the effect on sound distance perception is outstanding. To interpret these results (lower distance accuracy in real sound source condition compared to simulated condition), it is also worth looking at our previous findings on distance perception in VR.

We consistently found evidence for an overestimation of sound sources when using the visual presentation of a room via HMD (Kroczek et al., 2022; Roßkopf et al., 2023; Roßkopf et al., 2024). One explanation could be the visual distance compression in VR. A visual virtual room is perceived as about a fourth smaller than a corresponding real room (Renner et al., 2013). The acoustic perception of a sound source may interact with the compressed visual spatial impression and audiovisual integration may then lead to a distorted perception of distances. Furthermore, our measurement of localization – the placement task – could have contributed to these findings, since numerically higher overestimation was found in the placement paradigm compared to the eye-tracking paradigm. Assuming a visual compression of up to a quarter in VR, participants would have to place the agent much further away to set the desired distance. Based on the assumption that loudspeakers generate the desired sound source distance in the best possible way (and thus represent the baseline), this could mean that the simulated BRIR auralizations generated compressed distances (“too near”), which lead to a better agreement with the visually compressed distances. Furthermore, we found that participants improved the accuracy of distance estimations over time, which was not the case for azimuthal localization. Note, that no feedback was given on potential source positions. Potentially, they learned to rely more on the acoustic cues over the course of time or adapted to the visual compression by using different spatial cues. This should be clarified in future studies.

It should also be noted, that the majority of participants perceived all audio conditions (except the anchor) as elevated (supplementary material). The measHATS BRIRs were perceived as most elevated. These BRIRs were the only audio condition without different vertical orientations, which suggests a possible connection between head tracking and

elevated perception of simulated sound sources. Nonetheless, no detrimental effect of the lack of elevation angles in measHATS were found on their azimuthal localizability.

Last, we examined the influence of the Audio Conditions on subjective experience in VR. Not surprisingly, given the high internalization rates, anchor condition stimuli were rated as lowest concerning realism, but also concerning social presence. Social presence was furtherly rated higher in simIndivHRIRs in comparison to the simHATS condition. MeasHATS was found to be equivalent to simHATS and simIndivHRIRs concerning social presence ratings and subjective realism. Last, we gained evidence that for subjective realism, all three plausible binaural auralizations are equivalent to loudspeakers.

5.5.2 Eye-tracking paradigm

In this task, sound source localization was measured within a comparatively naturalistic paradigm using eye-tracking. We analyzed the rate of trials in which participants looked at the correct virtual agent (agent at sound source position) as a measure of sound source localization accuracy. Comparable to the placement paradigm, we found very low rates of externalization in anchor trials and consequently excluded this Audio Condition from further analyses on sound source localization accuracy.

Comparable to the placement paradigm, our hypotheses that the plausible binaural auralizations are equivalent to loudspeakers concerning sound source localization accuracy could not be confirmed. Higher rates of correct fixations (around 10 % more) were found in loudspeaker trials. We further could not gain evidence for an equivalency of the three plausible binaural auralizations (Bayes factors < 3), which is in contrast to the equivalency between auralizations which we found in the placement paradigm. However, no significant differences were found, indicating a need for further research.

When asking participants about their social presence in VR during the eye-tracking paradigm, an effect of audio condition was found. Participants again reported lower social presence in anchor trials. No other significant differences could be found. However, only for the hypothesis, that measured and simulated binaural auralizations are equivalent, evidence could be found.

Furthermore, differences between Audio Conditions also became apparent in the rating of realism. Again, anchor trials were rated worst. Here, we gained evidence for an equivalency of loudspeakers and all three plausible binaural auralizations concerning subjective realism and for the equivalency of binaural auralizations based on measured and

simulated BRIRs. The equivalency of simulated binaural auralizations based on generic BRIRs and individual BRIRs could neither be confirmed nor disproved.

To conclude, the hypothesis that the plausible binaural auralizations (and loudspeakers) are superior to the anchor could be confirmed on all outcome variables. For sound source localization, loudspeakers were in parts superior to all other conditions. Ambiguous evidence for equality between the plausible binaural auralizations was found. Yet, the binaural auralizations were equivalent to loudspeakers when it came to subjective realism in VR.

5.5.3 Implications

Our results provide evidence that the here investigated plausible binaural auralizations can be used to create an auditory impression that is very similar to that produced by a real sound source. These binaural auralizations include a room simulation using RAZR (Wendt et al., 2014), accurately measured HRIRs, and acoustic source directivity. In terms of externalization, the renderings are not inferior to the loudspeakers. Nonetheless, loudspeakers are still superior when it comes to sound source localization accuracy. However, clear evidence was gained, that the loudspeakers and the three plausible renderings are equivalent in terms of subjective realism underpinning previous work that found authenticity and plausibility of speech auralizations. In this study, the investigated scene incorporated a visual digital twin of the simulated room, the real room itself, in which participants were present, and multiple real sound sources alongside simulated ones. Therefore, the equivalency concerning subjective realism hints towards transfer-plausibility of the used binaural auralizations (Neidhardt et al., 2022; Wirler et al., 2020). Additionally, binaural auralizations are not inferior to loudspeakers in terms of social presence in VR.

One result that requires discussion is the better performance concerning distance estimation for simulated binaural auralizations in comparison to loudspeakers and measured BRIRs. As already briefly described above, one explanation could be that our simulated binaural auralizations generate distance impressions that seem nearer than intended. Potentially, using the measured room impulse responses the impression of distance could be created in a more realistic way than when using the simulated RIRs. In combination with visual distance compression when using an HMD, this could be an explanation for our findings. We used the HTC Vive Pro Eye as HMD. For the HTC Vive, a mean compression rate of 0.6 at a real distance of 5 m was reported (Buck et al., 2021). This is a comparatively high compression rate. In another study, in which the plausible binaural auralizations were

investigated and in which a room was also visually presented via HTC Vive, potentially similar evidence was found (Stärz et al., 2024). This implicates potential systematic influences of visual cues on auditory distance perception in VR (referring to ventriloquist effects (Jack & Thurlow, 1973). Participants reported higher perceived reproduction quality for simulated binaural auralizations (in comparison to loudspeakers and measHATS). The source distance of loudspeakers and measHATS was rated as too distant. This is in line with our findings. Even more interestingly, the source distance for the simulated binaural auralizations was perceived on average as more precise than loudspeaker and measHATS. In contrast, Blau et al. (2021) found no difference between measHATS and simulated binaural auralizations in terms of source distance or overall quality. While both studies used comparable BRIR sets, in the 2021 study, no visual virtual reality was presented. Here, no effect of visual compression should have occurred. To add, in the earlier study, participants should compare the binaural auralizations to a non-hidden loudspeaker reference. Furthermore, it is unknown whether participants rated the source distance of binaural auralizations as suboptimal because it is perceived as too close or too far. While the prior studies and this study used comparable binaural auralizations, measurement techniques, comparison conditions, and visual setup differed. More research is needed on audiovisual integration in VR and the influence of context and measurement techniques on source localization.

It should be noted that there is already research on ways to compensate for visual distance compression in VR. One approach used a similar mechanism, namely that audiovisual integration influences distance perception in VR, but in the opposite direction (manipulation of auditive distance to influence visual distance perception). In a recent study (Huang et al., 2021), the reverberation time of an auditory stimulus was manipulated to alter depth perception. Especially in the near field (up to 5 m), depth perception could be influenced by longer reverberation times. When very long reverberation times were used, participants put more trust in visual information and sensory segregation occurred. Compensation was also attempted by visual manipulations. For example, a kind of geometric minification (Li et al., 2015) or the reduction of eye height (Leyrer et al., 2015) could decrease (visual) distance underestimation. An interesting and open research approach would be the investigation of the plausible binaural auralizations in comparison to loudspeakers in a visually non-compressed (or compensated) virtual environment.

Besides the question of how accurate audio-rendered sound sources can be localized in comparison to loudspeakers, we investigated whether the three binaural auralizations were

equivalent. The measurement of BRIRs in a real room is time-consuming since many different head-above-torso orientations have to be adjusted and extensive acoustical technical equipment is needed. A further disadvantage is the limitation to existing rooms. Rendering of BRIRs of a not yet or not anymore existing room is not possible with this method. Auditory simulations of rooms based on simulated BRIRs can overcome these drawbacks. Furthermore, these simulated BRIRs can quickly be adjusted to new requirements (for example change of source positions). We gained evidence, that regarding subjective experience in VR, binaural auralizations based on BRIRS simulated using RAZR (Wendt et al., 2014) are equivalent to measured BRIRs. Furthermore, in the placement paradigm, equivalency could be shown for azimuthal localization. However, distances were estimated significantly worse in terms of overestimation when measured BRIRs were used. Since comparable effects were found for loudspeakers, the same underlying mechanisms (see above) can be assumed.

In the eye-tracking paradigm, no significant differences between measured and simulated BRIRs audio conditions were found. However, the evidence for equivalency was small, therefore we could not confirm the equivalency hypothesis. A further concern of this study was investigating how close binaural auralizations based on simulated BRIRs in combination with generic HRIRs come to the quality of simulated BRIRs which are calculated using individually measured HRIRs. While within the placement paradigm, we gained clear evidence for an equivalency of simHATS and simIndivHRIRs concerning localization accuracy, within the eye-tracking paradigm, only small evidence was gained for the equivalency hypothesis. However, no significant differences were found between simHATS and simIndivHRIRs in any of the localization variables. Looking at both, these results and the effort of individually measured HRIRs, a valid assumption is that for many areas of application, especially when using speech stimuli, the cost-benefit analysis is in favor of using generic HRIRs.

As briefly stated above, an area of application is creating a convincing audiovisual scene for multiuser interaction or virtual exposure therapy. Here, accurate source localizability can contribute to higher degrees of realism of the virtual scene. Another important goal besides high realism is a high degree of (social) presence (Slater, 2018). Concerning the feeling of being with another person present in virtual reality (social presence), higher levels can be reached using plausible auralizations and loudspeakers in comparison to anchor stimuli. In the placement paradigm, the simIndivHRIRs stimuli were rated significantly higher concerning social presence than simHATS. In the eye-tracking paradigm, there is no evidence for the equivalency of simHATS and simIndivHRIRs. This can be seen as an

indicator that using individually simulated BRIRs social presence in VR can be more enhanced than when using generic BRIRs (simHATS). Subjective rating of realism also differed between Audio Conditions. Here again, the anchor was inferior to all other Audio Conditions. Interestingly, social presence and subjective realism were strongly correlated. The correlation remained robust even when excluding the anchor condition (which prominently differed from the other conditions in terms of rating scores). This indicates that with a subjectively higher quality of the binaural auralization in a VR scene, higher levels of social presence can be induced (or vice versa). However, it is also possible that participants referred to a related construct, although the two rating items were formulated very differently. A comprehensive investigation of the interplay of subjective realism and social presence is pending, here a broader variety of items as well as inverted items should be implemented. It has to be pointed out, that no further validation of our rating items was conducted to investigate the internal reference to which participants of the main study referred to (only for the pilot study). Therefore, also implicit measures for social presence and subjective realism should be used and psychophysiological data could be included to gain a more robust evidence on this question.

We used two different paradigms to measure sound source localization accuracy of different Audio Conditions in VR. As can be seen in Figure 8, similar patterns for effects of Audio Conditions can be found independently of the measurement paradigm. This indicates towards validity of both paradigms. All in all, both, deviances in distance and angle estimation seem to be in trend higher in the placement paradigm. Besides the “device” for measurement (controller vs. gaze), the difference between both paradigms is the (non-) existence of predefined offered positions. In the eye-tracking paradigm, 16 different plausible source positions (16 agents in the room) were offered. Contrarily, in the placement paradigm, source positions could be estimated and placed anywhere (on the xy plane). We assume the placement paradigm to be more sensitive to finding differences. The advantage of the eye-tracking paradigm lies in the assumed higher external validity, e.g. when the goal is the assessment of audiovisual quality of a virtual classroom. Since visual feedback on the adjusted source position was only provided in the placement task, higher accuracy can be assumed for the placement task in comparison to eye-tracking. In comparison to traditional pointing methods transferred to a VR setting (e.g. using controllers, Ahrens et al., 2019; Huisman et al., 2021), by tendency lower levels of azimuthal accuracy were found either for the placement and the eye-tracking paradigm. This may be more due to methodological reasons (sample, outlier exclusion) than acoustic reasons (also for real loudspeaker).

A further important difference of the here used placement paradigm in comparison to traditional pointing methods transferred to a VR setting is that a visual virtual room but no possible source positions were displayed, which could also have an effect on azimuthal accuracy. To sum up, both paradigms have drawbacks and merits, the choice should also depend on the intended context. Nonetheless, with both paradigms, we only measured sound source localization in terms of distance and angle estimation. Deviances in elevation perception were not assessed within the paradigms. An additional task on elevation perception revealed in part significant differences between Audio Conditions (see supplementary material). This discrepancy must be taken into account. What else needs to be considered, is that we found differences in perceived loudness of the different Audio Conditions. Since loudness is an important factor for distance perception (Zahorik et al., 2005), this may have influenced sound source localization. However, only slightly different sound pressure levels were measured (see supplementary material). The difference in loudness perception is possibly due to different presentations (via loudspeakers and headphones).

The contribution of a possibly higher measurement error when using gaze behavior instead of placing an agent is not yet clear. It is also noteworthy, that our sample consisted of non-experts in acoustics. Participants were not consistently able to perceive how many different sound sources had been placed in the room, the estimations ranged broadly (see supplementary material). Furthermore, subjective assessment of own hearing capabilities was not relied to localization accuracy. Only for years of experience playing a musical instrument, a positive effect was found. To add, participants seem to progress during the tasks. The more trials they completed, the more accurate was localization. This is in line with previous studies on sound localization in three-dimensional virtual environments (Roßkopf, Kroczek, Stärz, Blau, Van de Par, et al., 2023). Not only that sound localization improve with duration of training (Rajguru et al., 2022), but localization training in VR can also compensate for the shortcomings of generic HRIRs in comparison to individual HRIRs (e.g. Steadman et al., 2019). This training effect is even more pronounced when head movements and therefore more information on head-tracked binaural auralizations are targeted. In the eye-tracking paradigm, in which 16 different plausible source positions were offered, less than a third of the fixations were on the correct position (in the best Audio Condition).

5.6 Conclusion

To conclude, with two different sound source localization paradigms we could show, that the three plausible binaural auralizations allow comparable accuracy. Nonetheless, localization

accuracy was not as high as for real sound sources. Concerning externalization rate, social presence and subjective realism, the plausible binaural auralizations are comparable to loudspeakers. The anchor condition (state-of-the-art 3D audio for the Unreal gaming engine) was inferior in all investigated aspects. Of particular interest is the correlation between social presence and subjective realism. Overall, the findings suggest that advanced binaural auralizations may improve emotional processing by increased presence levels that facilitate more seamless social interactions (Pfaller et al., 2021). Research in social or socioemotional contexts or in VR-based psychotherapy can benefit from this. Taking the cost-benefit ratio into account, binaural auralizations based on simulated BRIRS using RAZR (Wendt et al., 2014) in combination with generic HRIRs (simHATS) are considered the most recommendable of the here investigated three plausible binaural auralizations (measHATS, simIndivHRIRs, simHATS) for typical research in the clinical-psychological or other applied psychological fields, as long as the dominant auditory stimulus type is human speech.

6 Study 3: Hey AI: Can you trigger me? On the equivalency of Text-to-Speech synthesis and human speech in a virtual social stress paradigm

Sarah Roßkopf, Leon O.H. Kroczek, Theresa F. Wechsler, Felix Stärz, Matthias Blau, Steven van de Par, and Andreas Mühlberger

At the time the dissertation was published, Study 3 had still been under review (since 22nd November 2024) in a psychological journal. The official citation that should be used in referencing this material is: Roßkopf, S., Kroczek, L. O. H., Wechsler, T. F., Stärz, F., Blau, M., Van de Par, S., & Mühlberger, A. (2024). Hey AI: Can you trigger me? On the equivalency of Text-to-Speech synthesis and human speech in a virtual social stress paradigm [Manuscript submitted for publication]. Fakultät für Humanwissenschaften, Lehrstuhl für Klinische Psychologie und Psychotherapie, Universität Regensburg.

6.1 Abstract

While the use of text-to-speech synthesis (TTS) is high, evidence of its effects on presence in VR, for example, is scarce. It also remains unclear, whether virtual interactions using TTS can provoke psychosocial stress and evaluative threat as required for psychological paradigms like the Trier Social Stress Test (TSST) as the experimental gold standard to examine human stress reaction. Using TTS instead of prerecorded human speech (PHS) in the virtual version (VR-TSST) can be beneficial, e.g. by creating greater freedom and flexibility to react. In a randomized-controlled trial, we compared TTS stimuli to PHS within a VR-TSST. Participants' heart rate, trapezius muscle activity, and self-reports on stress, presence, and affective state were surveyed. In both speech conditions, profound stress reactions were provoked and no differences between audio versions were found for any variable. This indicates TTSs' comparability to PHS in socio-emotional experiences underlining its high utility in virtual social interactions.

6.2 Introduction

Text-to-Speech synthesis is widely used in various areas of applications such as e-learning (e.g. Herawati et al., 2022; Minder et al., 2012; Mishev et al., 2022) interactive interfaces e.g. in customer support (Qiu & Benbasat, 2005b, 2005a), and of course the most traditional use case – navigation (e.g. TomTom, GoogleMaps). Synthetic voices including voices generated with the help of artificial intelligence (AI) are also commonly heard when moving around in a

"smart home" or using a smartphone [e.g. Siri (Apple), Alexa (Amazon)]. Human voices do not only transport verbal information but are also critical carriers of emotion and personal identity (for a review: Belin et al., 2011). Similar to the processing of faces, humans are highly trained in recognizing familiar voices. The irritation that arises when a new speaker dubs a series character demonstrates the great influence of the voice on interpersonal interaction can be estimated well. Recent technological advances succeeded in creating "voices" so convincing that human listeners and even machines (software systems) can be fooled (Wenger et al., 2021). Synthetic speech can be simulated in such a way that it corresponds to specific needs (e.g. age, gender) or that it sounds like a specific person.

While the danger of identity theft and betrayal might not be neglected, TTS synthesis offers the opportunity to improve the life of people with e.g. visual impairments (Ani et al., 2017) and aid the education process of impaired or neurodivergent students (Herawati et al., 2022; Mishev et al., 2022). Also, for applications and research conducted in virtual reality (VR), the use of TTS yields advantages. The speech of virtual agents or spoken instructions can easily be adjusted in comparison to when human voice recordings are used. In contrast to the reactions of human experimental instructors, TTS reactions can also be automated and can be seen as more standardized. The interference variable "investigator" can therefore be reduced using TTS synthesis. Additionally, TTS might fit better with virtual agents than a prerecorded – often imperfect - human voice. Human perception is in clear favor of congruence over incongruence (e.g. auditory Stroop effect, Green & Barber, 1981). The perceptual processing flow is disturbed by incongruencies. Novel approaches suggest congruence and plausibility as the most important quale for an high qualitative experience in mixed reality (XR) (Latoschik & Wienrich, 2022). When expectations about cues are fulfilled, congruence can be assumed and thus, VR/XR users should experience higher plausibility of the virtual scene. One might assume that an evenly sounding TTS fits better to the appearance of virtual agents, which are often rendered symmetrically and smoothly (e.g. VALID see Do et al., 2023, or MakeHuman). If then, the virtual agent "speaks" with an accent or a syntactic peculiarity, which can be found especially for non-professional human speakers, a break in expectation might occur and may lower the quality of VR experience. Support for this assumption is given by a study which examined the effect of human vs. robot voice when a human vs. robot face was presented. Eeriness was highest in incongruent conditions (Mitchell et al., 2011). Such a break in expectation could also effect the uncanny valley phenomenon – a strong negative impression of human-like objects (Zhang et al., 2020). Since difficulties in categorization seem to contribute to the uncanny valley (Yamada et al., 2013), the

combination of (imperfect) human voices with visually rendered faces may even induce discomfort.

Creating highly standardized and at the same time flexible virtual interactions is particularly beneficial in the application field of clinical psychology research in VR. Virtual exposure therapies have been shown to be very effective and non-inferior to exposure therapies conducted in vivo (Wechsler et al., 2019). Only for treatment of social phobia, in-vivo therapy seemed to be slightly more effective, indicating a further need of improving virtual social interactions. Here, social presence - the sense of being together with another (Oh et al., 2018) - is an highly relevant construct. Since social anxious persons are afraid of being viewed critically by other people or of being rejected by them (American Psychiatric Association, 2013), the feeling of being together with another (social presence) can be assumed as a prerequisite for the induction of fear in an exposure scenario for social anxiety. Social presence was not only found to be essential for overcoming fear, it can also support well-being. For a non-clinical population (young South Koreans), it was found that social presence on massive-social virtual platforms (metaverse platforms) could help to reduce feelings of loneliness in everyday life (Oh et al., 2023). To sum up, for treatment of patients as well as for preventive purposes, research how social presence induction can be enhanced should be targeted.

The question arises to which extent the use of TTS effects experiences in VR. Virtual social interactions for which an agent is displayed and for which verbal content is used was shown to be effective in inducing social fear (Reichenberger et al., 2017) and for inducing a profound psychophysiological stress reaction (Liu & Zhang, 2020; Shiban et al., 2016; Zimmer et al., 2019). For all these studies, prerecorded human voices were used for the verbal interactions while the visual scene (appearance, lip movements) was rendered synthetically. To the best of our knowledge, no study before investigated the effect of TTS use on variables such as social fear, and stress reaction. Though, it was previously shown that the use of TTS increased perceived flow of live help in electronic commerce but did not affect social presence. Furthermore, the way a TTS algorithm is trained influenced the way in which participants reacted towards it. A TTS model trained with dialogues evoked more spontaneous listeners reactions (such as nods or “uh-huh”s) than a text trained model, indicating a more naturalistic interaction feeling (Misu et al., 2011). Furthermore, participants preferred chatting with more dynamic modelled TTS-bots which include naturalistic cues such as fillers (“umm”) or interjections (“wow”) (Cohn et al., 2019). In addition, participants preferred interacting with TTS voices which are modelled in a way that they match their own

personality (extraversion vs. introversion) (Nass & Lee, 2001). When users had to cooperate online, the partner was not rated more or less trustworthy or likeable when a human voice was used in comparison to a TTS (Jensen et al., 2000). For virtual agents supporting students with stress management in a non-immersive environment, a human voice was preferred over TTS (Abdulrahman & Richards, 2022). Nonetheless, students benefited equally from the aid of both versions. Also, co-presence and trustworthiness did not differ significantly between human and synthetic voice.

A study on the use of TTS and virtual agents in primary school education could find that while the children preferred human voices (in the voice only condition), the integration of a VR agent enhanced the perception of TTS synthesis. A virtual teacher with a human voice was not favored over a one with an advanced TTS "voice". On the other hand, the visual display of a VR agent lowered recall scores (memory task) for all voice conditions and especially for first-time VR users (Dai et al., 2024). These findings underline the assumption that when both modalities, vision and hearing, are rendered consistently (virtual), an agreeable and natural virtual scene /interaction can be created.

To sum up, while the interaction with virtual agents using TTS seems to follow similar rules as human-to-human interactions (e.g. Cohn et al., 2019; Nass & Lee, 2001) and while users benefit equally well from the aid of TTS-based agents (e.g. Abdulrahman & Richards, 2022; Misu et al., 2011), participants seem to still prefer human voices (e.g. Abdulrahman & Richards, 2022; Dai et al., 2024). The question arises whether we can find a difference between human voices and TTS in cases where complex reactions such as social evaluative threat and social stress are crucial. Are virtual agents with synthetic voices able to trigger fear of being viewed critically or fear of being perceived as embarrassing as needed for virtual exposure therapy of social anxiety disorders - or does a human voice bring that extra spark of natural human interaction that is necessary to set these processes in motion? The experimental gold standard to examine human stress reaction, especially on an endocrinological level (cortisol increase), is the Trier Social Stress Test (TSST, Kirschbaum et al., 2008). This laboratory paradigm includes demanding tasks in a mock job-interview in front of a committee. Not only psychobiological stress reactions but also social evaluative threat and feelings of uncontrollability can be induced via the TSST (Frisch et al., 2015). Also, virtual versions of the TSST were found to be effective (Zimmer et al., 2019).

In this study, we investigated in a VR version of the Trier Social Stress Test (VR-TSST) whether the use of advanced TTS stimuli using AI technology has comparable effects on stress, presence, and anxiety to the use of audio recordings of human voices. Furthermore,

we investigated the influence of the individuals' social anxiety level on the stress reaction and the relationship between stress and presence in VR. We expect that the virtual stress scenario evokes a profound psychological and physiological stress reaction in both used audio versions. The heartrate, trapezius muscle tension and subjective stress evaluation were supposed to increase from the baseline measurements to during or after stress induction measurement (manipulation check). We hypothesize no difference between audio versions, first concerning the stress reaction and second, concerning presence in VR. Equivalency tests were planned to obtain evidence on the equal usability of different audio versions in terms of the clinical psychologically relevant attributes subjective stress, physiological stress reaction and presence. Last, individuals, which generally tend to a higher social fear reactivity are supposed to show stronger stress reactions towards the VR scenario.

6.3 Methods

6.3.1 Sample

The sample ($N = 40$) consisted of 30 female and 10 male participants aged between 18 and 55 years ($M = 22.7$, $SD = 7.5$). Participants were excluded if self-reporting previous participation in a (virtual) stress test, neurological or central nervous system disorder, or pregnancy. Furthermore, exclusion criteria were the presence of acute depressive or manic episodes or suicidality assessed via a brief structured diagnostic interview (Sheehan et al., 1998) by a psychologist prior to the experiment. Participants were randomly assigned to the experimental conditions (TTS vs. PHS) stratified for gender. Randomization lists were generated using R (Development Core, 2019).

Paired *t*-tests and *chi-square* tests were conducted to test for systematic a-priori differences between the experimental groups. All participants gave written informed consent. The study was in line with the Declaration of Helsinki and approved by the local ethics committee (University of Regensburg, 22.04.2020, 20-1804-101). The privacy rights have been observed and written informed consent was given.

6.3.2 Materials

The social stress paradigm is based on the original version of the Trier Social Stress Test (TSST) of Kirschbaum, Pirke, and Hellhammer (2008) and the VR variant (Shiban, Diemer, et al., 2016), which includes three virtual agents (mixed-gender) as committee and one male agent as instructor. Under inspection of the agents which are dressed like managers and react in a very neutral way and a camera plus microphone "for further behavioral analyses"

participants have to perform as well as they can during a job interview scenario including a talk and an arithmetic task. The TSST was found to elicit a robust and reliable psychological, physiological and neuroendocrinological stress reaction, also in a virtual version (Q. Liu & Zhang, 2020; Zimmer et al., 2019). Psychosocial stress is induced by demanding tasks and especially high social evaluative threats. The evaluative threat is induced by emphasizing competitiveness and personal evaluation (Frisch et al., 2015). Furthermore, the TSST is suggested to be stressful due to the uncontrollability and the deliberate lack of feedback from the committee. The virtual agents (see Figure 19) are synchronized either with PHS or TTS. The instructor explains the goal of the interview (e.g. to perform as good as possible) and the procedure. The agents of committee specify the tasks and give short, neutral feedback (such as "Please calculate faster."). Further speech was included in the scenario via rating items (stress, presence). All speech parts were determined from the outset.

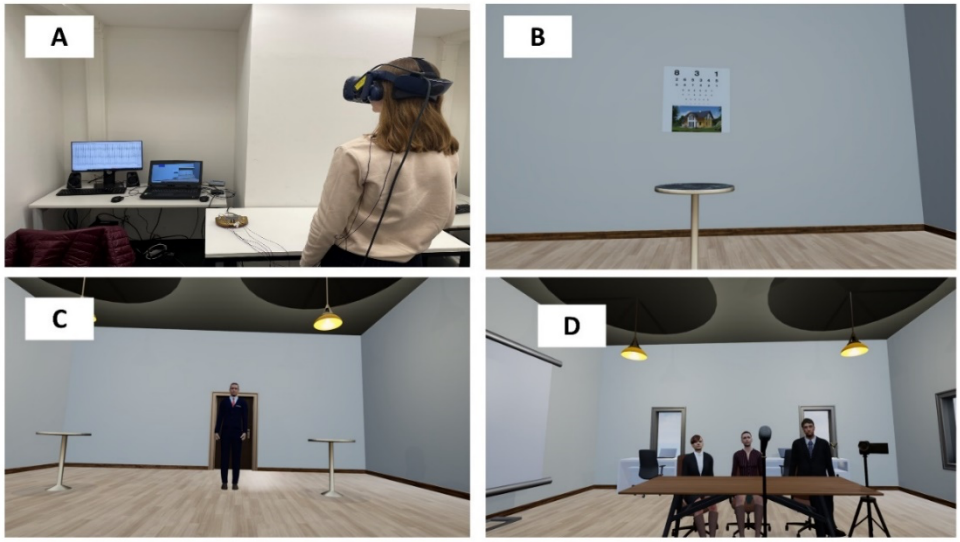

Figure 19: Experimental setup. A: Participant is standing and wearing a Head-Mounted-Display (HTC Vive pro Eye), the motion controller is used with the right hand. EMG and ECG electrodes are attached to the participant and controlled via Brain Vision Software; B – D: Virtual scene from participants' view, B: Baseline room, C: Instructor, D: Virtual Committee.

Overall, 341.9 s of TTS speech were created. The length and speed have been adapted to the PHS. The TTS stimuli were synthesized using the AI voice generator Murf (murf.ai, 2022, Utah, USA). Murf claims to provide high-quality natural-sounding AI voices in over 20 languages that can easily be customized. For the TTS audio content, the German voices of "Murf.ai" (one female, two male) were used. The PHS were dry recordings of three persons (one female, two male) with professional speaker backgrounds. All generated or recorded audio data were processed using Audacity (v 2.3.2, iWeb Media Ltd, Birkirka) to ensure similar length and quality (mono, 44100 H, 32-bit-floating point). The stimuli of both audio conditions were presented with normalized output levels, the amplitudes were RMS-

normalized to -24 dB (full scale). The speech animations of the virtual agents were manually adjusted to fit both audio conditions. There were 28 distinct speech sections ranging from one to eight sentences.

The virtual environment was created using the Unreal game engine (v 4.26, Epic Inc), Autodesk 3ds Max, MakeHuman (v1.2), and Blender (v2.79). The presentation of the VR was controlled using the Unreal game engine and SteamVR (Valve Corporation). The virtual environment was presented via a head mounted display (HMD; Vive Pro Eye, HTC).

Physiological data was recorded using Brain Vision Recorder (Version 1.20, Brain Products GmbH, Gilching). For heart rate data, three electrodes were attached to participants' sternum (ground), left and right lower costal arch. For muscle tension data, two further electrodes were applied to the left Musculus trapezius. Physiological data were amplified using a V-AMP 16 and analyzed using the Brain Vision Analyzer (V 2.2.2, Brain Products GmbH, Gilching, Germany). The Brain Vision Analyzer was also used for preprocessing of physiological data. A 50 Hz notch filter was applied to all data. For electrocardiogram (ECG) data furthermore a 0.159 Hz low cut-off filter with a time constant of 0.1 s and a 30 Hz high cut-off filter; for electromyography (EMG) data, a 499 Hz high cut-off filter, a 28 Hz low cut-off filter were applied. Data was segmented (see Figure 20) to pre, during and after stress parts via the Unreal Script. For heartrate, a custom algorithm was applied which uses the inspection of all r-waves within a segment to compute the mean beats per minute (bpm) per segment. For EMG, the moving average in µV over the segment was used.

Several questionnaires and interviews were used to assess psychological variables and subjective data. Three parts of the Mini international Neuropsychiatric interview (Sheehan et al., 1998) were used to exclude persons suffering from acute episodes of affective disorders or suicidality. A demographic questionnaire assessed general data on participants. For assessment of social anxiety, the social phobia inventory (SPIN) (Sosic et al., 2008), for general anxiety, the state-trait-anxiety inventory (STAI) was used (Spielberger et al., 1971). Before and after the VR scenario, mood was assessed using the positive-and-negative-effect-schedule (PANAS) (Krohne et al., 1996). After the VR scenario, physical and social presence during VR were assessed using the multimodal presence scale – MPS (Makransky et al., 2017). During the VR scenario subjective stress level as well as physical presence were assessed using verbal ratings ("How stressed did you feel [e.g. during the talk] on a scale from 0 – not at all stressed to 100 – maximally stressed?"; "How present do you feel in the virtual environment on a scale from 0 – not at all present to 100 – maximally present?" – translated from German).

6.3.3 Procedure

Figure 20 provides an overview of the experimental flow and the time points of assessments. After giving written informed consent, the corresponding parts of the clinical interview (see above) were carried out. If no exclusion criteria were met, the first (pre) assessment of subjective units of distress, furtherly called stress rating was done. Therefore, the experimenter asked the participants "How stressed do you feel at the moment on a scale from 0 – not at all stressed – to 100 – maximally stressed?" The verbal answer of the participants was registered. Then, participants filled in the questionnaires on demographic data, anxiety, and mood. Afterwards, the electrodes for assessment of physiological data were attached and impedances were kept below 50 kOhm. After asking for a second (pre-VR) stress rating, participants were introduced to the motion controller, the HMD (HTC Vive Pro Eye), and consequently to the virtual environment. Also, randomization was accomplished automated by starting the VR part. Participants entered a virtual room equipped with a high table, lamps, a door, and a poster with series of letters in descending size. When good vision within the HMD was confirmed by reading out aloud the letters, the pre-TSST stress rating was done.

Unlike before, the questions and all further speech of agents were not spoken by the experimenter, but via audio playback (TTS vs. human depending on group assignment). From the beginning of the VR scenario on, the acoustic VR scene differed between the experimental groups. Additionally to subjective stress, presence was assessed by the question "How present do you feel within the virtual environment on a scale from 0 – not at all present to 100 – maximally present?". Afterwards, a virtual agent wearing a black suit entered the room through a door and approached the participant. The virtual agent instructed participants that they would subsequently have to give a talk in front of a committee to convince the managers of the participant's suitability for their dream job. The participants were instructed to emphasize personal strengths and weaknesses and were informed about a second task that would be conducted after the job talk, but the exact task was not specified. Finally, participants were informed that their voice, behavior, and answers would be recorded throughout the job interview and that these recordings would be analyzed by an expert for behavioral analyses. Participants were then instructed to prepare their talk for three minutes (preparation). The virtual agent then left the room and after three minutes of preparation time, reentered. The agent guided the participant towards the interview room, while participants moved through the VR by using the motion controller. The participants were then positioned to stand at a defined position in front of the committee consisting of one female and two male agents in suits ("managers"). The virtual room was equipped like a conference room.

Additionally, a camera and a microphone were directed toward the participants, emphasizing social evaluation during the task. The female agent opened the interview instructing the participant to begin with the talk. The next five minutes, the participants had to present themselves while the agents did not react, except when the participants remained silent for at least 20 seconds, or only talked about professional expertise, etc. In that case, the agents gave short verbal instructions. The short instructions were controlled by the experimental conductor following predefined guidelines. After the talk, again subjective stress and presence had to be rated (talk). For the arithmetic task, participants were instructed to count backwards from 2023 in steps of 17. After a miscalculation they were interrupted by one of the agents and asked to start again from the beginning. If they calculated correctly nine steps in a row, they were asked to calculate faster or to speak more clearly. Following the arithmetic task, again subjective stress and presence (arithmetic) had to be rated and the participants left VR by taking off the HMD. The post-assessment questionnaires were conducted, starting with the MPS, the state anxiety, and the mood questionnaire, and ending with a general questionnaire on hypotheses, adverse effects, and open feedback on the experiment.

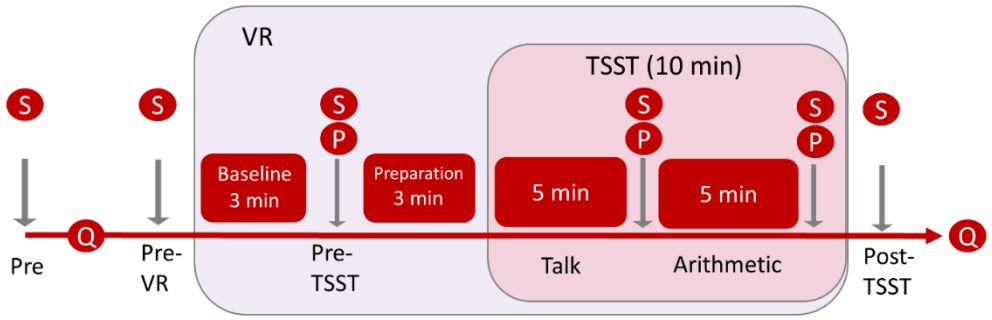

Figure 20: Experimental flow and time points of measurements. Letters in circles represent subjective assessments, S: Stress Rating, Q: Questionnaires, P: Presence Rating. Below the line the labels for time points of measurements are given. The rounded rectangles represent further labels of time points of physiological data (ECG, EMG), time spent in VR (violet) and time in stress test (light red).

6.3.4 Statistical analyses

Possible a priori differences between the experimental groups were analyzed using t-tests and chi-squared tests. To test whether stress was induced via the VR-TSST, repeated measures ANOVA were conducted comparing stress ratings, heartrate, and trapezius EMG at different time points (manipulation check). Audio version was included as between-subject factor to rule out the possibility of a main effect of audio condition and of an interaction effect of time point and audio version. To gain further evidence for the equivalency of the audio versions, two one-sided *t*-tests (TOSTs) were computed. For these tests, boundaries of equivalency must be defined a priori. When the confidence interval of found differences between two

groups lies within these boundaries, one can state that a possible effect is below the smallest effect size of interest. Therefore, regions of practical equivalence were defined as follows: for rating of stress and presence, a difference of less than 10 points (scales are from 0 – 100) was defined as practically equivalent; for heart rate and trapezius EMG, a difference less than one standard deviation; and for the MPS, a difference less than 5 points (scale is from 0 – 50). These boundaries were chosen as a trade-off between findable effect size and practical considerations. To test the effect of social anxiety on stress reaction, linear models were used. Additionally, Bayes factors of independent *t*-tests using non-informed prior distributions were computed when neither significant differences nor equivalency could be shown using the R package BayesFactor (Morey & Rouder, 2014) (not preregistered). Evidence for the null-hypotheses was provided by dividing 1 by the Bayes factor of independent *t*-tests.

Statistical analyses were conducted within the R environment (Development Core, 2019). Alpha level was set to 5%. Extreme outliers (+/- 3*SD) were excluded. Only for ECG data, extreme outliers were found and excluded ($N = 3$). Where needed, sphericity violations were corrected with the Greenhouse-Geisser method. All analyses were preregistered (https://osf.io/v5nby). Data and analysis scripts are publicly available at (https://osf.io/zv6wn).

6.4 Results

6.4.1 Baseline assessment

No significant a-priori differences between the TTS and the PHS group were found concerning age, gender, baseline heartrate, trapezius EMG, stress rating, or any of the anxiety and mood variables, all *p*s > .05. Therefore, randomization did work out and no a-priori differences between groups should account for potential differences in dependent variables.

6.4.2 Stress reaction

Figure 21 provides an overview of the physiological and psychological stress reaction in terms of an increase in heartrate and reported stress level. It displays these two dependent variables at different time points and separately for each audio version.

Repeated-measures ANOVAs revealed a significant main effect of time point on heart rate, $F(2.09, 79.59) = 49.43$, $p < .001$, $\eta_p^2 = 0.57$ and stress rating (SUDs), $F(2.51, 95.4) = 71.32$, $p < .001$, $\eta_p^2 = 0.65$, over both groups. Post-hoc comparisons showed that the heart rate increased from baseline ($M = 95.04$, $SD = 17.68$) to TSST measurements (mean

heartrate during talk and arithmetic task: M = 109.80, SD = 20.35), $t(39)$ = -8.21, p < .001, d = -1.30. Also subjective stress increased from baseline assessment (M = 27.2, SD = 22.87) to TSST (mean stress rating after talk and arithmetic task: M = 61.94, SD = 21.27), $t(39)$ = -9.49, p < .001, d = -1.50. Although for trapezius muscle tension measured via EMG, a significant main effect of time point was found, $F(2.85, 99.62) = 3.52$, p = .019, η_p^2 = 0.09, the direction was not as predicted. Contrarily to our expectations, the muscle tension did not increase from baseline (M = 15.75, SD = 12.02) to TSST assessments (mean heartrate during talk and arithmetic task: M = 13.58, SD = 12.88); $t(39)$ = 1.47, p = .151, but declined. Therefore, EMG data will not be analyzed further, since they do not reflect a stress reaction in this study.

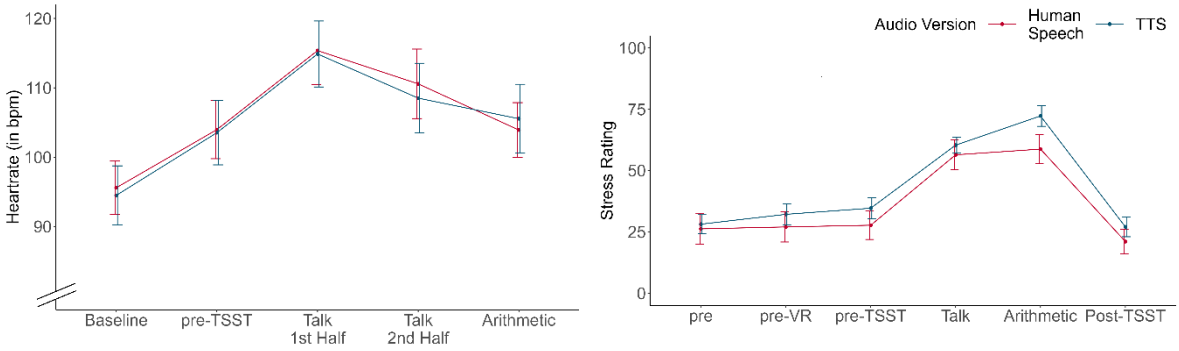

Figure 21: Mean heartrate in beats per minute (bpm), Stress Rating (from 0 – 100) respectively per time point of measurement and per audio version. Error bars indicate the standard error.

To sum up, the manipulation check revealed that the VR-TSST was able to evoke a profound stress reaction in terms of an increase in subjective stress and heartrate. Furthermore, the repeated-measures ANOVAs revealed no significant main effect of audio version, or an interaction effect of audio version and time point on heartrate, stress rating, EMG, or presence rating, all ps > .05, all $\eta_p^2 s$ < 0.03. As we did not find significant differences between audio versions and no interactions involving audio versions on all measures, equivalency tests were conducted in a second step.

6.4.3 Equivalency tests

Four two-one-sided-tests (TOSTs) were conducted to test equivalency of audio version in terms of stress reaction (rating and heartrate) and presence (rating and MPS). Boundaries of equivalency are explained in the methods section. To begin with, the increase of heartrate from baseline to peak (defined as physiological reaction measured via ECG) in the TTS group (M = 20.55, SD = 16.19) is equivalent to the increase of heartrate in the PHS group (M = 20.20, SD = 18.82). The difference lies within the upper bound of equivalency, for

which the mean standard deviation over both audio versions were taken (± 12.729), $t(38) = 3.03$, $p = .002$, and the lower bound, $t(38) = -3.21$, $p = .001$. The increase of subjective stress in the TTS group ($M = 45.25$, $SD = 22.50$) was not inferior to the increase in the PHS group ($M = 37.05$, $SD = 26.71$), since the difference was significantly smaller than the upper bound of equivalency (± 10), $t(38) = -2.33$, $p = .013$. The difference was negative (indicating numerically higher stress in the TTS group) and since it did not lie within the lower bound, $p > .05$, equivalency could not be confirmed.

Concerning presence ratings, equivalency between TTS group ($M = 61.08$, $SD = 18.94$) and PHS group ($M = 63.53$, $SD = 24.14$) could not be shown since the confidence interval of the between-groups difference exceeds the upper and lower bounds (± 10), ps > .05. For presence measured after the VR-TSST with the MPS, also no equivalency could be shown. With a mean sum of $M = 29.4$ ($SD = 28.5$), the TTS group did at least not report significantly lower levels of social and physical presence than the PHS group ($M = 31.3$, $SD = 30.5$). For equivalency (bounds ± 5), not enough evidence was gained, all ps > .05.

For comparisons, for which neither differences nor equivalency between the audio versions could be shown, we computed Bayes Factors of independent t-tests (not preregistered) to gain further evidence for or against equivalency. For subjective stress, the null hypothesis of no difference between groups a Bayes factor of 2.09 was found, implying that the data is about two times more likely under equivalency of audio versions than under a difference. For presence ratings, a Bayes factor of 3.08 was found, implying that the data are about three times more likely when presence is equivalent in TTS and PHS group than the opposite. For MPS sum values, a Bayes factor of 2.67 was found. A common way of interpreting Bayes factors is to suggest a Bayes factor 3 as threshold above which good evidence for a hypothesis is assumed (Jeffreys, 1998), whereas elsewhere it is advised against a reduction of Bayes factors to a dichotomous decision since this leads to a loss of information and still is arbitrary (Wagenmakers, 2007).

6.4.4 Presence, stress, affect, and anxiety

Interestingly, main effect of time on presence was found, $F(1.68, 63.75) = 7.99$, $p = .002$, $\eta_p^2 = 0.17$ (no preregistered hypothesis or analysis), indicating that the feeling of presence seemed to depend on the stressfulness of the time in VR. Additionally, ratings of presence and stress (averaged over all rating time points) were found to be correlated, $r(38) = .34$, $p = .031$

(see Figure 22), indicating that the higher participants were stressed the more present they felt in VR.

A linear model was specified to investigate the possible influence of social anxiety on stress reaction. Neither on any of the stress ratings nor on heart rate at any time point an influence of the SPIN (questionnaire) value was found, all *p*s > .05. Exploratorily (not preregistered), relationships between (social) anxiety, presence, and affect were investigated. No influence of social anxiety on physical, social, and general presence was found, $p > .05$. Hence, the individual day-to-day anxiety level (STAI trait anxiety, Spielberger et al., 1971) was correlated with the retrospectively reported arousal during the stress scenario, $r(38) = .31$, $p = .049$. Furthermore, the more negative affect participants reported before the experiment, the higher was their first presence rating in VR, $r(38) = .34$, $p = .031$. Negative affect after the VR scenario was also correlated with social presence, $r(38) = .41$, $p = .009$, and physical presence during VR, $r(38) = .46$, $p = .002$. Not surprisingly, social and general anxiety was correlated, $r(38) = .48$, $p = .002$. In contrast to the null findings concerning the relationship between social anxiety and social stress, it was found that the more socially anxious participants were the higher arousal they reported, $r(38) = .39$, $p = .012$. Furthermore, the higher the social anxiety of participants was, the more dominant they rated the virtual agents, $r(38) = .36$, $p = .021$, and the less pleasant they perceived the virtual interaction, $r(38) = -.37$, $p = .020$.

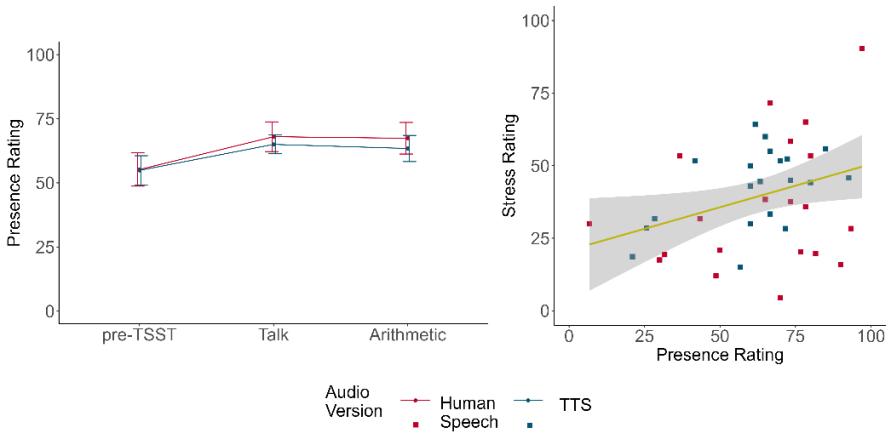

Figure 22: On the left: Mean presence rating as a function of time point and audio version. Error bars indicate the standard error. On the right: Correlation between presence and stress rating.

6.5 Discussion

This study investigated with a standardized virtual social stress test (VR-TSST) whether the use of virtual agents in combination with TTS synthesized "voices" can provoke stress reactions equivalent to those provoked by virtual agents synchronized with prerecorded

human speech (PHS). Psychological and physiological stress reactions as well as presence and further clinically relevant variables were assessed. Participants reported higher subjective stress during the VR-TSST than before and their heart rate increased significantly. No difference was found between the participants who interacted with the TTS-synchronized agents and those who interacted with the agents synchronized with human voice recordings for any of the stress parameters. Neither for presence ratings during VR nor physical or social presence evaluated afterwards with the MPS, an influence of audio version was found. Concerning trapezius muscle activity, which was previously found to be a biomarker of tension and mental stress (e.g. Wijsman et al., 2013), no stress reaction measured via EMG could be found in this study. This may be due to the fact that the baseline measurement took place at the beginning of the VR scenario and maybe the participants moved their heads involuntarily more here (orientation) than during the stress task (focus on the committee). Alternatively, the EMG data may also reflect anticipatory anxiety.

Equivalency tests showed equivalency of TTS and PHS concerning the evocation of stress in terms of an increase in heartrate. Furthermore, non-inferiority of TTS and PHS concerning the evocation of stress in terms of subjective stress levels was shown. Concerning presence, no equivalency could be shown using the TOST method. The reason why equivalency could be shown for heartrate, in part for subjective stress, and could not be shown for the presence ratings primarily lies in the defined boundaries. For the physiological data, boundaries of equivalency of one standard deviation were preregistered due to the experience that these data are commonly relatively noisy. This resulted in an effect size of $d = 1$ which was defined as smallest effect size of interest. Note that this effect size is normally interpreted as big. On the other hand, the boundaries defined for the subjective ratings, were defined in such way that only with almost exact same values, equivalency could have been shown for data based on a sample size of $N = 40$. The defined boundaries –chosen pragmatically on the basis of the scale size - come close to an effect size of $d = 0.4$ given the variance of the data. Vice versa, a sample size of $N = 266$ would have been necessary to find an effect of this size with a power of 90% and an alpha of 5% (according to a power analysis using G*Power 3.1, Faul et al., 2009). For the stress ratings, at least non-inferiority could be shown, due to the fact that in the TTS group numerically higher levels were assessed.

In reaction to these shortcomings, Bayes factors for independent samples t-tests were computed. This statistical approach is less susceptible to the influence of sample sizes and the output can be directly interpreted as the probability of data under e.g. the hypothesis of equivalency (Wagenmakers, 2007). For all variables for which neither statistical differences

between audio versions nor equivalency with TOSTs could be shown, Bayes factors > 2 were found. The here presented data are at least two times more likely under equivalency of TTS and PHS version than under the hypothesis of a difference. To sum up, while the equivalency tests using TOST method must be interpreted cautiously due to the (for this method) comparable small sample size and the partly broad equivalency boundaries, no statistically significant differences could be found between the TTS and PHS version. Furthermore, evidence was gained via Bayes factors that equivalency of the audio versions is more probable than a difference of interest.

The focus of this study was on the effects of different audio versions on clinically relevant psychological variables such as physiological and psychological arousal, anxiety, and presence. Therefore, comparably implicit measures were assessed for the experience in VR when TTS or PHS stimuli were used. While it can be cautiously concluded that TTS and PHS work equally well in supporting a stressful virtual interaction, no statements can be made about the preferences of participants concerning used speech synchronization of virtual agents. For laboratory stress scenarios (in which preference of voice does not matter) such as the TSST, low expressiveness of voices and neutral reactions are targeted which makes these scenarios all the more suitable for the use of TTS.

In scenarios for which high expressiveness characterized by pitch variations and emotional voices are needed, the use of TTS may be still inferior to human voices. Also, for VR scenarios, in which voices should be as pleasant and as appealing as possible, TTS may be inferior. The findings of previous studies indicate that humans show clear preferences for human voices (e.g. Abdulrahman & Richards, 2022; Dai et al., 2024). Especially children do not only respond more positively to expressive voices (Gustafson & House, 2001), but also benefit from more expressive voices by higher motivation and focus (Westlund et al., 2017). These results were gained by manipulating synthetic voices, nonetheless, even after more than three decades of research on artificial emotional intelligence (for a review: Schuller & Schuller, 2018), emotional voice conversion yields lower quality ratings than the human reference (Zhou et al., 2022).

Deep learning seems to be the most promising approach according to the frequency of its use (Triantafyllopoulos et al., 2023). Besides indistinguishability, the same effects of emotional TTS on human behavior as human affective voices are targeted (e.g. Cohn et al., 2024). To assess emotional TTS synthesis or conversion in the context of clinical psychology applications, affective learning mechanisms, such as social fear or appetitive conditioning could be used (e.g. Reichenberger et al., 2017, 2020). In such paradigms conducted in VR,

(virtual) opponents behave in a way that provokes strong feelings (e.g. insults or compliments). To sum up, the implementation of TTS is beneficial (see introduction) and does not lower the effectiveness of a virtual stress scenario. Hence, in the field of virtual clinical psychology, research is still needed on the advantages or side effects of emotional TTS conversion.

Furthermore, a relationship between presence and stress was found. Presence increased with the stressfulness of the task and hence, stress ratings were positively correlated with presence ratings. In addition, negative affect was interrelated with presence in VR. These findings fit well into a framework that describes a close relationship between presence and affective involvement in VR. The perception of arousal in combination with immersion leads to the judgment that one experience higher levels of presence in VR (Diemer et al., 2015). At the same time, presence was discussed as a prerequisite for the evocation of feelings via virtual scenes.

While in this study, the level of social anxiety of participants did not influence their stress reaction or feeling of presence, an influence on the perception of the virtual interaction was found. These results, however, must be interpreted cautiously since they were only investigated exploratorily and no correction for alpha inflation was done. Considering that the induction of fear was found to increase presence (Diemer et al., 2015), a comparison of TTS and human voices in a more neutral non-affective VR scenario could furtherly help to clarify whether the use of these voice categories result in equivalent virtual interactions. The focus may lay then on the investigation of social presence.

6.6 Conclusion

Research on the improvement of text-to-speech synthesis is ongoing and resulted in an up-to-date high quality, e.g. shown by partly indistinguishability with human voices. Especially the use of artificial intelligence and deep learning approaches enables natural voices adaptable to specific needs. In the application field of clinical psychology, the implementation of TTS has the potential to enhance and flexibilize virtual interactions as used e.g. for research paradigms to examine human stress reactions and, carrying on, virtual exposure therapy of social anxiety disorders. This study aimed to investigate the equivalency of TTS stimuli to prerecorded human speech in a social stress scenario (VR-TSST). Virtual agents were either synchronized with TTS stimuli or PHS. Participants had to accomplish stressful tasks in front of the virtual agents and high social evaluative threat was targeted. Psychological as well as physiological stress reactions could be induced. Neither the increase in participants' heart rate nor their

subjective stress level depended on the audio version (no significant interaction effect). Additionally, no significant differences concerning physical or social presence during VR were found between the participants in the TTS or the PHS group. Also, equivalency tests yielded equivalency for heart rate and non-inferiority for subjective stress. For presence, no statistically significant equivalency could be shown using the two one-sided t-tests (TOST) method but support for the hypothesis of equivalency was found when computing Bayes factors for the probability of the data under equivalency. Interesting open questions on the use of TTS in virtual clinical psychology concern on the one hand emotional voice conversion for highly affective scenarios and the effects of TTS on non-affective and non-arousing virtual scenarios, since emotion induction itself was shown to enhance the feeling of presence in VR.

7 Study 4: Aurally impressed, yet not more stressed: On the relationship between audiovisual realism, social anxiety, and presence in a virtual social stress scenario

Sarah Roßkopf, Andreas Mühlberger, Felix Stärz, Matthias Blau, Steven van de Par, and Leon O.H. Kroczek

The following article was accepted on 08 March 2026 and published online on 23 March 2026 in PLoS One following peer review. The official citation that should be used in referencing this material is: Roßkopf, S., Mühlberger, A., Stärz, F., Blau, M., Van de Par, S., & Kroczek, L. O. H. (2026). Aurally impressed, yet not more stressed: On the relationship between audiovisual realism, social anxiety, and presence in a virtual social stress scenario. PLoS One, 21(3): e0345565. https://doi.org/10.1371/journal.pone.0345565.

7.1 Abstract

Binaural auralizations can create spatial hearing impressions that closely resemble real sound sources, enhancing immersion and realism in virtual environments. Although social interactions often involve emotional responses such as stress (e.g., during a job interview), the interplay between emotion and binaural auralizations in virtual social interactions remains underexplored. Therefore, we investigated the effects of audiovisual realism in a virtual social stress scenario based on the Trier Social Stress Test. Acoustic realism was manipulated between subjects using head-tracked binaural auralizations and a diotic condition. For binaural auralizations, simulated binaural room impulse responses were based either on individual or generic head-related impulse responses. Stressfulness was also varied: a control group performed a task with reduced cognitive demand and social-evaluative threat by only "testing" a virtual job interview scenario and reading aloud preformulated answers. Social presence, stress responses (measured by salivary cortisol, heart rate, and self-reports), and gaze behavior were assessed in 78 participants. The virtual scenario reliably induced stress across all audio conditions compared to the control version. Binaural auralizations were rated as more externalized and realistic than diotic audio, but did not significantly influence social presence, stress responses, or gaze behavior. Social presence increased with higher social-

evaluative threat and over time. Social anxiety was associated with greater social presence, altered gaze behavior (shorter latencies), and, to some extent, stronger stress responses. It also interacted with the auralization type in affecting social presence. Overall, enhancing acoustic realism with externalized auralizations did not affect stress or presence in the virtual scenario. Elevated stress levels also in the control condition may have masked potential audio effects, implicating the need for investigating binaural auralizations in less stress-related social contexts.

7.2 Introduction

Virtual Reality (VR) is a technology used to simulate compelling three-dimensional scenes evoking the feeling of actually being there, known as presence. Although interactivity and multisensory stimulation is known to improve presence, typically visual attributes of the virtual scene are modulated to enhance presence, such as the display's resolution (Lee, 2004; Slater, 2018). Recently, the implementation of binaural auralizations, and therefore enhancements of the virtual acoustic scene, has been used to improve audio-visual plausibility and presence (Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2024). However, the impact of binaural auralizations on presence and emotional responses in an interactive virtual social scenario which is associated with stress, is still unknown.

Virtual acoustics aim to simulate how an auditory event would sound in a specific environment. Often, the virtual sound sources are reproduced via headphones (Werner, 2018). An important attribute to describe realistic binaural hearing is the externalization of the perceived auditory events. Externalization refers to the phenomenon where a sound source is perceived as coming from the surrounding environment, rather than from inside the head, as is typically the case with headphone playback of stereo or mono sound (Best et al., 2020). Therefore, both surrounding and vividness of VR are increased by implementing spatialized sound, positively affecting immersion and realism (Agrawal et al., 2020). Typically, data on the user's head orientation is retrieved from the head-mounted display (HMD) and combined with the corresponding binaural room impulse response (BRIR), allowing the human sensory system to perceive a stable sound source which is located in the virtual room. Despite the improvement in the realism of VR due to the use of spatialized sound, it remains unclear whether and how these improvements modulate user experience in VR. On the one hand, presence was found to be enhanced by spatial sound (Bormann, 2005; Freeman & Lessiter, 2001; Kern & Ellermeier, 2020). On the other hand, no consistent effects were found on social presence, a sub-aspect of presence which is especially relevant for VR applications in the

context of, e.g., virtual social interactions. Social presence, the sense of being with another, as defined by Oh et al., is also influenced by the immersive qualities of VR (C. S. Oh et al., 2018). In their review, the positive effects of higher audio fidelity on social presence were summarized (C. S. Oh et al., 2018). However, only non-immersive environments presented on 2D screens or audio-only VR were investigated. Regarding immersive environments, in a recent study, binaural auralizations resulted in higher social presence ratings in direct comparison to less immersive sound in a virtual seminar room scenario (Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2024). Also, a positive effect of including individual head-related impulse responses was found, although this required a time-consuming measurement process. Contrarily, no benefits of binaural auralizations on social presence and communication behavior were found in a dyadic communication VR scenario (Immohr, Rendle, Lammert, et al., 2024). Overall, implementing binaural auralizations can be expected to improve the quality of VR experience in terms of realism and presence, but the effects on social presence in virtual social interactions remain unclear.

Human social interactions are complex and challenging, often evoking emotional responses (Shiban, Diemer, et al., 2016; Zimmer et al., 2019). In VR, these affective experiences can be explored in a standardized manner, as even simply designed avatars can evoke substantial feelings of social presence (Latoschik et al., 2017; Yoon et al., 2019). Virtual social interactions are not only a helpful tool in communication and collaboration research, but can be used for basic research on stress and its neurophysiological basis (Shiban, Diemer, et al., 2016; Zimmer et al., 2019). The Trier social stress test (TSST) is the laboratory gold standard for investigating acute stress, especially concerning the neuroendocrinological domain (Allen et al., 2017). Typically, an increase in salivary cortisol is used as an indicator of stress. Furthermore, physiological stress reactions are investigated using the TSST, e.g., by measuring heart rate (Kupper et al., 2021), skin conductance (Zimmer et al., 2019), myoelectrical activity, or body temperature (Pribék et al., 2021). Finally, self-reports on affect and stress are typically collected, e.g. (Shiban, Diemer, et al., 2016). The complexity of human stress is reflected in differential response patterns, which are influenced not only by inter-individual traits but also by situational and contextual factors (Kupper et al., 2021; Shiban, Diemer, et al., 2016).

While conducting the TSST in VR offers several advantages in terms of logistical effort and standardization (Allen et al., 2017), not all studies found that the VR-TSST evoked a (neuroendocrinological) stress response. In a direct comparison study, it was further found that the VR-TSST evoked equal (or even higher) subjective stress, but decreased cortisol

responses compared to the in-vivo version (Shiban, Diemer, et al., 2016). These findings were linked to the degree of immersion. More specifically, low levels of social presence in some VR-TSST studies were suggested to contribute to the differential effects (Shiban, Diemer, et al., 2016). The stress induction via TSST is mainly driven by social evaluative threat and uncontrollability (Allen et al., 2017; Frisch et al., 2015). In order to trigger this threat, subjects need to feel that others are present, and thus, a certain degree of social presence is required.

The impact of social evaluative threat on stress is also reflected in TSST studies, including patients with social anxiety disorders. These patients experience an intense fear of social situations involving potential judgment or embarrassment, resulting in daily-life impairments (American Psychiatric Association, 2013), and show stronger subjective stress reactions than healthy controls, but unaltered cortisol responses to the TSST (Grace et al., 2022). Furthermore, socially anxious participants were found to experience higher levels of social presence (Felnhofer et al., 2019). In general, the effectiveness of VR applications, whether for the treatment of mental disorders (Wechsler et al., 2019) or for social skills training (Howard & Gutworth, 2020), increases with immersion (Wiebe et al., 2022).

The implementation of spatial audio, by using head-tracked binaural auralizations, may help to increase social presence within virtual scenes and therefore the impact of virtual social situations. If the sound of speech is perceived in the same location as a virtual agent, it may feel more like a naturalistic interaction compared to seeing a speaking agent but perceiving the produced sound inside the head. It is therefore to be investigated whether increased audiovisual realism affects social presence and, in turn, increases social stress reactions in a stressful virtual interaction. While the effectiveness of an intervention in VR seems to depend on (social) presence, it was also found that the immersivity of a virtual environment may be less relevant in demanding situations especially when anxiety is high (Gorini et al., 2011). The affective states, especially arousal and fear, are in a mutual relationship with presence (Diemer et al., 2015). Initial increases in fear positively affect presence, and enhanced presence in turn intensifies the fear reaction towards virtual phobic stimuli (Peperkorn et al., 2015). Therefore, the effect of binaural externalized auralizations in virtual social interactions is to be investigated under different stressful conditions. The potential positive effect of spatial sound may be higher in a social situation with comparable low social evaluative threat due to the interrelation of immersion, arousal, and presence. Furthermore, potential influences of the participants' social anxiety trait should be considered. Since a relevant application of virtual social interactions is in the treatment of social anxiety

disorders, differential effects of binaural auralizations on stress response and social presence are to be investigated.

To the best of our knowledge, no systematic evaluation of the influence of binaural auralizations on the induction of stress and social evaluative threat has been conducted. Therefore, this preregistered study (https://osf.io/7gy3p) aimed to investigate how binaural auralizations in a virtual social stress scenario (VST) affects presence and social presence as well as stress reactions (neuroendocrinological, physiological, and subjective). If the binaural auralizations are perceived as externalized, we refer to them as externalized auralizations. For that reason, we manipulated the social stress scenario (low vs. high-stress) and investigated low- versus high-socially anxious participants. A subsequent research question was whether individual acoustic measurements are necessary to simulate the externalized auralizations to maximize the effects. We derived the following hypotheses:

H1: Externalized auralizations increase social presence in virtual interactions compared to non-externalized ones.

H2: The VST evokes stronger stress responses in terms of higher increases of a) salivary cortisol, b) heart rate, and c) stress ratings from baseline to post-stress measurements (or during stress measurement for heart rate) when externalized auralizations are used.

H3: Concerning gaze behavior, we expect enhanced visual spatial attention when externalized auralizations are used, in the form of shorter latencies for the first fixation on virtual speakers.

H4: We hypothesize that the difference in levels of social presence between externalized and non-externalized auralizations is lower in the high-stress condition compared to the low-stress condition.

H5: In the high stress condition involving the externalized auralizations, social anxiety is a stronger predictor for social stress than in the high stress condition involving the non-externalized ones.

H6: Equivalency of individualized and generic externalized auralizations: Based on our previous findings (Blau et al., 2021; Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2024; Stärz et al., 2025), we expect that using individualized measurements for the externalized audio condition will not result in further improvements concerning social presence, stress reactions, or visual spatial attention.

7.3 Methods

7.3.1 Participants

Our sample (N = 78) consisted of 52 female and 26 male participants. No one identified as non-binary. Due to legal and hormonal reasons, only adults between 18 and 55 years were included. Our sample consisted mainly of young adults aged between 18 and 39 (M = 23.9, SD = 3.8). The sample size was based on a power analysis conducted with G*Power 3.1 (Faul et al., 2009), indicating N = 42 (14 non-externalized vs. 14 externalized–individual and 14 externalized-generic participants) to be sufficient to detect an effect size of d = 1.10 with *alpha* set at .05 and 1 – *beta* = .95 for a one-sided paired sample t-test (externalized vs. non-externalized). In a previous study (Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2024), we found effect sizes of d > 1.10 for the comparison of externalized auralizations with the anchor control condition concerning social presence (primary outcome variable). We increased the sample size to N = 78 to have at least 13 participants per *Stress x Audio* group. The majority of participants were students (n = 71). Table 6 shows demographic, psychological, and further relevant characteristics of the experimental groups. We examined whether the experimental groups differed prior to the manipulations. As shown in Table 6, no differences emerged regarding outcome variables. Groups were also comparable in demographic and clinical characteristics, except for negative affect.

Table 6. Participants' characteristics per experimental conditions.

	high-stress (n = 39)				low-stress (n = 39)							
	non-externalized		externalized		non-externalized		externalized		Stress		Audio	
	(n = 13)		(n = 26)		(n = 12)		(n = 27)		t or χ^2	p	t or χ^2	p
Sex, male, n (%)	4	31	9	35	4	33	9	33	0	1	<0.01	1
Women using hormonal contraception, n (%)	5	38	6	23	5	42	9	33	0.36	.549	1	.317
Women in luteal phase or irregular cycle, n (%)	4	31	13	50	4	33	11	41	0.02	.882	0.16	.692
Age, y, M (SD)	24.1	3.2	23.8	3.6	24.1	2.8	23.8	4.8	0	1	-0.37	.710
Depression Screening (BSI-D), M (SD)	1.4	0.4	1.2	0.3	1.3	0.4	1.3	0.4	0.41	.682	-1.06	.296
Social Anxiety (SPIN), M (SD)	35.2	11.7	29.1	11.6	27.9	9.8	28.2	11.5	1.17	.244	-1.11	.274
Positive Affect (PANAS), M (SD)	3.1	0.7	2.8	0.5	2.8	0.7	2.9	0.6	-0.02	.985	-0.49	.625
Negative Affect (PANAS), M (SD)	1.5	0.4	1.5	0.4	1.2	0.2	1.3	0.2	2.78	.007	0.20	.840

Baseline												
Salivary Cortisol Level in nmol/l, M (SD)l	3.4	2.7	5.0	5.0	3.9	3.5	3.0	1.7	1.49	.143	0.36	.719
Stress Rating, M (SD)	41.3	26.9	38.3	23.9	27.5	17.6	41.2	24.6	0.49	.628	0.92	.360
Heart Rate, M (SD)	92.8	15.0	86.5	14.3	83.0	12.4	87.5	13.4	0.69	.490	-0.30	.763

Note. Bold values indicate significant differences
Abbreviations: BSI-D, Brief-Symptom-inventory-Depression; SPIN, Social phobia Inventory; PANAS, Positive Affect Negative Affect Scale; χ^2-tests were conducted for categorical data, and t-tests were conducted for continuous variables.

Participants were recruited via the university's participant management system and social media. All reported unimpaired hearing, normal or corrected vision, and at least five years of German-speaking experience (two were non-native speakers). No participant met criteria for a current affective episode, generalized anxiety disorder, or acute suicidal tendencies as confirmed with the Mini International Neuropsychiatric Interview (M.I.N.I., Sheehan et al., 1998). None reported current psychotherapy, psychotropic medication, cardiovascular or neurological conditions, tinnitus, or acute respiratory, sinus, or ear infections.

Furthermore, measures were taken to reduce disruptive influences on salivary cortisol levels. Self-reporting pregnancy, lactation, or intake of medication containing glucocorticoids such as cortisol were defined as exclusion criteria, as well as regular smoking (more than 5 cigarettes per day). To control for menstrual cycle effects on cortisol, female participants were tested during the luteal phase (2–3 weeks after self-reported cycle onset). Females using hormonal contraception (n = 23) or self-reporting no regular cycles were tested independently of the current cycle. To control for circadian effects on cortisol, especially the cortisol awakening reaction (Goodman et al., 2017), the experiments took place between 1 and 8 p.m. Participants were instructed to abstain from cannabis or any other psychotropic substances for three days, and from nicotine and alcohol for one day prior. Ninety minutes before testing, participants were instructed not to brush their teeth or eat a large meal. During the experiment, only water was allowed.

The study was realized in compliance with the Declaration of Helsinki and was approved by the ethics committee of the University of Regensburg (Ref-No.: 20-1804-101). All participants gave written informed consent. The study was conducted from December 2023 to July 2024. Participants received financial compensation, or psychology students, if preferred, course credits.

7.3.2 Study design

The study employed a between-subjects design manipulating Audio (non-externalized, externalized-individual, externalized-generic) and Stress (low vs. high). The low-stress condition involved reduced social-evaluative threat and cognitive demand. Time was a within-subjects factor due to the repeated measures (stress responses, ratings). Primary outcomes included self-reported social presence via questionnaire (MPS), salivary cortisol increase, heart rate, and subjective stress. Gaze behavior was analyzed via first fixation latency, dwell time, and accuracy of first fixations on virtual agents. Secondary outcomes included in-VR social presence ratings, cortisol responder rates, adverse effects, and perceived audio quality. Social anxiety was analyzed as an individual difference factor.

7.3.3 Materials

7.3.3.1 Low vs. high social stress manipulation

To induce psychosocial stress in a controlled and standardized manner, we used an adaptation of the virtual reality version of the Trier Social Stress Test (VR_TSST, Kirschbaum et al., 2008; Shiban, Diemer, et al., 2016). To enhance the salience and potential impact of the audio condition, the VST included a higher proportion of committee speech. An introductory phase was added, during which virtual agents presented their roles and expertise. Instead of the standard VR-TSST math task, participants completed a question-and-answer (Q&A) round with 30 challenging job interview questions, each followed by 20 seconds for spontaneous responses. At the start of the VR, participants received condition-specific instructions. The high-stress group was told they would undergo a job interview for their “dream job” and should perform at their best. The low-stress group was informed they were testing a VR training scenario and should simply read prewritten answers aloud.

7.3.3.2 Virtual reality set-up

The virtual committee consisted of three males and one female who were formally dressed (suits, costumes, see Figure 23). They were created using MetaHumans (MetaHuman Creator, Unreal Engine, & Quixel Bridge; Epic Games) and Blender (v2.79, Blender Foundation). Their expressions were emphatically neutral, and they gave no feedback throughout the interaction to trigger social evaluative threat. All verbal interactions were pre-recorded voiceovers triggered by the VR game engine (see section Audio set-up). Lip synchronization was realized using Audio2Face AI (NVIDIA OmniverseTM).

Figure 23: Virtual Stress Scenario. Left: high-stress; right: low-stress condition.

The VST took place in a small seminar room of the University of Regensburg. We created two photorealistic models of this seminar room with the Unreal Game Engine (v 4.27, Epic Inc.) and Blender (v 2.79). One room was used for the job interview, and was equipped with the committee behind tables with tablets, writing materials, a whiteboard, a camera, etc. The other room was the preparation room and was equipped with a table, chair, and a notebook on which further instructions were written. The visual virtual environment was presented via an HMD (Vive Pro Eye, HTC). An inaudible work station with passive cooling was used (Silentmaxx PC Kenko S-770i). The starting position of participants in the physical room was matched to the corresponding position in the virtual visual room model via an in-house-developed two-point calibration technique using custom-made mounts for the HTC motion controller (Kroczek et al., 2023).

7.3.3.3 Audio set-up

Three different auralization types were used. Two of the three auralizations (individual and generic HRIRs) were simulated in such a way that they evoke highly spatialized, realistic, and externalized hearing impressions ("externalized auralizations"). The third auralization was a "diotic" rendering, which should evoke a non-externalized hearing impression since binaural cues were eliminated. The auralizations were generated based on BRIRs simulated with RAZR (v0962b; Wendt et al., 2014). The simulations incorporated the dimensions of the experimental room (6.8 m × 4.8 m × 3.3 m), source directivity of loudspeakers (Genelec 8030b, Genelec Oy), and frequency-dependent absorption coefficients averaged in octave bands per room wall. The simulated reverberation time (T20 = 0.8s) was fitted to the physically measured monaural impulse responses of the experimental room.

For the individualized auralization, BRIRs were simulated based on individual HRIRs, whereas the generic condition, HRIRs from a head-and-torso simulator (KEMAR Type 45BB, GRAS Sound and Vibration A/S, Holte, Denmark) were used. All HRIRs were recorded using a measurement setup that replicated the system developed at Jade Hochschule Oldenburg, see

(Blau et al., 2021) for details. The simulations covered 37 azimuthal orientations (−90° to +90° in 5° steps) and nine elevation angles (−30° to +30° in 7.5° steps), with a fixed ear height of 1.60 m. For the diotic auralizations, the left and right BRIRs of the generic condition were averaged. The auralizations were combined with individualized headphone equalization and real-time head tracking via the HMD.

Audio was presented through extra-aural headphones (AKG K1000, AKG Acoustics GmbH, Vienna, Austria) mounted on the HMD using custom 3D-printed holds (Stärz et al., 2023), powered by a headphone amplifier (Lake People G103P, Lake People Electronic GmbH, Konstanz, Germany) and an external audio interface (RME Fireface UC, Audio AG, Haimhausen, Germany).

Dry recordings of four trained speakers were used to generate the auralizations. These recordings were loudness-normalized using the integrated loudness function from MATLAB's Audio Toolbox™ (following EBU R 128) to minimize loudness differences between individual speakers and Hann-windowed (10 ms per flank) to prevent onset or offset artifacts. The total duration of the speech was 5 minutes and 19 seconds.

7.3.4 Procedure

7.3.4.1 Audio measurements and externalization instruction

The experimental procedure comprised two appointments. During the first, participants gave written informed consent and completed psychoacoustic measurements based on their assigned audio condition. Those in the externalized-individual group underwent the measurement of HRIRs, which took approximately 30 minutes (for further details, see Blau et al., 2021). All participants underwent a headphone impulse response measurement (about 5 min). They were then introduced to the concept of externalization to prepare them for the related ratings during the second appointment. They were shown the externalization rating scale (see Table 10, 10.1) ranging from "0: fully inside the head" to "100: fully outside". The instructor explained that typical television or a loudspeaker sound corresponds to full externalization (100), while headphone audio is usually perceived non-externalized (0), with intermediate perceptions also possible. To familiarize participants with the scale, they rated three binaural auralizations and one non-spatialized audio sample. For this procedure, no head-tracking was used, and the auralizations, audio stimuli, and headphones (model HD 800, Sennheiser electronic GmbH & Co. KG, Wedemark, Germany) were different from those of the VST.

7.3.4.2 Main experiment

Pre-assessments and preparation

The second appointment started with general instructions, a check for exclusion criteria, and three sections of the neuropsychiatric interview. Afterwards, participants completed questionnaires (for further information on all used questionnaires see section Measurements – Self-Reports), first the demographic, followed by the brief symptom inventory (BSI)-18, the PANAS, and then the SPIN. Next, the first cortisol saliva sample was collected, the electrocardiogram (ECG) electrodes were attached, and psychophysiological recording started. The participants were then familiarized with the HMD and the controller before entering the experimental room, guided to the starting position by the experimenter and virtual footprints. Only then was the virtual replication of the room displayed. After eye tracking calibration, a 90-second ECG baseline was recorded while participants stood still and upright. Practice trials introduced the rating procedure, followed by the first ratings (Figure 24).

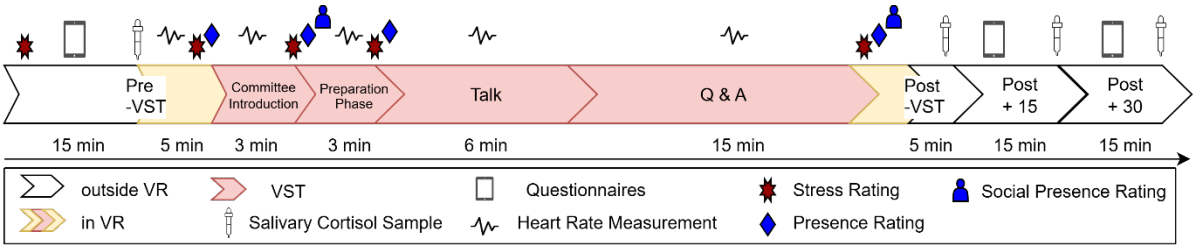

Figure 24: Experimental Procedure of the VST. Key measurement time points inside and outside VR.

VST

After the first ratings, VR scenes varied according to the experimental condition. In the low-stress group, participants were told they were testing a VR job interview training tool and instructed to read aloud predefined answers and a written talk. In the high-stress group, participants were asked to imagine applying for their dream job, to perform at their best, deliver a talk, and answer questions spontaneously. After confirming the instructions via button press, all participants were teleported to the virtual interview room (see Figure 23), where virtual agents introduced themselves (committee introduction) and instructed them to prepare a short talk. Participants were then teleported to a virtual preparation room with a desk and notebook, aligned with physical furniture. There, they completed further ratings and had three minutes to either prepare their talk (high-stress) or read the predefined version (low-stress), followed again by ratings (preparation phase). Participants then stood up, returned to the interview room, and began their talk upon instruction. After six minutes, a virtual agent

ended the talk and initiated the Q&A with 30 challenging job interview questions (see Table 9, 10.1) adapted from an online source (Mai, 2021). Each question began with a turn-taking sequence between the previous and current virtual speaker. Participants had 20 seconds to respond, spontaneously in the high-stress condition, or by reading predefined answers in the low-stress condition. The questioning agent ended each sequence with a brief, neutral remark (e.g., “Mhmm, thank you.”). After the final question and a closing statement by an agent, participants completed ratings on the experience of stress, VR and subjective audio quality features (see Table 10, 10.1) . They then left the room and removed the HMD.

Post-assessment

Immediately after the VST, the second saliva sample was collected. Participants then completed further questionnaires via tablet: MPS, SSQ, qualitative questions on (auditory) VR experience, PANAS, SVF-78, hearing-related questions (e.g., musical experience, audio sensitivity), and the DAS-18. To assess peak cortisol response, two additional saliva samples were collected 15 and 30 minutes post-VST, following Dickerson & Kemeny, 2004. Finally, the instructor checked participants’ current affective state and offered a referral to the university outpatient clinic if needed (no cases occurred).

7.3.5 Measurements and preprocessing

7.3.5.1 Self-reports

Subjective data was assessed using several questionnaires and analog rating scales implemented within the VR scene. Each of these scales included a rating item (see Table 10, 10.1), with verbal anchors at both ends, 0 ("not at all") and 100 ("very much"), and a slider that participants could adjust continuously using the HTC motion controller. Outside VR, all subjective data were collected using a tablet (Apple iPad Pro, 12.9-inch, 4th generation, model year 2020) and SoSci Survey (Version 3.1.06, Leiner, 2019).

Stress was rated before VR (slider rating on tablet): “How stressed do you feel on a scale from 0: not stressed at all - to 100: maximally stressed?”, and during VR with the three stress items depending on the time point of measurement (see Table 10, 10.1). For each participant, the peak subjective stress level was defined as the highest self-reported stress following the onset of the VST. The German version of the MPS, the multimodal presence scale (Makransky et al., 2017; Volkmann et al., 2018), was used for standardized (social) presence measurement. Also, the presence ratings within the VR scene (see Table 10, 10.1) were based on subitems of the MPS.

Further questionnaires were used. We used the SSQ, the simulator sickness questionnaire (Kennedy et al., 1993), to assess possible adverse effects of VR. The occurrence of psychopathological symptoms was assessed with the BSI-18, the brief symptom inventory (Franke et al., 2011); the current affective state with the PANAS, the positive and negative affect schedule (Krohne et al., 1996); and social anxiety with the SPIN, the social phobia inventory (Sosic et al., 2008). The subscale of the Sensorik Inventar for hearing (Zamoscik et al., 2017) was used as an indicator of audio sensitivity. To gain insights on stress-management, we assessed coping strategies with the SVF-78, the "Stress Verarbeitungsfragebogen" (Janke & Erdmann, 2008), and dysfunctional cognitions with the DAS-18, the dysfunctional attribute scale (Rojas et al., 2022). Figure 24 gives an overview of the measurements at several time points.

7.3.5.2 Heart rate

To assess heart rate as a physiological indicator of social stress, we continuously recorded ECG data throughout the VST. Therefore, three self-adhesive electrodes (Ag / AgCl, Ø = 40 mm; Diagramm Halbach GmbH & Co. KG, Schwerte, Germany) were attached to the participant, one on the sternum and one on each side of the lower costal arch. ECG data were collected using a portable wireless sensor (PLUX – Wireless Biosignals, S.A., Lisbon, Portugal). Data acquisition and storage were managed using the OpenSignals software (PLUX) and LabRecorder (Lab Streaming Layer, GitHub repository, 2014). The ECG recordings were analyzed offline using a custom MATLAB script (v R2022a, The MathWorks, Inc., Natick, MA, USA). Heart rate data were segmented into 30-second intervals, and the mean beats per minute (bpm) were computed for each segment and labeled with the respective experimental phase using markers sent by the VR engine. Further preprocessing was performed in the R statistical environment, Version 2024.04.2, (Development Core, 2019). The heart rate during the first 90 s of each segment was averaged and used for further analyses, as this corresponded to the length of the baseline measurement. Data from 10 participants had to be excluded due to technical errors, missing data, or markers.

7.3.5.3 Cortisol

Salivary samples to determine cortisol levels as a neuroendocrinological indicator for social stress were taken at four time points (pre-VST, post-VST, post+15, post+30) using salivette collection tubes (Sarstedt AG & Co., Nümbrecht, Germany). After the experiment, the saliva samples were stored at -20°. They were analyzed in single determination (standard) at the laboratory of Prof. Dr. Clemens Kirschbaum in Dresden, which provided the following

rationale: “After thawing, samples were centrifuged at 3,000 rpm for 5 min, which resulted in a clear supernatant of low viscosity. Salivary concentrations were measured using a commercially available chemiluminescence immunoassay with high sensitivity (Tecan - IBL International, Hamburg, Germany; catalogue number R62111). The intra- and interassay coefficients of variance were 2.2% and 2.9%. Three of the saliva samples were missing data (from a total of 312 samples). Participants (n = 2) with missing cortisol data (at baseline or peak) were excluded from analyses concerning salivary cortisol levels.”

Salivary cortisol levels were log-transformed (base 10) to normalize data. For statistical analyses, self-reported gender, age, and hormonal contraception (hc, 3 factors: male, female-no-hc, female-hc) of participants were included as covariates (Bärtl et al., 2024). Furthermore, for each participant, the peak cortisol level of salivary samples measured after the VST (post-VST, post+15, post+30) was computed. The difference between the peak level and pre-VST sample cortisol level was taken as an indicator of individual cortisol increase. Proportions of responders vs. non-responders were compared. Responders were defined as participants with a minimal cortisol increase of 15.5% from the pre-VST level to the maximal response level (Miller et al., 2013).

7.3.5.4 Gaze behavior

We used the eye-tracking system implemented within the HMD (VIVE SRanipal SDK, HTC corporation) for measurement of gaze behavior. Areas of interest (AOIs) were predefined and attached to all objects and agents in the virtual room. Gaze behavior was analyzed offline using a custom MATLAB script (v R2022a, The MathWorks, Inc., Natick,MA, USA) which categorized gaze as fixation or saccade behavior. Fixations were defined using both velocity (<75°/s) and gaze duration (>=140 ms) criteria (Holmqvist et al., 2011). We computed the latency from speech onset until the first fixation on the currently speaking agents, percentage of (in)correct fixations, as well as the dwell time on speaking agents / social AOIs. Eye tracking data from two participants had to be excluded due to technical errors.

7.3.6 Statistical analyses

Statistical analyses were conducted using the R environment (Development Core, 2019). For all hypotheses, first, mixed ANOVAs were computed, and Greenhouse-Geisser correction was applied in cases of violations of sphericity. Then, post-hoc t-tests were computed to follow up on significant effects. Directed one-sided t-tests for independent samples were computed to gain evidence on the superiority of externalized auralizations compared to non-externalized ones (H1, H2, and H3). When the requirements for parametric tests were not fulfilled, the

Mann-Whitney-U-Test, as a non-parametric equivalent, was computed and the Wilcox test (*W*) was reported. Holm procedures were used to correct for multiple comparisons (H2). For all hypothesis tests except for H6, audio manipulation was analyzed with two levels: non-externalized vs. externalized, including data from individual and generic audio groups. Interaction effects of Stress-by-Audio-by-Time on social presence were analyzed for the hypothesis on differential effects of auralizations and stress (H4). For the hypothesis tests concerning a higher stress reaction in the externalization group (H2), we only included data from the high-stress group, and exploratory investigated possible effects for the high- and low-stress groups. For the tests of non-superiority of simulations based on individualized HRIR in comparison to generic HRIR (H5), independent sample t-tests were computed with regard to the above-mentioned outcome variables. If one model resulted in significant differences, the equivalency hypothesis was rejected. Null hypothesis significance testing does not allow a conclusion on equivalence. Therefore, we additionally computed Bayes Factors for independent *t*-tests to investigate whether the equivalency hypothesis is more probable than the difference hypothesis. The BayesFactor package (Morey & Rouder, 2014) with the default Cauchy prior distribution was used, and the null-hypothesis was tested against the directed hypothesis of superiority either of externalized auralizations (H1) or individual BRIRs (H5). Bayes Factors greater than three were regarded as confirmatory since indicating moderate evidence (Jeffreys, 1998). For H6 (differential effects of social anxiety), general linear models on social stress with the predictor of social anxiety were computed for both audio condition groups.

7.4 Results

7.4.1 Manipulation check

First, we checked whether our intended manipulations were successful (see Figure 25). Indeed, the high-stress group reported significantly more maximal stress than the low-stress group ($W = 450.5$, $p = .002$, $d = 0.351$, $n1 = 39$, $n2 = 39$). Also, the audio manipulation was successful; the externalized auralizations (including individual and generic BRIRs) were rated significantly higher as externalized than the non-externalized one ($W = 1044$, $p < .001$, $d = 0.464$, $n1 = 53$, $n2 = 25$).

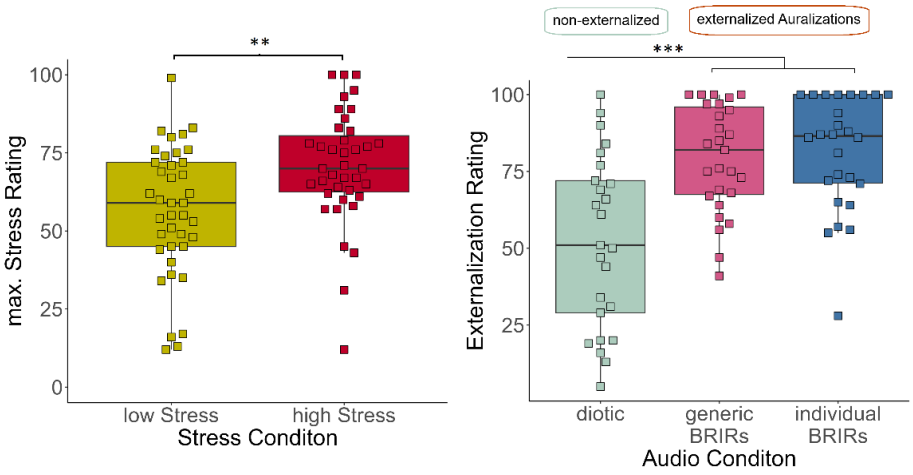

Figure 25: Manipulation check. On the left, maximal stress ratings per stress condition are displayed, on the right, externalization ratings per audio condition.

7.4.2 Social presence

Neither audio nor stress condition nor their interaction affected social presence measured via the MPS questionnaire conducted after the experiment, see Figure 26, Audio: $F(1, 74) = 0.11$, $p = .739$, $\eta_p^2 < 0.01$; Stress: $F(1, 74 = 0.00$, $p = .994)$, $\eta_p^2 < 0.01$; *Audio x Stress*: $F(1, 74) = 0.15$, $p = .702$, $\eta_p^2 < 0.01$. Social presence was not significantly higher in participants listening to externalized auralizations ($M = 2.73$, $SD = 0.93$) compared to those listening to non-externalized auralizations ($M = 2.66$, $SD = 0.89$); $t(48.65) = -0.35$, $p = .365$, $d = 0.08$. In addition to the MPS questionnaire, social presence was assessed with a single-item rating in VR directly after an interaction (two times). Again, neither a significant main effect of *Audio* was found, $F(1, 74) = 0.14$, $p = .709$, $\eta_p^2 < 0.1$; nor of *Stress*, $F(1, 74) = 2.81$, $p = .098$, $\eta_p^2 = 0.04$; nor a significant interaction between *Audio and Stress*, $F(1, 74) = 0.56$, $p = .458$, $\eta_p^2 < 0.1$. The mean social presence rating during VR was not higher in the externalized audio group ($M = 49.1$, $SD = 20.4$) than in the non-externalized audio group ($M = 50.8$, $SD = 20.8$), $t(46.32) = -0.32$, $p = .373$, $d = -0.08$. To sum up, we could neither confirm the hypothesized superiority of externalized auralizations concerning social presence, nor any differential effects of *Audio and Stress* on social presence.

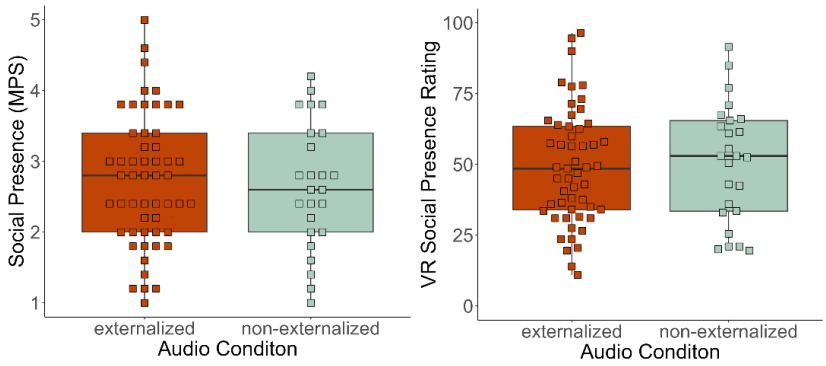

Figure 26: Social Presence. On the left: measured with the subscale of the MPS; on the right: with rating scales within the VR scene.

We additionally computed Bayes factors (BF) for independent *t*-tests to gain further insights on the (null-)effects of auralizations in our VST. For social presence measured via the MPS, a BF = 3.82 for equivalency of audio conditions was found, and for social presence measured via VR rating, a BF = 3.10 was found. Therefore, moderate evidence was gained that both audio conditions evoked equivalent levels of social presence.

Furthermore, a repeated-measures ANOVA was computed, including the different time points of social presence ratings to gain insights into the time course of social presence and a possible interaction with *Stress*. A significant main effect of *Time*, $F(1,148) = 7.90$, $p = .006$, $\eta_p^2 = 0.05$; and *Stress*, $F(1,148) = 4.71$, $p = .032$, $\eta_p^2 = 0.03$, was found. Figure 27 indicates that social presence increases throughout the VST and is higher in the high-stress group.

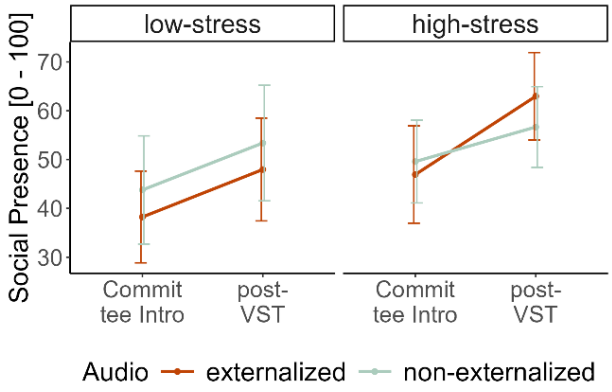

Figure 27: Social Presence rating as a function of stress and audio manipulation at two different measurement time points. Error bars indicate the standard error.

We conducted exploratory analyses on additional indicators of the quality of experience in VR, including physical presence, perceived realism, and subjective audio quality. Detailed results of these analyses are provided in the supporting information (see 10.1). Similar to social presence, physical presence also showed time-related effects, with a general increase observed over time (see Figure 32, 10.2). Presence (social and physical) positively correlated with acoustic realism and acoustic presence. Interestingly, acoustic realism and tone richness, but not acoustic presence, were positively affected by externalized

auralizations. Finally, audio liking and speech intelligibility were affected by externalization and stress, with these two conditions interacting marginally significantly (see Figure 33, 10.2). Furthermore, time-lagged Pearson's correlations across five measurement points between stress and social or physical presence were computed to investigate their causal relationship (see Figure 34, 10.2). Numerically, the strongest relationship between both social and physical presence and stress was found between the rating after the committee introduction and the subsequent stress rating (post-VST). This indicates that the more participants experienced the initial virtual social interaction as if really being in front of a job interview committee, the more stress they experienced later during the VST.

7.4.3 Stress induction

7.4.3.1 Salivary cortisol levels

A repeated measures ANOVA was conducted to examine the effect of *Time* (of salivary sample), *Stress* (high vs. low), *Audio* (externalized vs. non-externalized), and their interactions on the salivary cortisol level, while controlling for sex, hormonal contraception, and age. We found a significant interaction effect of *Time and Stress*, $F(1, 73)=7.49$, $p = .008$, $\eta_p^2 = 0.09$, confirming that the cortisol increase specifically occurred in the high-stress condition (see Figure 28). An effect size of $d = 0.41$ of VST (pre vs. peak) on salivary cortisol was found in the high-stress group. While sex also significantly influenced the salivary cortisol level, $F(1, 70) = 6.94$, $p = .010$, $\eta_p^2 = 0.09$, neither *Audio*, $F(1, 70) = 0.27$, $p = .605$, $\eta_p^2 < 0.01$, nor *Time*, $F(1, 70) = 0.03$, $p = .864$, $\eta_p^2 < 0.01$; nor any of the other covariates, HC: $F(1, 70) = 0.64$, $p = .437$, $\eta_p^2 < 0.01$, Age: $F(1, 70) = 0.640$, $p = .167$, $\eta_p^2 = 0.03$, had a significant main effect on salivary cortisol level. Also, the interaction effect between *Audio and Stress* did not reach significance, $F(1, 73) = 3.23$, $p = .077$, $\eta_p^2 = 0.04$. Furthermore, neither the interaction between *Audio x Time* was significant, $F(1, 73) = 0.05$, $p = .828$, $\eta_p^2 < 0.01$; nor the three-way interaction between *Audio x Time x Stress*, $F(1, 73) = 0.05$, $p = .826$, $\eta_p^2 < 0.01$. Contrasting our hypotheses, the increase in salivary cortisol level was not significantly higher in the externalization audio group ($M = 0.17$, $SD = 0.68$) than in the non-externalization group ($M = 0.17$, $SD = 0.42$), $t(37) = -0.02$, $p = .507$.

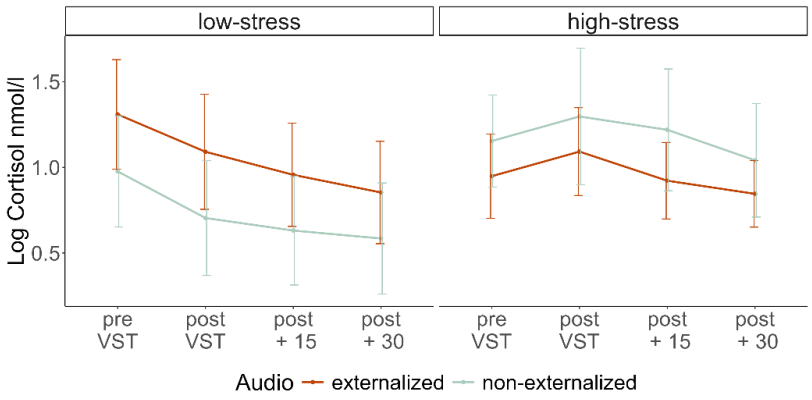

Figure 28: Log-transformed mean salivary cortisol (nmol/l) in response to the VST as a function of audio and stress. Error bars indicate the standard error.

Additionally, we analyzed whether the responder rates differed between experimental groups (see Table 7). A binary logistic regression model revealed that participants in the high-stress condition had a significantly higher probability of being classified as responders compared to the low-stress condition, $b = 1.09$, $SE = 0.54$, $z = 2.02$, $p = .043$, $OR = 2.99$. Neither an effect of *Audio* nor an interaction between *Audio and Stress* was found for the cortisol response in responders only (see Figure 35, 10.2.) .

Table 7: Cortisol responder rate (in %) per experimental condition

	High Stress (n=39)	Low Stress (n=39)	
Externalized Auralizations (n = 53)	42 %	19 %	30 %
Non-Externalized Auralizations (n = 25)	42 %	15 %	28 %
	41 %	18 %	

7.4.3.2 Heart rate

A repeated-measures ANOVA including the measurements from all five time points revealed a significant interaction between *Time and Stress*, $F(4, 256) = 2.64$, $p = .035$, $\eta_p^2 = 0.04$, as well as a significant main effect of *Time*, $F(4, 256) = 31.61$, $p < .001$, $\eta_p^2 = 0.33$. When only including two measurements (as preregistered), the baseline and the individual peak heart rate after start of the VST, the model resulted in different significant effects. While *Time* also significantly affected heart rate, $F(1, 64) = 139.35$, $p < .001$, $\eta_p^2 = 0.69$; this was not found for the interaction between *Time and Stress*, $F(1, 64) = 0.16$, $p = .692$, $\eta_p^2 < 0.01$. Neither Stress, $F(1, 64) = 0.37$, $p = .547$, $\eta_p^2 < 0.01$; nor *Audio*, $F(1, 64) = 0.17$, $p = .679$, $\eta_p^2 < 0.01$; nor *Audio x Stress*, $F(1, 64) = 1.17$, $p = .283$, $\eta_p^2 < 0.01$; nor *Audio x Time*, $F(1, 64) = 3.778$, $p = .056$, $\eta_p^2 = 0.06$; nor *Audio x Time x Stress*, $F(1, 64) = 1.98$, $p = .164$, $\eta_p^2 = 0.03$; significantly affected heart rate. As illustrated in Figure 29, the difference between the two models might mainly be due to the preparation phase. During the preparation phase, the heart rate seems to decrease only in the low-stress group (sitting down and reading the answers)

while remaining relatively constant in the high-stress group (also sitting down but preparing the talk). This may reflect differences in cognitive demand during the preparation (Solhjoo et al., 2019). Again, contrary to our hypotheses, the increase in heart rate was not significantly higher in the externalized audio group (M = 14.8, SD = 11.5) than in the non-externalization group (M = 13.7, SD = 6.8, t[31] = 0.28, p = .392).

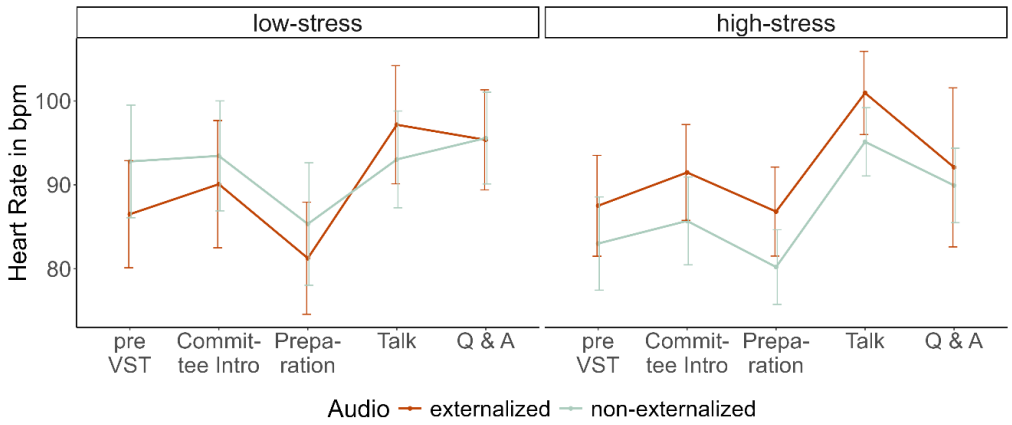

Figure 29: Mean heart rate in beats per minute in response to the VST as a function of stress and audio manipulation. Error bars indicate the standard error.

7.4.3.3 Subjective stress

A repeated-measures ANOVA including all five measurement times of stress ratings revealed a significant main effect of *Stress*, F(1, 74) = 5.90, p = .018, η_p^2 = 0.07; and *Time*, F(4, 296) = 15.65, p < .001, η_p^2 = .17. Also, a significant interaction between *Time and Stress* was found, F(4, 296) = 4.91, p < .001, η_p^2 = 0.06. As can be seen in Figure 30, subjective stress levels are higher and increase more continuously in the high-stress group.

As preregistered, only the baseline and the individual peak after stress instruction rating were included to test the hypothesis of a stronger subjective stress reaction in the externalized auralizations condition. This model again revealed a significant effect of *Time*, F(1, 74) = 99.34, p < .001, η_p^2 = 0.57, a significant interaction between *Time x Stress* F(1, 74) = 11.19, p = .001, η_p^2 = 0.13, and of *Time x Stress x Audio*, F(1, 74) = 5.05, p = .028, η_p^2 = 0.06. Neither a main effect of *Stress*, F(1, 74) = 1.80, p = .184, η_p^2 = 0.02; nor of *Audio*, F(1, 74) = 1.48, p = .227, η_p^2 = 0.02; nor an interaction between *Audio x Time*, F(1, 74) = 0.01, p = .938, η_p^2 < 0.01; nor between *Audio x Stress*, F(1, 74) = 0.20, p = .658, η_p^2 < 0.01; was found. Again, contrasting our hypotheses, the increase in subjective stress was not higher in the externalized auralizations group (M = 29.8, SD = 23.9) than in the non-externalized group (M = 42.3, SD = 13.3); t[37] = -1.69, p = .951, d = -0.59.

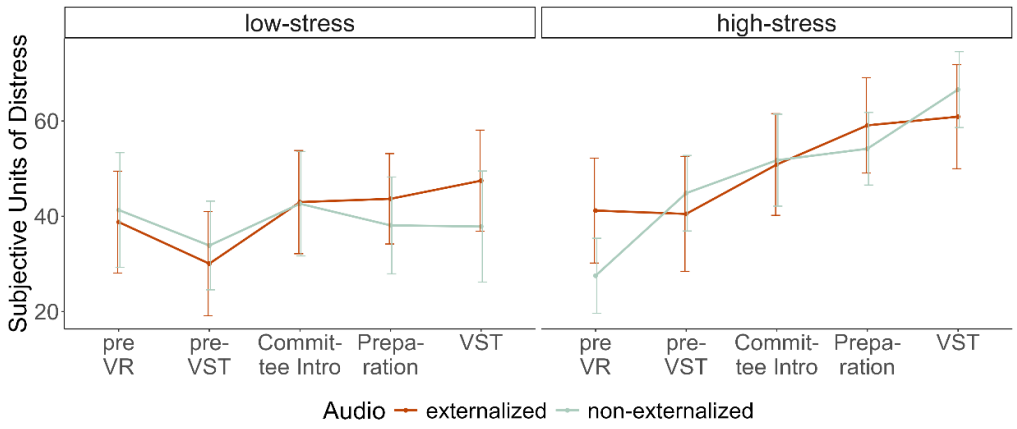

Figure 30: Mean stress rating as a function of stress and audio manipulation at five different measurement time points. Error bars indicate the standard error.

7.4.4 Gaze behavior

Only data from the high-stress group was analyzed due to the methodological differences (reading of preformulated answers in the low-stress group). Eye-tracking data from the question and answer phase were analyzed since offering 30 similar trials. In the externalized audio group, the mean latency of the first fixation (in ms) on the speaking agent was not significantly shorter (M = 1127, SD = 213) than in the non-externalized audio group (M = 1078, SD = 193); $t(20) = 0.69$, $p = .752$, $d = 0.24$.

Exploratorily, possible differences in the number of fixations on correctly identified speaking agents were investigated. Again, no significant difference was found between the externalized audio group (M = 25.6, SD = 3.7) and the non-externalized audio group (M = 24.2, SD = 8.0); $\chi 2\ (11) = 16.14$, $p = .136$.

7.4.5 Social anxiety

As preregistered, we analyzed whether the regressional weight of participants' social anxiety on social stress is differentially dependent on the externalized auralizations (and the stress manipulation), using three general linear models. Social anxiety, indexed by the continuous SPIN total score, was not a significant predictor of the increase in salivary cortisol, heart rate, or subjective stress from baseline to post or during VST measurement, nor was the interaction with audio or stress.

Furthermore, we exploratorily analyzed whether social presence, gaze, behavior, or stress response varied as a function of social anxiety and audio condition. For this purpose, a median split was conducted to classify participants as lower or higher socially anxious. Detailed results are provided in the supporting information. Briefly, for social presence, an interaction between social anxiety and auralizations was found, with higher socially anxious

participants reporting higher social presence, but only when externalized auralizations were used, see Figure 31. Moreover, higher socially anxious participants showed shorter latencies from speech onset until first fixation on the speaker, possibly reflecting hypervigilance (N. T. M. Chen & Clarke, 2017). Concerning stress indicators, a main effect of social anxiety (low vs. high) was found for heart rate and subjective stress, but no interaction effect between social anxiety and time or audio.

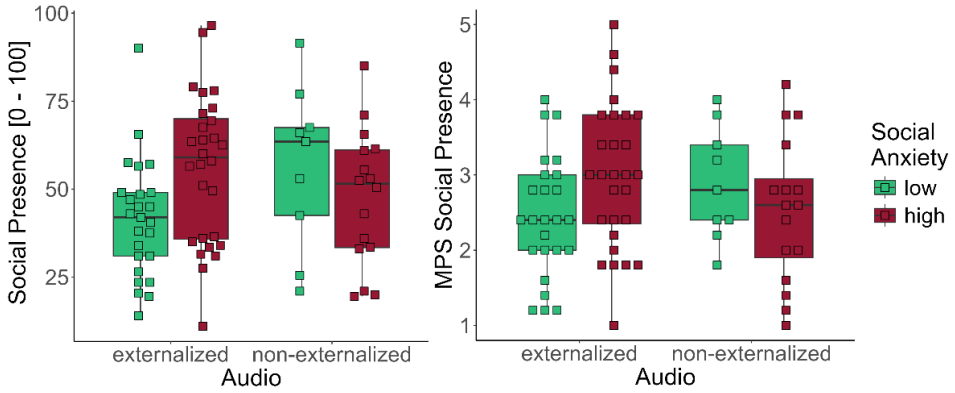

Figure 31: Social Anxiety and Audio. Social presence rating (on the left) and subscale of the MPS (right) as a function of audio and social anxiety (SPIN median split).

7.4.6 Equivalency of externalized auralizations

We expected non-superiority of auralizations based on individual HRIRs in comparison to generic HRIRs concerning all outcome variables. In addition to independent sample t-tests, we computed Bayes Factors on the probability of the equivalency hypothesis. These results were presented at the 51st German Acoustical Society meeting (DAGA) and published (non-peer-reviewed) within the conference proceedings (Roßkopf et al., 2025). As can be seen in Table 8, the Bayes Factors for H0 for all outcome variables are larger than 1, implicating at least anecdotal evidence in favor of equality versus a difference (Jeffreys, 1998). However, the critical threshold of 3 was only reached for social presence and heart rate.

Table 8: Means of outcome variables in individual and generic HRIRs audio conditions and the Bayes factor for an independent samples t-test of H0.

	Individual HRIRs (n = 26)	Generic HRIRs (n = 27)	Bayes Factor H0
Social presence (MPS)	2.62	2.84	6.21
Stress Response - Increase of			
Salivary cortisol	0.44	-0.26	1.73
Heart rate	13.22	17.78	4.53
Subjective stress	29.15	21.67	1.16
Visual spatial attention	1128	1073	1.97

7.5 Discussion

7.5.1 Summary

We investigated the effects of binaural auralizations which are perceived as externalized on presence and stress reactions in a virtual stress scenario. Participants completed a virtual job interview, which was either intended to induce high or low social stress, while the auditory scene was manipulated as externalized and realistic or as a non-externalized control condition. As intended, all three indicators for stress, salivary cortisol, heart rate, and ratings, reflected a response to the VST. Also, as expected binaural auralizations were perceived as more externalized and with higher acoustic realism compared to the diotic control condition. However, there was no effect of audio condition on social presence and neither on measures of stress response such as salivary cortisol, heart rate, subjective stress ratings, and visual spatial attention. Social presence was higher in the high-stress group and increased during the VST. Also, physical presence was found to increase with time spent in VR, or alternatively, as the VST progressed. Although neither social presence nor presence ratings were affected by the audio conditions, an interaction effect between audio and the participants' level of social anxiety level was found on an exploratory basis. Social presence was increased in participants with high social anxiety, but only when externalized auralizations were used. Also, the latencies of first fixations on agents after their speech onset were lower for highly socially anxious participants. Social anxiety (high vs. low) had a main effect on heart rate and stress ratings, but not on salivary cortisol. We also exploratorily evaluated the subjective quality of the acoustic scene and the VR experience via ratings. Interestingly, acoustic realism was positively correlated with both physical and social presence. A similar pattern was observed for acoustic presence. Furthermore, high stress decreased speech intelligibility and audio liking compared to low stress, while externalization increased speech intelligibility and audio liking compared to non-externalized sound. Last, evidence was gained that individualization of binaural auralizations is not superior to the use of generic binaural auralizations concerning all measured variables. To sum up, the present study does not support the claim that externalized binaural auralizations increase stress responses and social presence in a stressful virtual interaction. Instead, we identified specific relations between binaural auralizations and quality of experience in VR, as well as between interindividual differences related to social anxiety and stress responses.

7.5.2 Effects of binaural auralizations on virtual interactions

In contrast to our hypotheses, social presence did not differ between externalized and non-externalized auralizations. As outlined in the beginning, we expected increased social presence due to increased social realism when speech is perceived at the position where agents are located. Although the audio manipulation was effective, with binaural auralizations perceived as externalized and the respective virtual scene as more realistic, it had no measurable impact on presence. In a review, a positive effect of audio quality on social presence was summarized (C. S. Oh et al., 2018). Non-VR applications such as a first-person shooter video game (Skalski & Whitbred, 2010) and business-teleconferences (Christie, 1974) were investigated. In complex audiovisual scenes, the impact of externalized auralizations may be more limited. This may be particularly true for high-arousal scenarios (Diemer et al., 2015) like the VST. Indeed, in the current study, not only the high-stress group but also the low-stress group experienced increases in subjective distress (on average by 30 %), which was also found to affect social presence (Diemer et al., 2015). Differences in immersion were found to have larger effects on presence in non-emotion VR scenarios (Gorini et al., 2011). Possibly, (between-subject) differences in the acoustic scene may not be salient enough when embedded in a photorealistic visual scene in which demanding tasks must be accomplished. Furthermore, sound externalization was not task relevant. Under increased arousal, participants may have allocated their limited cognitive resources to threat-relevant stimuli, potentially the committee's neutral, feedback-free behavior. Attention is probably shifted towards nonverbal social rather than spatial cues. Notably, the TSST has proven effective even in teleconference formats without spatial co-presence of the committee, highlighting the task-irrelevance of spatial audio for eliciting social threat (Gunnar et al., 2021). Regarding the impact of binaural auralizations on social presence in complex audiovisual environments, findings are inconsistent. While one study found increased social presence with externalized auralizations in a virtual seminar room where participants had to localize the speaker (Roßkopf, Kroczek, Stärz, Blau, Van De Par, et al., 2024); another VR study involving dyadic problem-solving found no such effect, even when participants were encouraged to move around in order to experience spatial audio (Immohr, Rendle, Lammert, et al., 2024). In the former, externalization was task-relevant; in the latter, it was not.

Beyond the elevated arousal in our study and the task-irrelevancy of spatial audio, conceptual aspects of presence should be addressed to clarify the role of externalized auralizations in VR. The association between acoustic realism and physical presence was stronger than with social presence. This suggests that participants linked realistic sound more

to the overall VR environment than to the virtual agents. The more the speech was perceived as occurring in a real room, the stronger the reported sense of 'being there.' Thus, spatial audio may influence physical presence more than social presence. While the MPS social presence subscale was found to be sensitive to (large) social realism manipulations in a previous study, specificity to physical presence manipulations was low (Volkmann et al., 2019). This indicates that future studies should investigate the effects of spatial audio on experience in VR in a broader sense and concerning spatial presence.

Overall, moderate levels of social presence were found, with average ratings near the scale midpoint. While consistent with previous VR studies (Pfaller et al., 2021; Volkmann et al., 2019), this suggests that about half the participants lacked a clear sense of co-presence, indicating room for improving virtual social interactions. Alternatively, such levels may be expected when no deceptive instruction suggests interaction with a real human. Social presence encompasses an increasingly broad range of phenomena (Cummings & Wertz, 2023), and without deception, low ratings may reflect perceived non-human actorhood (Cummings & Wertz, 2023) regardless of audio realism. Future studies should complement the MPS with more specific items on salience, social realism, and involvement to better assess implementing binaural auralizations in virtual social interactions. Although integrating AI to simulate artificial humans may enhance interaction (Kroczek et al., 2025), it requires careful monitoring through refined social presence measures.

Next, we expected the VST to evoke stronger emotional responses due to increased social presence when immersivity is higher (by implementing externalized auralizations) as suggested by previous work (Felnhofer et al., 2019; Lønne et al., 2023; C. S. Oh et al., 2018; Slater, 2018). However, it was also found that not the presence of others, but rather the evaluative component of social presence, determined the response to stressors (Dickerson et al., 2008). As no audio effects were found on social or physical presence, nor visual attention, the absence of an effect on stress aligns with these findings.

7.5.3 VST paradigm

The present study presents a modified version of the TSST which is adapted for the manipulation of audiovisual VR. Our findings demonstrate a robust stress reaction in this paradigm which was observed on subjective, physiological, and neuroendocrine measures. Concerning subjective stress, as expected, the high stress group showed a higher increase than the low stress group, but unexpectedly, stress also increased in the low-stress group by up to 30% on average. Qualitative reports suggest that the reading-aloud task in front of the

committee also triggered social evaluative threat in some of the low-stress participants. Furthermore, the context of a job interview could have been generally perceived as stress-related. In contrast to subjective stress, cortisol stress increased in response to the VST selectively in the high-stress group, but decreased in the low-stress group. Neuroendocrinological measures may capture acute social stress and evaluative threat more distinctly and specifically than ratings.

In line with the literature the current VST resulted in overall cortisol responder rates of 41%; confer rates of 57% in a neuroimaging version of the TSST (Bärtl et al., 2024) and 33% to 86% for virtual and in-vivo TSST (Shiban, Diemer, et al., 2016). Similarly, the absolute cortisol increase (effect size of d = 0.41 in the high-stress group) aligns with prior findings; a meta-analysis (Helminen et al., 2019) reported average VR stress reactivity of d = 0.65 (range: 0.21–1.65).

Cortisol reactivity appears to be modulated not only by the virtual nature of the TSST but also by demographic variables, with greater responses typically observed in males and individuals under 25 (Goodman et al., 2017). In the present study, the predominance of female participants, despite hormonal control, may have resulted in reduced cortisol responsivity. On the other hand, the young age of the current participants and the high immersivity of the VST could have counteracted this effect (Goodman et al., 2017).

The present paradigm adapted the TSST to increase audio-visual components, and this resulted in deviations from the procedure in the traditional paradigm. The present VST lasted longer (6 min talk, 15 min Q&A) than the traditional TSST (5 min talk, 3 min arithmetic). However, we sampled salivary cortisol about 25 minutes (+ 40, + 55) after the onset of acute stress, which is within the best sample period (30 – 45 min), with peak responses occurring on average 38 min after TSST onset (Goodman et al., 2017). Furthermore, the TSST seems to be fairly robust to variations in the length of periods (Goodman et al., 2017). All in all, we provided a modified version of the VR-TSST, where stress induction is comparable to previous work and where a higher proportion of speech was held by virtual agents, allowing to investigate audio manipulations.

Furthermore, by implementing binaural auralizations, we provided a virtual acoustic environment which was superior to a non-externalized acoustic scene regarding all subjective audio quality features (externalization, liking, intelligibility, acoustic realism, acoustic presence, tone richness). Also, a naturalistic and realistic virtual scene was provided, which was confirmed by qualitative assessments and the VR ratings. Therefore, the current VST can be seen as a helpful tool for investigating the effects of audio (e.g., speech manipulations,

synthetic voices, spatialization) on stressful virtual interactions. However, the current findings suggest sound externalization has no substantial effect on stress response.

7.5.4 Stress and presence

While no main effect of the stress manipulation was found on the mean score of social presence, the effects become significant when the time point of measurement is taken into account. Social presence was higher in the high stress group and increased in both groups during the VST. As mentioned above, the low-stress group also reported increased subjective stress from pre- to post-VST. Also, physical presence increased throughout the VST. Therefore, the increase of presence over time may be due to increasing arousal, which was found to mediate presence (Diemer et al., 2015). It is not only stated that presence is the basis on which a VR scene results in "real" emotions (Peperkorn et al., 2015), but also the other way round was found. When arousal is induced, e.g., by displaying a phobia-relevant stimulus, presence in turn increases. The stronger participants' actual emotional experience is in VR, the more presence they report. Indeed, our supplementary time-lagged correlation analyses indicate effects in both directions. An alternative explanation would be that the more time is spent in VR, the higher the feeling of being there and the feeling of being with others. Since the experimental manipulation of stress level affected (social) presence in the expected direction, arousal and stress are suggested as mediators for higher presence.

7.5.5 Social anxiety

This implies that participants who react more adversely towards socially stressful situations, meaning participants with high levels of social anxiety, also report more (social) presence. Nonetheless, we found no consistent relationship between social anxiety and social presence. On an exploratory basis, social presence was higher in high socially anxious participants but only in the externalized auralizations condition. These results should be interpreted cautiously, but may indicate a need for investigating differential effects of audio externalization depending on traits.

Concerning stress indicators, heart rate and subjective stress were influenced by social anxiety, whereas cortisol was not. This again implies that these response domains may differentially reflect specific aspects of stress. Cortisol again seems to be a more specific indicator of the biological stress reaction, whereas heart rate may reflect the stressor itself even more. Subjective stress appears as an adequate and sensitive measure of how individuals experience a (social) situation. Indeed, a blunted cortisol stress reaction was found for patients with social anxiety disorder, while subjective stress reports were increased (Klumbies et al.,

2014). This dissociation between subjective and cortisol stress reaction may also manifest in participants with varying levels of sub-clinical social anxiety as investigated in the current study. However, this pattern emerged only in the exploratory analyses and not in the preregistered models. These discrepancies likely stem from methodological differences: the preregistered analyses treated social anxiety as a continuous predictor, whereas the exploratory analysis relied on a median split, which may have distorted effect estimates.

7.5.6 Limitations and future research

The main goal of the study was to investigate the effects of binaural auralizations on presence and specifically social presence, and subsequently social stress and behavior (gaze) in virtual social interactions that induce social evaluative threat. Hence, a specific focus was set on VR applications in the context of social fear. While on an exploratory basis, acoustic realism was correlated with presence, the experimental manipulation of sound externalization had no effects. This implies that in stressful virtual interactions, the implementation of spatialized sound does not make an important contribution to the effectiveness of the scenario. On the one hand, this finding is surprising since audio quality was found to enhance presence (Freeman & Lessiter, 2001; Kern & Ellermeier, 2020) and social presence. On the other hand, to the best of our knowledge, this is the first study to examine effects of externalized auralization on stressful virtual social interactions. Furthermore, previous work indicated a limited effect of increased immersivity on presence in virtual scenarios which induce high levels of arousal (Gorini et al., 2011). A similar relation can now be assumed for social presence. In our stressful VR application, no effect of increased immersivity in terms of more realistic and spatial audio was found. Although our study design included a low-stress control condition, this group similarly reported a 30% increase in stress. The problem of designing an appropriate “placebo TSST”, which includes a comparable task but without social evaluative threat, has been discussed before (Het et al., 2009). Often, the social component is removed by omitting the committee. This was not feasible in the current study, given the research goal of investigating audio effects in different stressful social interactions. Therefore, future studies should investigate the effects of binaural auralizations in socially relevant but less stressful situations and further social contexts. Furthermore, specific items for salience, social realism, and involvement should be used (Cummings & Wertz, 2023), as outlined above. Also, contexts in which a stronger influence of the room acoustics can be assumed should be investigated – e.g. concert halls for musicians with stage fright or auditoria for students with public speaking anxiety.

Since sound was found to attract attention to speakers (Foulsham & Sanderson, 2013), we also expected differences in visual attention depending on the spatiality of sound. Since only in the externalized audio conditions, the direction of the sound source – and therefore a cue about the speaking agent – can be perceived at the moment of the sound onset, we expected a shorter latency of first fixation on speaking agents for this group, and furthermore, sustained attention. The fact that we did not find these effects could be due to the lip synchronization of the agents. These were located in front of the participants, and all of them were within the field of view. It might be that the visual information was so effective that the additional auditory information was not relevant (e.g., ceiling effect). Future studies should evaluate the effect of externalized auralizations on visual attention in virtual interactions in which the speaker location is not immediately visually apparent, making externalized audio more task-relevant.

7.6 Conclusion

We investigated the effect of sound externalization by implementing binaural auralizations in a stressful virtual social interaction. While the VST was efficient in evoking a stress response and the binaural auralizations were shown to be highly realistic and externalized, no audio effects on social presence, stress induction, and visual spatial attention were found. Exploratorily, acoustic realism correlated with presence, and social anxiety interacted with the effects of externalized sound. Implementing spatial sound may not be needed in VR applications in the context of social fear, but it may enhance the realism and the acoustic quality of the virtual environment. Strong evidence is gained that individualization of binaural auralizations is not needed for virtual social interactions. Overall, only medium levels of social presence indicate a need for improvement of the virtual social interactions and a further systematic investigation on factors determining the feeling of being with another when interacting with artificial humans. Future studies should investigate the effects of binaural auralizations on social presence and behavior in virtual social interactions in which sound spatialization may be more salient, task-relevant and crucial for visual spatial attention, and with social presence measurements tailored for interactions with artificial humans.

8 General discussion

8.1 Summary

This dissertation project was guided by the overarching objective of examining the impact of audio renderings on socio-cognitive processing, presence, and affective responses within social virtual environments. Underexplored aspects of the interaction between auditory and visual immersion were addressed, with a particular focus on enhancing social presence and thereby improving the effectiveness of virtual social interactions.

Specifically, **Study 1** investigated how the localization of physical sound sources is affected by the visual scene presented via an HMD and the measurement method employed. Results indicated that audiovisual mismatches impaired localization accuracy, although physical presence and perceived audio quality remained unaffected. Additionally, a relationship was observed between visual distance compression in HMDs and auditory distance perception, suggesting that virtual psychoacoustic experiments should account for VR-specific visual distortions. The study further discussed the strengths and limitations of the measurement methods used, emphasizing that the specific goals and requirements of each study should guide method selection. Overall, the findings underscore the importance of considering the interaction between visual and acoustic elements in VR environments.

Study 2 explored the extent to which virtual sound sources approximate physical ones. Localization accuracy was compared between several HTBAs, loudspeakers, and a reference condition, serving as an indicator of auditory realism. Participants also rated social presence and perceived audio quality. No significant differences were found between physical sound sources and HTBAs in terms of externalization, azimuthal localization, perceived audio quality, or social presence, whereas the reference condition performed significantly worse. Interestingly, physical sound sources led to greater distance misperception, which may be best explained by the findings of Study 1, suggesting that visual displays distort the perception of physical sound sources. In total, the findings imply that enhancing auditory immersion can increase social presence in VR, thereby improving the quality of virtual social interactions.

Study 3 examined whether advanced TTS synthesis can match human voice recordings in eliciting stress, presence, and anxiety during a socially stressful virtual interaction modeled after the TSST. The use of synthetic speech offers potential benefits for flexibility and standardization in social VR applications. Comparable subjective and physiological stress responses were found for human voice recordings and TTS synthesis. Also, no differences were observed in physical or social presence between the audio

conditions. A positive correlation emerged between presence and subjective stress, and participants' levels of social anxiety influenced their perception of the virtual interaction. These results suggest that TTS systems can serve as effective and practical alternatives to natural human speech in virtual socio-emotional paradigms designed to induce social-evaluative threat.

Study 4 investigated the impact of audiovisual realism within two virtual social stress paradigms, both based on the TSST, and designed to elicit either high or low levels of stress. The study compared externalized HTBAs with non-externalized diotic auralizations for synchronizing the virtual agents' speech. Although HTBAs were rated higher on subjective realism, acoustic presence, and audio preference, these differences did not translate into effects on social presence or affective responses, regardless of stress level. Nevertheless, social and physical presence were positively associated with ratings of acoustic quality. A positive relationship between presence and stress was again observed, with tentative evidence suggesting that higher social presence during the initial interaction phase was linked to stronger stress responses during the subsequent stressor phase. The stress scenario reliably induced objective and subjective stress responses, including elevated salivary cortisol, heart rate, and subjective stress increases. Notably, even the low-stress condition evoked substantial subjective stress, which may have attenuated the influence of the audio condition. These findings suggest that in virtual social interactions designed to elicit strong affective responses, high auditory immersion may not be a critical factor.

8.2 Integration of findings

In the following, the central conclusions derived from the research project examining the impact of audio renderings on socio-cognitive processing, presence, and affective responses are presented and discussed. The first study established an essential basis for auditory research in VR by examining how the visually presented virtual scene influences auditory processing. The findings highlight that both the visual context and the measurement metrics must be carefully considered when conducting auditory experiments using HMDs. Also, the findings align well with classical non-VR literature on the effects of multimodal integration, such as the ventriloquist effect (Jack & Thurlow, 1973). Moreover, the importance of integration effects in the virtual acoustics is underscored. Prior work has shown that the visual context in which virtual sound sources are presented significantly affects auditory perception (Werner et al., 2016). For instance, the perception of externalized sound can be disrupted by incongruent visual scenes.

Study 1 extends this by demonstrating that visual context also influences sound source localization. Interestingly, while incongruent visual scenes reduced localization accuracy, they did not affect perceived social presence or acoustic realism, suggesting that these dimensions are more robust to cross-modal inconsistencies. This finding is consistent with previous research showing that incongruent visual cues can reduce scene plausibility without impairing embodiment or physical presence (Mal et al., 2023). Presence is considered to result not only from the illusion of being situated in a virtual environment, but also from the plausibility of the experience, which is supported by sensorimotor contingencies, such as the expectation that turning the head leads to a corresponding change in the visual or auditory scene (Slater, 2009). In line with this, plausibility in VR has been described as a multi-layered construct processed across cognitive, perceptual, and sensory levels (Brübach et al., 2022), which helps to explain why certain violations of physical realism may reduce plausibility without necessarily diminishing the sense of presence. Similarly, cognitive manipulations (e.g., narrative framing) can only partially compensate for such violations. Given that the study addressed presence and subjective audio quality rather than plausibility, the results are in line with previous findings.

The second study confirmed the high perceptual realism of several binaural auralization methods. Notably, the least resource-intensive approach, simulated rendering based on generic HRTFs, performed comparably to the more complex approaches. The only exception was a minor difference in one sub metric, social presence observed in the first block, where individual simulated HTBAs showed a slight advantage. This is a significant step toward the practical application of HTBAs, as individualized HRTF measurements are time-consuming, require specialized expertise, and often depend on elaborate measurement setups. Although various approaches have been proposed to simplify or automate the individualization of HRTFs (Guezenoc & Seguier, 2018), the need for individualized HTBAs would pose a substantial challenge to their adoption in psychological and clinical research.

Although there is a growing body of evidence supporting the efficacy of VR in clinical and therapeutic contexts (Wechsler et al., 2019), including its integration into German clinical guidelines for anxiety disorders (Bandelow et al., 2015, 2022), its implementation in practice remains limited. For example, only 10% of the 690+ clinicians surveyed reported using therapeutic VR (Felnhofer et al., 2025). Technological complexity and insufficient training are among the main barriers. Considering these findings, placing additional demands on clinicians, particularly regarding acoustic individualization in VR applications, is difficult to

justify. The finding that individualization is not necessary is therefore encouraging, especially in the context of efforts to simplify HTBA implementation.

In this regard, also the demonstrated equivalence between measured and simulated BRIRs is of relevance. HTBAs based on measured BRIRs require time-consuming acoustic recordings and specialized expertise within the target environment, which limits their practical use in psychological research. Additionally, they restrict applications to existing physical spaces, reducing both flexibility and generalizability of the audio renderings.

A key finding of Study 2 is that increased externalization in audio renderings led to higher ratings of both social presence and realism. This result provided a foundation for subsequent investigations into the application of HTBAs in complex social interactions within VR. An interim conclusion is that audiovisual virtual scenes employing advanced rendering techniques can achieve high levels of perceptual realism. Specifically, HTBAs appear to enhance social presence, thereby reinforcing the established notion that the immersive quality of VR positively influences presence, the feeling of "being there" (Slater & Wilbur, 1997). This finding extends previous research to the clinically relevant dimension of social presence.

Alternatively, one could argue that externalized auralizations offer greater social realism than non-externalized ones, thereby contributing to increased social presence, as previously suggested in a review (C. S. Oh et al., 2018). The accuracy of sound source localization serves as an indicator of auditory scene realism, and close-to-real sound source localization was observed. Notably, no significant differences were found between fully virtual audiovisual environments and those combining virtual visuals with a physical audio scene in terms of realism, social presence, and, to some extent, sound source localization. This suggests that one sensory modality can be effectively substituted by virtual stimulation. However, VR-specific characteristics, individual differences, and the strengths and limitations of measurement methods should be considered. Interestingly, participants demonstrated lower accuracy in estimating the distances of physical sound sources (i.e., loudspeakers) than of virtual ones. This finding, although initially counterintuitive, aligns with results from Study 1, which also revealed a consistent overestimation of physical sound source distances, particularly when using the placement task. The combination of a physical audio scene with a compressed virtual visual scene appears to contribute to this overestimation (compare Study 1). In contrast, virtual sound sources were perceived as closer to their actual positions, possibly due to a shortened distance impression. Alternatively, since both visual and auditory distance perceptions are based on illusion, a more coherent integrated percept may emerge when both sensory inputs are virtual.

Across Studies 1 and 2, four different localization methods were employed, each with distinct advantages and limitations. The eye-tracking task arguably offers the most naturalistic approach, as it reflects typical behavior such as looking at the speaker (Foulsham & Sanderson, 2013). However, eye-tracking data are susceptible to measurement errors and may vary due to individual differences in eye-region anatomy and device calibration. The placement task was the most frequently used method (in three out of four experiments) and was well accepted by participants. Its interactive nature likely contributed to enhanced presence, consistent with prior literature (Slater & Wilbur, 1997). At the same time, it appears to require a high degree of audiovisual integration, which could render it especially susceptible to VR-specific factors. The walking task also promoted interactivity and presence but was the most demanding in terms of physical space and effort, and it induced the highest levels of simulator sickness. The verbal estimation task, while less affected by the visual scene, was comparatively imprecise as indicated by the largest standard deviations and was mainly disliked by participants.

In both Study 1 and Study 2, the role of social anxiety in sound distance localization was examined as a factor in socio-cognitive processing. It was hypothesized that individuals with higher levels of social anxiety would perceive speakers as being closer. This assumption was based on prior findings indicating that phobic stimuli are processed differently, leading to perceptions of increased size (Shiban et al., 2016) or proximity (Givon-Benjio & Okon-Singer, 2022). However, no significant effect of social anxiety on the estimation of egocentric distances of speech sound sources was observed. Similarly, a study investigating the perception of angry voices found that such stimuli were localized at greater distances than neutral voices, potentially due to top-down knowledge about vocal effort and loudness (Kroczek et al., 2024). The lack of a social anxiety effect in that study aligns with the present findings, suggesting that this factor is of limited relevance in auditory distance perception.

The findings of Study 3 are particularly relevant for advancing research on virtual social interactions. To the best of my knowledge, this study is the first to systematically and experimentally demonstrate the equivalence of human voice recordings and AI-synthesized speech in eliciting social presence and affective responses during socially stressful virtual interactions. Notably, AI-generated speech proved effective in a context involving social-evaluative threat, which typically evokes complex emotional and interpersonal reactions. These results significantly facilitate the design and implementation of future experimental paradigms in both basic and applied psychological research addressing practicability concerns in VR.

The relevance of these findings is highlighted by the rapid proliferation of human–AI communication in recent years. Since the release of ChatGPT in November 2022, AI has become increasingly embedded in everyday life. Within just two months, the platform had attracted over 100 million users (Saini, 2023). This development reflects a broader trend: prior positive experiences with AI are associated with more frequent use, which in turn contributes to the growing integration of AI technologies into daily routines and professional contexts (Wang et al., 2025). Importantly, it has been shown that human–AI interactions often follow the same psychological principles as human–human interactions (Kroczek et al., 2025), further supporting the ecological validity of using AI-generated speech in social psychological research. Recent advances in large language models have enabled the development of conversational agents capable of engaging in dynamic, face-to-face interactions within VR. Using verbal prompts, the personality traits of conversational agents can be flexibly adjusted, enabling naturalistic social-emotional processing and behavior during virtual interactions. One particularly promising application of such agents in clinical psychology is in e-mental health.

In Germany, digital health applications (Digitale Gesundheitsanwendungen, DiGA) have been integrated into the statutory healthcare system since 2020 and can be prescribed by physicians or psychotherapists (Schmitz et al., 2025). Numerous applications have demonstrated clinical efficacy, including those targeting anxiety disorders such as social anxiety (Schreiter et al., 2023). These interventions often include psychoeducational components, providing users with evidence-based information about the development and maintenance of mental disorders, as well as treatment options and mechanisms of change. For such purposes, as well as for structured diagnostic interviews, conversational agents equipped with TTS synthesis offer a valuable tool for bridging the gap between static information and interactive, user-centered communication. Conversational agents have already shown promise in educational contexts (Gillies & Pan, 2018), and their potential in mental health education warrants further exploration. A recent review found that while such agents are widely used in mental health apps for psychoeducation, symptom assessment, and counseling, AI technologies are rarely employed. Most interactions are based on fixed algorithms and follow pre-scripted dialogue paths (Parmar et al., 2022). This highlights the potential of AI to enhance flexibility, personalization, and user engagement, all factors that may improve the efficacy and acceptance of digital interventions. Nevertheless, the integration of AI into clinical applications must be accompanied by strict adherence to data protection regulations to ensure user safety and prevent misuse. Ethical considerations, particularly regarding

transparency, informed consent, and data privacy, remain essential in the development and deployment of AI-based tools in healthcare.

In conclusion, study 3 revealed a fundamental insight: AI-synthesized and human-recorded speech are equivalent in eliciting social and emotional responses. From a speech and audio perspective, there appear to be no substantial barriers to integrating TTS into clinical applications such as virtual exposure therapy, conversational agent interactions, or digital health interventions. These findings support the continued exploration and responsible implementation of AI technologies in psychological research and practice, as well as the practicability of implementing audio renderings in virtual interactions.

Studies 3 and 4 both contribute to understanding the role of audio renderings in socially demanding virtual interactions, with Study 3 focusing on the viability of TTS audio and Study 4 exploring the potential advantages of HTBAs. Building on findings from Study 2, which demonstrated that HTBAs can enhance realism and social presence in VR, the seminar room scenario served as a basis for exploring HTBA in more complex virtual social interactions involving social evaluative threat relevant to the treatment of social anxiety. For this purpose, Study 4 employed the simplest, most effective rendering method identified in Study 2: simulated HTBAs, which yielded highly externalized hearing impressions. Additionally, it introduced a comparison between individual and generic HTBA to assess whether previously observed differences in social presence could be replicated.

The results from Studies 2 and 4 suggest that HTBAs, including generic HRTFs, are entirely sufficient for virtual social scenarios involving speech-based audio stimuli. Only one minor difference was observed, a slight advantage of individual over generic HTBAs on one social presence rating, which could not be replicated across other measures. The evidence strongly supports that generic HTBAs can achieve high localization accuracy, realism, and externalization.

The superiority of externalized over non-externalized auralizations in terms of social presence, which was observed in Study 2, was not replicated in Study 4. Several factors may account for this discrepancy. In Study 2, the audio condition was manipulated using a between-subjects design. Multiple trials presented audio conditions in randomized order, allowing participants to compare different auralizations directly. The primary task involved sound source localization, and ratings of social presence and acoustic realism were collected pseudo-randomly after several trials. This design is well-suited for detecting subtle differences between rendering methods. However, it likely heightened participants' sensitivity to such differences, as the study's focus on audio rendering may have led them to pay close attention

to even minor variations. Notably, the social presence rating specifically targeted the subjective impression that a present person could have spoken to the participant. Thus, while Study 2 provided strong evidence that externalization can influence social presence, this effect may be too subtle to manifest in more complex scenarios such as those used in Study 4. This assumption is further supported by the fact that Study 4 employed a between-subjects design for audio manipulation, meaning that participants were exposed to only one audio condition throughout the scenario. As a result, they had no opportunity to directly compare or contrast different auralizations, which may have reduced their sensitivity to subtle acoustic differences.

Although the visual room models in both studies were similarly detailed, featuring photorealistic surfaces and consistent light-shadow rendering techniques, the interaction scenario in Study 4 was considerably more dynamic. Participants moved through various positions within the virtual room and responded to questions posed by virtual agents. These agents were also visually more realistic. Study 4 agents were created using MetaHumans Creator (Epic Games), whereas Study 2 relied on alternative software (MakeHuman) that offered lower visual realism. In Study 2, virtual agents were passively placed in the room using a controller to indicate sound source locations, resulting in low social realism. In contrast, the agents in Study 4 exhibited socially realistic behavior, including synchronized lip movements and turn-taking cues. In the eye-tracking paradigm of Study 2, agents remained static and wore face masks to avoid unrealistic lip movements, which could have revealed the sound source location. This design choice in Study 2 likely increased the potential impact of auralization, as other cues indicating the co-presence of human-like beings were minimized.

Previous research that reported effects of audio quality on social presence often lacked either visual components or interactive elements. In some instances, no visual stimuli were presented at all (Christie, 1974), while other studies did not incorporate any form of social interaction (Freeman & Lessiter, 2001; Skalski & Whitbred, 2010). More recent work found no enhancement in perceived social realism during videoconferencing when surround sound systems were compared to conventional two-channel setups (Nowak et al., 2023). Likewise, studies investigating binaural auralizations in complex virtual social scenarios did not reveal significant effects on social presence (Immohr et al., 2023; Immohr, Rendle, Kehling, et al., 2024; Immohr, Rendle, Lammert, et al., 2024). Notably, these studies involved interactions with real individuals, suggesting that mere awareness of engaging with actual humans may be sufficient to evoke a strong sense of social presence, regardless of acoustic fidelity. It is conceivable that individuals have become accustomed to communicating via media that offer

limited acoustic realism, such as telephony, voice messaging, or even text-based formats. This habitual exposure may diminish the perceptual relevance of audio quality in shaping social presence within virtual environments.

However, in my view, the most critical factor explaining the differing results between Study 2 and Study 4 is the level of arousal experienced by participants. Although arousal was not directly measured during these experiments, indirect indicators suggest that arousal in Study 2 was low to moderate. Pre- and post-experimental assessments of affect revealed no increase in negative affect, while positive affect showed a slight rise. The localization task used in Study 2 did not seem to elicit socially or emotionally charged responses, and participants reported feeling relatively unaffected afterward. In contrast, Study 4 elicited significant emotional responses. Participants in the high-stress condition exhibited stress reactions across multiple domains, and even those in the low-stress condition reported elevated stress levels. This outcome reflects a common challenge in stress research: designing a paradigm that avoids social-evaluative threat while maintaining consistency across other experimental variables.

Previous research has shown that the immersive quality of virtual environments may be less influential under conditions of heightened anxiety (Gorini et al., 2011). Affective states, particularly arousal and fear, are closely intertwined with the experience of presence (Diemer et al., 2015). Initial increases in fear have been found to enhance presence, which in turn amplifies fear responses to virtual phobic stimuli (Peperkorn et al., 2015). Evidence for this dynamic also comes from studies involving socially anxious individuals, who report higher levels of social presence during virtual social interactions and cue exposures compared to healthy controls (Felnhofer et al., 2019). In the present research project, a similar association between social anxiety and presence was observed. However, this relationship was not consistently replicated across all measurements. This inconsistency is likely due to the methodological approach: rather than comparing clinical and non-clinical groups, the studies assessed social anxiety as a continuous variable without applying diagnostic thresholds. Consequently, the findings reflect correlational patterns within a subclinical or non-clinical spectrum of social anxiety.

Finally, revisiting the conceptual framework introduced in Figure 3, findings provide evidence for the expected relationship between realism and quality of experience, particularly for social presence, but also for user preference, perceived realism, and other subjective qualities such as acoustic preference. Localization accuracy of HTBAs was comparable to real sound sources, underscoring a high degree of realism. Moreover, HTBAs have been found to

enhance social presence, particularly when directly compared to non-externalized auralizations and in scenarios with relatively low levels of social interaction and arousal. In contrast, the impact of audio renderings appears limited in emotionally intense and stressful virtual social interactions, as in Studies 3 and 4. The results do not support the hypothesized effects of social anxiety on auditory distance perception and only partly support its influence on social presence. More broadly, interrelations between arousal and presence were observed, and the enhancing effect of social evaluative threat and arousal on social presence could be replicated. Finally, by examining the impact of audiovisual integration and practicability concerns in VR, this thesis provides implications and considerations for future research and applications in clinical psychology, such as the treatment of social anxiety.

8.3 Strengths and limitations

The following section presents a reflection on the strengths and weaknesses of the dissertation project. One of the significant strengths of this work lies in its interdisciplinary approach. From the outset, the project was informed by a broad literature review that extended beyond psychological sources to include key publications in VR and virtual acoustics. Core texts and studies provided by project partners were instrumental in building a solid understanding of core technological principles, not only of binaural hearing but also of signal processing and rendering techniques. This interdisciplinary grounding enabled a more nuanced interpretation of the findings, particularly in Studies 1 and 2, which would not have been achievable within a purely psychological framework.

It also enabled the use of innovative and complex methods, such as the setup of an HRTF measurement system within our Department of Clinical Psychology and substantial enhancements of the virtual social scenarios. This level of technical immersion required expertise beyond psychology and is reflected in the publication of Studies 1 and 2 in reputable journals in computer science and acoustics. While the project is interdisciplinary, with Study 1 most closely aligned with VR and computer science, and Study 2 focusing on advanced acoustic methods beyond the typical scope of psychology, the findings remain highly relevant to psychological research. It is therefore hoped that these insights will be effectively disseminated across all three disciplines.

The final two studies focused more explicitly on psychological constructs, methods, and applications. However, the quality of their experimental setups and manipulations would not have been achievable without interdisciplinary collaboration. This underscores a central strength of the dissertation: its integration of expertise from psychology, particularly cognitive

and clinical psychology, as well as acoustic engineering and VR/computer science. Studies 2 and 4 would not have reached their current level of sophistication without the involvement of acousticians.

A further key strength of the research project lies in its broad methodological scope. The localization studies employed four distinct measurement techniques, which were not only applied but also, in part, validated. Across all experiments, social presence was assessed both retrospectively using standardized questionnaires and concurrently during the VR experiences via single-item rating scales derived from validated instruments and implemented within the virtual scene. Although not all effects were consistently reflected in both types of measures, the observed trends were largely congruent, supporting the reliability of the findings. This dual approach also enabled an analysis of the temporal dynamics of spatial and social presence.

In Study 4, time-lagged correlation analyses were conducted to explore potential antecedent and subsequent factors in the relationship between arousal and presence. The results aligned well with existing literature, confirming the hypothesized bidirectional relationship between arousal and presence (Diemer et al., 2015). Furthermore, exploratory evidence suggested that higher levels of social presence during initial virtual interactions may contribute to increased arousal or social fear in subsequent phases. While this specific directionality has not, to the best of current knowledge, been previously demonstrated for social presence, it resonates with earlier findings indicating that temporal dynamics shift throughout a fear-inducing virtual scenario (Peperkorn et al., 2015). It has been shown that physical presence during early spider phobia exposures predicts subsequent fear responses, which the present findings support and potentially extend to encompass social presence and social anxiety.

A further strength of this project is that, in addition to sound source localization and qualitative aspects of the VR experience, affective responses were assessed using a diverse methodological toolkit. Standardized questionnaires were complemented by single-item rating scales tailored to specific affective variables, enabling near real-time assessment in VR and a more detailed analysis of affective processes. Beyond subjective measures, physiological and neuroendocrinological data were collected. Muscular activity in the trapezius muscle, associated with mental stress and fear (Pribék et al., 2021; Wijsman et al., 2013), was recorded, alongside heart rate, which typically increases under stress and cognitive load, though it is subject to comparably large variability (Santl et al., 2019; Solhjoo et al., 2019; van Dammen et al., 2022). Importantly, salivary cortisol levels were also measured. Despite

being sensitive to circadian rhythms, gender, medication, and other factors (Goodman et al., 2017), all of which were carefully controlled or excluded, cortisol proved to be a particularly informative biomarker. Also in study 4, it effectively distinguished between the high-stress and low-stress groups, a differentiation not as clearly reflected in the subjective or heart rate data. Cortisol is a central neuroendocrinological marker that links chronic stress to various psychiatric and psychological disorders (Grace et al., 2022; Zorn et al., 2017). The integration of these diverse methods and interdisciplinary expertise substantially enhances the validity of the findings and supports their generalizability across multiple research domains.

Nevertheless, certain limitations must be acknowledged. The sociodemographic composition of the sample constrains the generalizability of the results. Although the overall sample size of 227 participants is robust and power analyses were conducted to ensure adequate sensitivity for the research questions, the sample was predominantly composed of university students, with a substantial proportion studying psychology. The median age was in the early twenties, and the majority of participants were female, rendering the sample non-representative of the general population. While it is unlikely that sociodemographic factors significantly influenced the core finding that visual scenes affect auditory distance perception, they may have impacted participants' general ability to engage with VR and the quality of their VR experience (Méndez et al., 2025). Also, stress responses and social anxiety are known to be modulated by demographic variables such as gender and age (Allen et al., 2017; Goodman et al., 2017; Santl et al., 2019). These variables were addressed through gender-based randomization procedures. Nonetheless, the interaction between the stressfulness of the virtual scenes and the acoustic manipulations may have manifested differently in this young, predominantly female academic sample than in a more representative population.

From a technological perspective, one notable limitation of the project was the use of the Genelec loudspeakers' directivity as a simplified representative for a human's directivity in the auralizations of the virtual agents. This approach was based on procedures established by the Oldenburg project partners and was also necessary for Study 2, in which these loudspeakers served as a comparison condition. However, the intended goal, particularly in Study 4, was to evoke auditory impressions resembling human speech, as if produced by a physically present speaker. It is plausible that acoustically sensitive individuals may have perceived the loudspeakers' specific acoustic signature. The extent to which auralizations must be refined to simulate a human speaker rather than a loudspeaker remains a subject of ongoing debate in the field of acoustics. For virtual loudspeakers, equivalent auditory impressions have been demonstrated, particularly for speech stimuli (Blau et al., 2021;

Brinkmann et al., 2017; Stärz et al., 2022). A particularly challenging aspect of simulating human speech is the dynamic nature of directivity. Human speakers are rarely static while speaking, resulting in time-varying changes in sound radiation. Additionally, directivity varies across phonemes due to articulatory movements, such as lip modulation, which in turn affect acoustic radiation patterns (Pörschmann & Arend, 2020). Research has investigated whether acoustic properties, such as directivity, can be adjusted to resemble a human speaker rather than a loudspeaker. The evidence remains mixed. One study concluded that optimal auralization of speech in VR should incorporate interindividual differences and articulation-dependent characteristics (Pörschmann & Arend, 2020). However, these conclusions were based on signal processing analyses, and it remains unclear whether the identified differences are perceptually relevant. Conversely, there is evidence suggesting that speech directivity may have a limited impact in interactions with embodied conversational agents. For instance, incorporating head orientation into directivity simulations did not significantly affect perceived social presence or the realism of agent voices (J. Wendt et al., 2019). These findings suggest that while acoustic fidelity is essential, certain perceptual aspects of social interaction in VR may be robust to variations in directivity modeling.

A further shortcoming is the use of proprietary AI-based software alongside open-source tools and custom-made software. A literature review revealed that many of the highest-quality TTS systems are commercial products. While this approach does not fully align with open-science principles, it provided a validated, user-friendly, and efficient method for generating highly naturalistic synthetic voices for Study 3. With additional time and effort, it might have been possible to employ open-source alternatives developed by academic institutions or non-profit organizations. A similar consideration applies to the visual assets used in the virtual environments. To ensure visual fidelity and efficiency, high-quality commercial assets were selected, a choice that limits adherence to open-science principles.

Another limitation concerns the observed levels of social presence. One of the central aims of this dissertation project was to investigate the role of externalized auralizations in virtual social interactions, with social presence serving as a core variable. Overall, moderate levels of social presence were observed, with mean scores clustering around the midpoint of the scale. The present levels are comparable to previous VR studies (Pfaller et al., 2021; Volkmann et al., 2019), yet they also indicate that a substantial proportion of participants did not experience a strong sense of social presence. This points to the potential to enhance social interaction in VR. At the same time, such outcomes may be typical in scenarios where participants are not led to believe they are interacting with a real person. Given the evolving

conceptualization of social presence, lower ratings, particularly in the absence of deceptive cues, may reflect the perception of engaging with a non-human agent, even when the auditory realism is high.

While the strength of the project lies in its interdisciplinary expertise and broad methodological spectrum, this diversity also introduces certain limitations. The research yielded numerous relevant findings for auditory VR in the context of social interaction. However, the transition from the perceptually oriented studies in the first half of the project to the more specific, stress-inducing paradigms in the latter half may be considered relatively abrupt. While this dissertation addresses essential research questions, many opportunities remain to examine intermediate steps and smoother transitions between these domains. These gaps present promising directions for future investigations.

8.4 Future research

Based on the findings of Studies 1 and 2, auditory distance perception and sound source localization in VR appear to be complex and, to some extent, differ from real-world perception. Indeed, auditory distance perception has been shown to vary depending on whether virtual or physical sound sources are presented (Stodt et al., 2024). However, the underlying mechanisms of audiovisual integration in VR remain insufficiently understood. Systematic manipulation of both visual and acoustic parameters could provide deeper insights into these processes. While Study 1 employed real loudspeakers and Study 2 combined real loudspeakers with virtual sound sources, the visual scene was always virtual. Future research should therefore investigate comprehensive comparisons of virtual and physical environments across visual and auditory dimensions. In particular, examining auditory distance perception and sound source localization in real-world visual contexts and comparing these results with those obtained in HMD-based visual scenes could help disentangle effects specifically induced by HMDs.

Further effort should be made to identify and evaluate strategies to create coherent audiovisual scenes. Compensation techniques, applied either acoustically or visually, may reduce mismatches between modalities and enhance perceptual consistency. Their influence on perception warrants systematic investigation. Additionally, meta-analyses of the growing body of research on auditory distance perception and sound source localization in VR are needed to identify moderating factors, generalize findings across different HMDs and VR configurations such as CAVE systems, and establish universal principles of multisensory integration in VR. Expanding this research to include other sensory modalities would also be

valuable. Interactions among haptic, olfactory, and proprioceptive cues are plausible and warrant investigation in future research.

Importantly, future studies should furtherly investigate the role of measurement methods, for which not only significant main effects but also interactions with visual scene characteristics have been demonstrated. Developing and validating VR-specific measurement paradigms is essential, as appropriate methods may substantially reduce perceptual distortions. Moreover, potential tasks should be investigated that capture all three dimensions of spatial hearing, distance, azimuth, and elevation, rather than focusing solely on the first two.

Finally, open questions on adaptation to the VR environment should be addressed. Results from Studies 1 and 2 provide evidence that training effects occur and that familiarization with virtual environments can mitigate distortion. This should be examined systematically, and standardized procedures for future studies should be proposed. Incorporating extended feedback and training phases at the beginning of audiovisual VR experiments may help minimize or even eliminate VR-specific perceptual biases.

Furthermore, the combination of binaural auralizations with TTS renderings warrants investigation. Based on the findings of this dissertation, there is no indication that such a combination would be inferior to the well-studied pairing of binaural auralizations with human speech recordings. Nevertheless, it remains essential to examine whether localization accuracy and externalization are maintained when TTS stimuli are used.

In Study 3, a deliberately neutral and non-spatial audio setting was implemented. Future research should explore the limitations of TTS in contexts involving emotional voices or clearly modulated voices, as preferred by children in affective scenarios (Gustafson & House, 2001). Given the rapid advancements in artificial intelligence, it is plausible that deep learning techniques will enable highly convincing emotional modulation of TTS voices. Evaluating their integration into clinical psychological applications could offer greater flexibility and standardization.

Additionally, practical solutions for implementing binaural auralizations should be developed to make these techniques accessible to researchers without specialized acoustical expertise or collaborations with acousticians. The methods employed in Studies 2 and 4 of this project would not have been feasible without such cooperation, highlighting the need for simplified approaches that facilitate broader adoption.

In Studies 3 and 4, the effects of audio renderings on stressful social interactions in VR were examined, and no direct impact of audio was observed. However, acoustic realism

was positively correlated with presence, and stress was associated with arousal. These findings, consistent with previous research (Diemer et al., 2015; Felnhofer et al., 2019; Peperkorn et al., 2015), suggest a strong relationship between presence and affective involvement in VR. This relationship should be explored in greater depth, with particular attention to the role of the acoustic environment. As noted earlier, the low-stress condition in the stress paradigm was not entirely non-stressful, which may have masked potential audio effects. Future studies should investigate the influence of TTS and externalized sound on affective experiences in VR under genuinely low-stress conditions or with improved manipulation of emotional intensity. Scenarios in which sound source localization is critical should also be considered, for example, by reducing visual cues or presenting auditory cues outside the initial field of view.

In addition to external factors, internal factors such as interindividual differences and personality traits should be considered in future VR research. No evidence was found for an influence of social anxiety on perceived distances of speech sound sources. This was unexpected given prior findings of reduced egocentric distance and increased size estimations for phobic stimuli among highly anxious individuals compared to low-anxious individuals (Quadflieg et al., 2008; Taffou & Viaud-Delmon, 2014). However, social anxiety affected the subjective experience of virtual social interactions, influencing not only perceived pleasantness but also perceived social and physical presence. These relationships were not consistently observed across all measurements, possibly due to limited variance in anxiety levels within the samples. Previous research comparing individuals with social anxiety to healthy controls reported higher social presence among anxious participants (Felnhofer et al., 2019). While a similar trend was observed here, it was less pronounced, likely reflecting sample characteristics.

More broadly, binaural auralizations should be examined across diverse contexts and environments. While this project focused on a seminar room and a small conference room, future research should include office settings, large auditoriums, and outdoor scenes to enhance generalizability. Several paradigms relevant to clinical psychology also merit investigation, such as training programs, psychoeducation, and empathy-based interventions. Furthermore, the potential of externalized auditory distractors in attention or neuropsychological assessments, such as those for attention-deficit/hyperactivity disorder, warrants further investigation. Comparisons between neutral and positive emotional contexts, as well as conditions with and without threat, should be included. Expanding the scope to include socio-cognitive processes such as empathy or theory of mind represents a promising

direction for future research. An intriguing question would be whether audio spatialization or the use of audio-rendered agents influences empathy during social interactions in virtual environments. Such investigations could significantly broaden the understanding of how auditory renderings impact social cognition in VR. Of particular importance are both social presence and physical presence, as these dimensions are likely to interact with acoustic realism and affective engagement.

A further requirement concerns the conceptualization of social presence in interactions with non-human entities and embodied conversational agents. When participants are explicitly aware that they are interacting with a machine, even one with a human-like appearance, existing measures of social presence appear insufficient. Most current questionnaires do not differentiate between interactions with real humans and those with computer-driven agents, leaving a significant gap that should be addressed through rigorous psychological and social-psychological methods. Tailored questionnaires and concise one-item scales need to be developed and validated for these contexts. In addition, implicit measures of social presence should be explored further. Implicit indicators of physical presence include behaviors that transfer virtual experiences into real-world actions, such as finger tapping or sitting on a virtual chair without verifying its physical existence (Hinkle et al., 2025). Similarly, implicit measures for social presence could include prosocial behaviors, empathy, or replication of typical social norms and behaviors, as outlined in the introduction. Attempts have already been made to cluster behavioral indicators of social presence, including self-disclosure, laughter, and voice modulation (Hayes et al., 2022). A promising conceptualization suggests social presence as the perceptual salience of another social actor (Cummings & Wertz, 2023), which calls for the development of appropriate behavioral and attentional measurement paradigms. Both subjective and objective measures should be expanded to capture interactions with AI-generated embodied conversational agents, as these may currently represent the closest approximation to human-to-human interaction. It is plausible that incorporating highly realistic and socially adaptive agents will enhance social presence, provided that the construct is clearly defined and measured appropriately.

Finally, the homogeneity of the sample should be acknowledged. Future research should include a more diverse range of participants across sociodemographic, ethnic, and gender characteristics. Clinical samples should also be considered, as individuals with social anxiety disorders may respond differently to audiovisual paradigms and could benefit in distinct ways compared to non-clinical or subclinical populations.

In summary, future studies should explore VR-specific integration effects, address methodological limitations, and challenges related to technical feasibility. Broader contexts, varied room characteristics, advanced methods, and AI-based interactions should be employed to examine the effects of audio rendering on socio-cognitive processing, presence, and affective responses in VR. Representative and clinical samples should be investigated, as should refined indicators of social presence in interactions with embodied conversational agents, which are essential for advancing this field.

8.5 Practical implications

The findings of this dissertation project yield several implications, both for conducting experiments on auditory perception in VR and for applications in social VR contexts. First, when investigating auditory perception in VR, or more broadly, perceptual processes in VR, with an emphasis on audiovisual integration, the interaction between visual and acoustic scenes must be considered. It was clearly demonstrated that an altered visual scene not only influences auditory distance perception but also that the visibility of the virtual room interacts with the task performed in VR. As emphasized earlier, perception is not absolute. The human brain integrates physical cues with expectations and prior experiences to generate estimations, such as the perceived location of a sound source (Blauert & Braasch, 2020; Knill & Pouget, 2004). Auditory distance perception, and to a lesser extent sound source localization, are comparatively imprecise relative to their visual counterparts (Carlini et al., 2024).

Beyond these findings, acoustic factors in VR environments warrant careful consideration. Physical sound sources originating outside the VR context can disrupt place illusion or interfere with the sense of presence (Brübach et al., 2022). Conversely, auditory cues embedded within the virtual scene significantly shape the overall experience (e.g. Study 2, Freeman & Lessiter, 2001). Interestingly, the interaction between visual and acoustic scenes can be leveraged to mitigate specific visual distortions inherent to VR, such as visual distance compression. For example, adjusting perceived auditory distance by manipulating reverberation time within the virtual acoustic environment can compensate for visually compressed spaces (Huang et al., 2021).

Also, including a familiarization phase with binaural auralizations may reduce the slight differences observed between auralizations and loudspeakers, as well as between individualized and generic auralizations. As previously noted, one practical approach to mitigating the adverse effects of non-individualized HRTFs is training (Guezenoc & Seguier,

2018). There is also anecdotal evidence suggesting that binaural auralizations perform better for expert listeners or individuals accustomed to virtual acoustics (Brandenburg et al., 2020).

Moreover, substantial interindividual variability emerged in ratings of social and physical presence, audio quality, and externalization. For instance, externalization ratings of non-spatial (diotic) auralizations ranged from 0 to 100, possibly indicating limited familiarity with the concept of externalization. Although explanatory instructions and examples were provided, these did not appear to sustain comprehension throughout the experiment. This may indicate conceptual ambiguity or the inherent volatility of auditory experiences and memory (Kraus & White-Schwoch, 2015; Larsson et al., 2008). Future studies should therefore ensure that participants receive clear and repeated explanations of such constructs to minimize variability and improve data reliability.

It is essential not only to provide clear explanations of constructs related to acoustic quality for participants unfamiliar with these concepts, but also to ensure that the measurements themselves are as precise as possible. In particular, constructs such as social and physical presence should be assessed using well-defined, carefully formulated items, ideally based on validated instruments. However, it may be more beneficial to employ a thoughtfully refined and targeted measure of social presence rather than a broadly validated questionnaire that captures a wide range of phenomena. Researchers should first determine which specific facet of social presence is of interest and then consider validating a tailored rating scale, e.g., through pilot testing.

An important implication of the current findings is that all measurement constructs, task designs, and manipulations in VR should be considered holistically. Given that in Study 1, differential effects of the task on symptoms such as nausea and vertigo were found, potential effects on simulator sickness also require attention. When simulator sickness occurs, even the most realistic audiovisual environment cannot preserve a positive experience, turning VR use unpleasant or intolerable (Kemeny et al., 2020).

Although not a direct conclusion of this project, the broader issue of multimodal interference in VR deserves consideration. Evidence suggests that perception in VR is inherently altered (Buck et al., 2021; Buck et al., 2018; Creem-Regehr et al., 2023; Huang et al., 2021). The combination of physical and virtual sensations may produce unpredictable perceptual outcomes. Conceptual frameworks classify virtual scenarios along a continuum from augmented reality, in which only limited sensory input is virtual, to full VR, in which all sensory modalities are simulated (Davis et al., 2003). In most VR applications, particularly those using HMDs, substantial sensory input is virtual. At the same time, certain physical

sensations, such as ambient sounds, smells, or temperature, remain partially perceptible. Therefore, the effects of multimodal integration, especially the interplay between physical and virtual cues, should be carefully examined.

This consideration is particularly critical for VR applications in diagnostic and training contexts. For instance, training programs for children with auditory processing disorders (Ramírez et al., 2024) should account for localization blur induced by virtualization. While this example pertains to early childhood, the implications extend across the lifespan. Similarly, VR-based neuropsychological assessments are increasingly relevant (Buchmann & Randerath, 2017). Specific impairments with real-world consequences, such as difficulties in judging whether an object is within arm's reach, are closely tied to distance estimation (Randerath et al., 2021). VR assessments are suggested to offer higher ecological validity than traditional paper-and-pencil or computer-based tests because they allow testing more closely related to daily life functioning (Parsons, 2011). However, internal validity can only be maintained if distortions introduced by virtualization are recognized and appropriately compensated. This highlights the importance of avoiding disciplinary tunnel vision. VR research is inherently interdisciplinary, requiring insights from psychology, such as auditory and visual perception, and technological expertise, for example, from acousticians.

The extensive body of VR research conducted primarily by computer scientists and graphic designers has not only made VR systems increasingly accessible for psychological research but also provided essential knowledge for interpreting findings and defining prerequisites for successful applications in mental health. Conversely, computer science has produced numerous feasibility studies on VR-based interventions for mental health, many of which show considerable promise. Yet, without rigorous validation by clinical psychologists and related experts, these approaches are unlikely to progress beyond preliminary stages. Overall, interdisciplinary collaboration, particularly in audiovisual VR but also in broader domains where humans interact with computer simulations, offers the most significant potential to advance the field.

A seamless integration of multiple domains should be a primary objective when creating convincing virtual environments. This is especially critical when the goal is to replicate or pre-simulate real-world spaces. In such cases, visual and acoustic components must be closely aligned. Alongside the work presented in this dissertation, a calibration method was developed to synchronize virtual and physical spaces using fixed reference points and motion controller tracking (Kroczek et al., 2023). This ensured that participants' positions in the physical environment corresponded accurately to their positions in the virtual

environment. Such alignment is essential for localization and perceptual experiments, as well as for scenarios combining physical and virtual stimuli, a common feature in most VR settings (Davis et al., 2003). Beyond matching room dimensions and accounting for potential compression effects, participant positioning must also be considered. The calibration technique proposed here addresses these requirements and is recommended for future studies aimed at aligning physical and virtual positions.

8.6 General conclusion

This dissertation project investigated how audio renderings influence socio-cognitive processing, presence, and affective responses in VR. Across four empirical studies, the findings reveal a complex interplay between auditory and visual scene and their combined impact on affective experience and overall quality of experience. The first key conclusion is that incongruence between visual and auditory scenes negatively affects auditory distance perception, underscoring the importance of carefully considering VR-specific effects in research on auditory cognition and virtual acoustics. Furthermore, the task used to assess sound source localization interacted with the visual scene to shape localization accuracy and affected levels of adverse effects and presence. Auditory realism was positively associated with both physical and social presence. HTBAs have been shown to evoke spatialized and externalized hearing impressions, thereby enhancing perceived realism in VR. However, their effect on social presence varied by context. While HTBAs enhanced social presence in basic perceptual tasks, this effect did not extend to social stress paradigms, underscoring both the potential and context-dependence of virtual acoustics in social VR. In high-arousal contexts, audio renderings did not significantly influence affective responses or social presence. Stress was positively correlated with social presence, which may explain the limited influence of HTBAs under demanding conditions. Additionally, the use of AI-generated TTS proved equivalent to human voice recordings in eliciting social evaluative threat and social presence. This finding supports the integration of TTS in social VR interactions, including experiments involving conversational agents. It highlights its potential as a practical alternative to recorded voices in VR-based research. The broader relevance of this work lies in its connection to social anxiety and its potential application in therapeutic and training interventions. Although social anxiety did not affect auditory distance perception, individual differences influenced both social and physical presence and partially moderated the effects of the HTBAs. Ratings of acoustic features and auralization characteristics varied substantially, likely reflecting participants' limited familiarity with virtual acoustics and HTBA technology. These findings

emphasize the need for clearly defined constructs and carefully selected measurement metrics. Future research should incorporate refined assessments of social presence, including implicit measures, and extend investigations to additional socio-cognitive processes such as empathy. Investigating audio renderings in interactions with AI agents and tailoring studies to human-computer interaction contexts will be particularly valuable.

In conclusion, audio renderings can enhance social presence and realism in virtual social interactions, but their effects and relevance are highly context-dependent. Auditory immersion appears most promising in scenarios with moderate affective arousal. Researchers should consider the interplay between VR-specific visual features and auditory variables. Practicality aspects, such as the use of TTS and the further simplification of HTBAs for applied settings, will facilitate the integration of advanced acoustic techniques into psychological research. By combining state-of-the-art audio technologies with well-established, validated psychological paradigms, this dissertation established a basis for implementing advanced audiovisual VR environments and provides recommendations for their application in both basic research and clinical psychology.

9 Literature

Abdulrahman, A., & Richards, D. (2022). Is natural necessary? Human voice versus synthetic voice for intelligent virtual agents. *Multimodal Technologies and Interaction*, *6*(7), 51.

Adams, H., Stefanucci, J., Creem-Regehr, S., & Bodenheimer, B. (2022). Depth perception in augmented reality: The effects of display, shadow, and position. *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 792–801.

Agrawal, S., Simon, A., Bech, S., Bærentsen, K., & Forchhammer, S. (2020). Defining Immersion: Literature Review and Implications for Research on Audiovisual Experiences. *Journal of the Audio Engineering Society*, *68*(6), 404–417. https://doi.org/10.17743/jaes.2020.0039

Ahrens, A., Lund, K. D., Marschall, M., & Dau, T. (2019). Sound source localization with varying amount of visual information in virtual reality. *PLOS ONE*, *14*(3), e0214603. https://doi.org/10.1371/journal.pone.0214603

Alcañiz, M., Parra, E., Chicchi Giglioli, I. A., & García, A. (2024). Chapter 7 Using Virtual Reality for Leadership Assessment and Training Through Behavioral Biomarkers. In L. Moutinho & M. Cerf (Eds.), *Biometrics and Neuroscience Research in Business and Management* (pp. 141–170). De Gruyter. https://doi.org/10.1515/9783110708509-007

Allen, A. P., Kennedy, P. J., Dockray, S., Cryan, J. F., Dinan, T. G., & Clarke, G. (2017). The Trier Social Stress Test: Principles and practice. *Neurobiology of Stress*, *6*, 113–126. https://doi.org/10.1016/j.ynstr.2016.11.001

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences*, *4*(7), 267–278. https://doi.org/10.1016/S1364-6613(00)01501-1

Altenhoff, B. M., Napieralski, P. E., Long, L. O., Bertrand, J. W., Pagano, C. C., Babu, S. V., & Davis, T. A. (2012). Effects of calibration to visual and haptic feedback on near-field depth perception in an immersive virtual environment. *Proceedings of the ACM Symposium on Applied Perception*, 71–78. https://doi.org/10.1145/2338676.2338691

American Psychiatric Association, D. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (Vol. 5). American psychiatric association Washington, DC.

Andrist, S., Gleicher, M., & Mutlu, B. (2017). Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2571–2582. https://doi.org/10.1145/3025453.3026033

Ani, R., Maria, E., Joyce, J. J., Sakkaravarthy, V., & Raja, M. A. (2017). Smart Specs: Voice assisted text reading system for visually impaired persons using TTS method. *2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT)*, 1–6. https://doi.org/10.1109/IGEHT.2017.8094103

Bahr, L.-M., Maurer, F., Weigl, J., Weber, K., Melchner, D., Dörfelt, A., Wechsler, T. F., Bauer, O., Reinders, J., & Milenkovic, V. M. (2021). Dissociation of endocrine responses to the Trier Social Stress Test in Virtual Reality (VR-TSST) by the benzodiazepine alprazolam and the translocator protein 18 kDa (TSPO) ligand etifoxine. *Psychoneuroendocrinology*, *124*, 105100.

Bandelow, B., Lichte, T., Rudolf, S., Wiltink, J., & Beutel, M. E. (2015). The German guidelines for the treatment of anxiety disorders. *European Archives of Psychiatry and Clinical Neuroscience*, *265*(5), 363–373. https://doi.org/10.1007/s00406-014-0563-z

Bandelow, B., Werner, A. M., Kopp, I., Rudolf, S., Wiltink, J., & Beutel, M. E. (2022). The German Guidelines for the treatment of anxiety disorders: First revision. *European Archives of Psychiatry and Clinical Neuroscience*, *272*(4), 571–582. https://doi.org/10.1007/s00406-021-01324-1

Bandelow, B., Wiltink, J., & Watzke, B. (2014). *S3-Leitlinie Behandlung von Angststörungen: Langfassung*. https://www.zora.uzh.ch/id/eprint/125722/

Bärtl, C., Henze, G.-I., Peter, H. L., Giglberger, M., Bohmann, P., Speicher, N., Konzok, J., Kreuzpointner, L., Waller, L., & Walter, H. (2024). Neural and cortisol responses to acute psychosocial stress in work-related burnout: The Regensburg Burnout Project. *Psychoneuroendocrinology*, *161*, 106926.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience*, *7*(11), 1190–1192. https://doi.org/10.1038/nn1333

Begault, D. R., Wenzel, E. M., & Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, *49*(10), 904–916.

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding Voice Perception. *British Journal of Psychology*, *102*(4), 711–725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Berkman, M. İ., & Çatak, G. (2021). I-group presence questionnaire: Psychometrically revised English version. *Mugla Journal of Science and Technology*, *7*, 1–10.

Best, V., Baumgartner, R., Lavandier, M., Majdak, P., & Kopčo, N. (2020). Sound Externalization: A Review of Recent Research. *Trends in Hearing*, *24*, 233121652094839. https://doi.org/10.1177/2331216520948390

Bidelman, G. M., Bernard, F., & Skubic, K. (2025). Hearing in categories and speech perception at the "cocktail party." *PLOS ONE*, *20*(1), e0318600. https://doi.org/10.1371/journal.pone.0318600

Biocca, F., Harms, C., & Burgoon, J. K. (2003). Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators and Virtual Environments*, *12*(5), 456–480. https://doi.org/10.1162/105474603322761270

Biocca, F., & Levy, M. R. (Eds.). (2013). *Communication in the Age of Virtual Reality* (1st ed.). Routledge. https://doi.org/10.4324/9781410603128

Blau, M., Budnik, A., Fallahi, M., Steffens, H., Ewert, S. D., & Van De Par, S. (2021). Toward realistic binaural auralizations – perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario. *Acta Acustica*, *5*, 8. https://doi.org/10.1051/aacus/2020034

Blauert, J. (1997). *Spatial hearing: The psychophysics of human sound localization*. MIT press.

Blauert, J., & Braasch, J. (2020). *The technology of binaural understanding*. Springer Nature.

Bombari, D., Schmid Mast, M., Canadas, E., & Bachmann, M. (2015). Studying social interactions through immersive virtual environment technology: Virtues, pitfalls, and future challenges. *Frontiers in Psychology*, *6*, 869.

Bormann, K. (2005). Presence and the utility of audio spatialization. *Presence: Teleoperators & Virtual Environments*, *14*(3), 278–297.

Brandenburg, K., Klein, F., Neidhardt, A., Sloma, U., & Werner, S. (2020). Creating Auditory Illusions with Binaural Technology. In J. Blauert & J. Braasch (Eds.), *The Technology of Binaural Understanding* (pp. 623–663). Springer International Publishing. https://doi.org/10.1007/978-3-030-00386-9_21

Braren, H. S., & Fels, J. (2021). Towards child-appropriate virtual acoustic environments: A database of high-resolution HRTF measurements and 3D-scans of children. *International Journal of Environmental Research and Public Health*, *19*(1), 324.

Brinkmann, F., Lindau, A., & Weinzierl, S. (2017). On the authenticity of individual dynamic binaural synthesis. *The Journal of the Acoustical Society of America*, *142*(4), 1784–1795. https://doi.org/10.1121/1.5005606

Brodoehl, S., Klingner, C. M., & Witte, O. W. (2015). Eye closure enhances dark night perceptions. *Scientific Reports*, *5*(1), 10515. https://doi.org/10.1038/srep10515

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica*, *86*(1), 117–128.

Brübach, L., Westermeier, F., Wienrich, C., & Latoschik, M. E. (2022). Breaking Plausibility Without Breaking Presence—Evidence For The Multi-Layer Nature Of Plausibility. *IEEE Transactions on Visualization and Computer Graphics*, *28*(5), 2267–2276. https://doi.org/10.1109/TVCG.2022.3150496

Brück, C., Kreifelts, B., & Wildgruber, D. (2011). Emotional voices in context: A neurobiological model of multimodal affective information processing. *Physics of Life Reviews*, *8*(4), 383–403. https://doi.org/10.1016/j.plrev.2011.10.002

Buchmann, I., & Randerath, J. (2017). Selection and application of familiar and novel tools in patients with left and right hemispheric stroke: Psychometrics and normative data. *Cortex*, *94*, 49–62. https://doi.org/10.1016/j.cortex.2017.06.001

Buck, L. E., Paris, R., & Bodenheimer, B. (2021). Distance Compression in the HTC Vive Pro: A Quick Revisitation of Resolution. *Frontiers in Virtual Reality*, *2*.

Buck, L. E., Young, M. K., & Bodenheimer, B. (2018). A Comparison of Distance Estimation in HMD-Based Virtual Environments with Different HMD-Based Conditions. *ACM Transactions on Applied Perception*, *15*(3), 1–15. https://doi.org/10.1145/3196885

Cabero-Almenara, J., Barroso-Osuna, J., Llorente-Cejudo, C., & Fernández Martínez, M. del M. (2019). Educational Uses of Augmented Reality (AR): Experiences in Educational Science. *Sustainability*, *11*(18), Article 18. https://doi.org/10.3390/su11184990

Cacioppo, J. T., & Decety, J. (2011). *The Oxford handbook of social neuroscience*. Oxford University Press.

Cambre, J., Colnago, J., Maddock, J., Tsai, J., & Kaye, J. (2020). Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content.

Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3313831.3376789

Carl, E., Stein, A. T., Levihn-Coon, A., Pogue, J. R., Rothbaum, B., Emmelkamp, P., Asmundson, G. J. G., Carlbring, P., & Powers, M. B. (2019). Virtual reality exposure therapy for anxiety and related disorders: A meta-analysis of randomized controlled trials. *Journal of Anxiety Disorders*, *61*, 27–36. https://doi.org/10.1016/j.janxdis.2018.08.003

Carlini, A., Bordeau, C., & Ambard, M. (2024). Auditory localization: A comprehensive practical review. *Frontiers in Psychology*, *15*, 1408073. https://doi.org/10.3389/fpsyg.2024.1408073

Chalmers, A., & Ferko, A. (2008). Levels of realism: From virtual reality to real virtuality. *Proceedings of the 24th Spring Conference on Computer Graphics*, 19–25. https://doi.org/10.1145/1921264.1921272

Chen, L., & Vroomen, J. (2013). Intersensory binding across space and time: A tutorial review. *Attention, Perception, & Psychophysics*, *75*(5), 790–811. https://doi.org/10.3758/s13414-013-0475-4

Chen, N. T. M., & Clarke, P. J. F. (2017). Gaze-Based Assessments of Vigilance and Avoidance in Social Anxiety: A Review. *Current Psychiatry Reports*, *19*(9), 59. https://doi.org/10.1007/s11920-017-0808-4

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*, 975–979.

Christie, B. (1974). Perceived usefulness of person-person telecommunications media as a function of the intended application. *European Journal of Social Psychology*, *4*(3), 366–368. https://doi.org/10.1002/ejsp.2420040307

Cohn, M., Bandodkar, G., Sangani, R. B., Predeck, K., & Zellou, G. (2024). Do People Mirror Emotion Differently with a Human or TTS Voice? Comparing Listener Ratings and Word Embeddings. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3613905.3650757

Cohn, M., Chen, C.-Y., & Yu, Z. (2019). A large-scale user study of an Alexa prize chatbot: Effect of TTS dynamism on perceived quality of social dialog. *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 293–306.

Cooper, E., Huang, W.-C., Tsao, Y., Wang, H.-M., Toda, T., & Yamagishi, J. (2024). A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, *45*(4), 161–183.

Corporation, V. (2022). [Computer software] (Version 4.1.4). Valve Corporation. https://valvesoftware.github.io/steam-audio/doc/capi/index.html

Creem-Regehr, S. H., Stefanucci, J. K., & Bodenheimer, B. (2023). Perceiving distance in virtual reality: Theoretical insights from contemporary technologies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1869), 20210456. https://doi.org/10.1098/rstb.2021.0456

Cuevas-Rodríguez, M., Picinali, L., González-Toledo, D., Garre, C., Rubia-Cuestas, E. de la, Molina-Tanco, L., & Reyes-Lecuona, A. (2019). 3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation. *PLOS ONE*, *14*(3), e0211899. https://doi.org/10.1371/journal.pone.0211899

Cummings, J. J., & Wertz, E. E. (2023). Capturing social presence: Concept explication through an empirical analysis of social presence measures. *Journal of Computer-Mediated Communication*, *28*(1), zmac027. https://doi.org/10.1093/jcmc/zmac027

Dai, L., Kritskaia, V., Van Der Velden, E., Vervoort, R., Blankendaal, M., Jung, M. M., Postma, M. Š., & Louwerse, M. M. (2024). Text-to-speech and virtual reality agents in primary school classroom environments. *Journal of Computer Assisted Learning*, jcal.13046. https://doi.org/10.1111/jcal.13046

Darling, K. (2015). "Who's Johnny?" Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. In *ROBOT ETHICS 2.0*. Oxford University Press. http://www.ssrn.com/abstract=2588669

Davis, L., Rolland, J., Hamza-Lup, F., Ha, Y., Norfleet, J., & Imielinska, C. (2003). Enabling a continuum of virtual environment experiences. *IEEE Computer Graphics and Applications*, *23*(2), 10–12. https://doi.org/10.1109/MCG.2003.1185574

Denk, F., Brinkmann, F., Stirnemann, A., & Kollmeier, B. (2019). The PIRATE: An anthropometric earPlug with exchangeable microphones for Individual Reliable Acquisition of Transfer functions at the Ear canal entrance. *Fortschritte Der Akustik—DAGA*, 18–21.

Development Core, R. T. (2019). *R: A language and environment for statistical computing* [Computer software].

Dicke, C., Aaltonen, V., Rämö, A., & Vilermo, M. (2010). Talk to me: The influence of audio quality on the perception of social presence. *Proceedings of HCI 2010*. https://www.scienceopen.com/hosted-document?doi=10.14236/ewic/HCI2010.36

Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, *130*(3), 355.

Dickerson, S. S., Mycek, P. J., & Zaldivar, F. (2008). Negative social evaluation, but not mere social presence, elicits cortisol responses to a laboratory stressor task. *Health Psychology*, *27*(1), 116.

Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., & Mühlberger, A. (2015). The impact of perception and presence on emotional reactions: A review of research in virtual reality. *Frontiers in Psychology*, *6*, 26.

Do, T. D., Zelenty, S., Gonzalez-Franco, M., & McMahan, R. P. (2023). VALID: A perceptually validated Virtual Avatar Library for Inclusion and Diversity. *Frontiers in Virtual Reality*, *4*. https://doi.org/10.3389/frvir.2023.1248915

Durlach, N. I., Shinn-Cunningham, B. G., & Held, R. M. (1993). Supernormal auditory LocalizationI. general background. *Presence: Teleoperators & Virtual Environments*, *2*(2), 89–103.

Ebner, P., & Szczuka, J. (2025). *Predicting Romantic Human-Chatbot Relationships: A Mixed-Method Study on the Key Psychological Factors* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2503.00195

Emmelkamp, P. M. G., Meyerbröker, K., & Morina, N. (2020). Virtual Reality Therapy in Social Anxiety Disorder. *Current Psychiatry Reports*, *22*(7), 32. https://doi.org/10.1007/s11920-020-01156-1

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864.

Etchemendy, P. E., Spiousas, I., Calcagno, E. R., Abregú, E., Eguia, M. C., & Vergara, R. O. (2018). Direct-location versus verbal report methods for measuring auditory distance perception in the far field. *Behavior Research Methods*, *50*(3), 1234–1247. https://doi.org/10.3758/s13428-017-0939-x

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Felnhofer, A., Hlavacs, H., Beutl, L., Kryspin-Exner, I., & Kothgassner, O. D. (2019). Physical Presence, Social Presence, and Anxiety in Participants with Social Anxiety Disorder During Virtual Cue Exposure. *Cyberpsychology, Behavior, and Social Networking*, *22*(1), 46–50. https://doi.org/10.1089/cyber.2018.0221

Felnhofer, A., Pfannerstill, F., Gänsler, L., Kothgassner, O. D., Humer, E., Büttner, J., & Probst, T. (2025). Barriers to adopting therapeutic virtual reality: The perspective of clinical psychologists and psychotherapists. *Frontiers in Psychiatry*, *16*. https://doi.org/10.3389/fpsyt.2025.1549090

Felton, W. M., & Jackson, R. E. (2022). Presence: A Review. *International Journal of Human–Computer Interaction*, *38*(1), 1–18. https://doi.org/10.1080/10447318.2021.1921368

Finisguerra, A., Canzoneri, E., Serino, A., Pozzo, T., & Bassolino, M. (2015). Moving sounds within the peripersonal space modulate the motor system. *Neuropsychologia*, *70*, 421–428. https://doi.org/10.1016/j.neuropsychologia.2014.09.043

Fiske, S. T. T., & Taylor, S. E. (2020). *Social cognition: From brains to culture* (4th Edition). SAGE Publications Ltd.

Foulsham, T., & Sanderson, L. A. (2013). Look who's talking? Sound changes gaze behaviour in a dynamic social scene. *Visual Cognition*, *21*(7), 922–944. https://doi.org/10.1080/13506285.2013.849785

Franke, G. H., Ankerhold, A., Haase, M., Jäger, S., Tögel, C., Ulrich, C., & Frommer, J. (2011). The usefulness of the Brief Symptom Inventory 18 (BSI-18) in psychotherapeutic patients. *Psychotherapie, Psychosomatik, Medizinische Psychologie*, *61*(2), 82–86.

Freeman, J., & Lessiter, J. (2001). Here, there and everywhere: The effects of multichannel audio on presence. *Proceedings of ICAD*, 231–234.

Frisch, J. U., Häusser, J. A., & Mojzisch, A. (2015). The Trier Social Stress Test as a paradigm to study how people respond to threat in social interactions. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00014

Fujihira, H., Itoi, C., Furukawa, S., Kato, N., & Kashino, M. (2022). Sensitivity to interaural level and time differences in individuals with autism spectrum disorder. *Scientific Reports*, *12*(1), 19142. https://doi.org/10.1038/s41598-022-23346-y

Funk, H., Kuhn, C., Nielsen, L., Rische, K., Lex, B., Redecker, B., & Winzer-Kiontke, B. (2013). Studio" 21": Das Deutschbuch: Deutsch als Fremdsprache.

Gagnon, K. T., Geuss, M. N., & Stefanucci, J. K. (2013). Fear influences perceived reaching to targets in audition, but not vision. *Evolution and Human Behavior*, *34*(1), 49–54. https://doi.org/10.1016/j.evolhumbehav.2012.09.002

Gardner, M. B. (1968). Proximity image effect in sound localization. *The Journal of the Acoustical Society of America*, *43*(1), 163. https://doi.org/10.1121/1.1910747

Getzmann, S., Schneider, D., & Wascher, E. (2023). Selective spatial attention in lateralized multi-talker speech perception: EEG correlates and the role of age. *Neurobiology of Aging*, *126*, 1–13. https://doi.org/10.1016/j.neurobiolaging.2023.02.003

Ghazanfar, A. A., Neuhoff, J. G., & Logothetis, N. K. (2002). Auditory looming perception in rhesus monkeys. *Proceedings of the National Academy of Sciences*, *99*(24), 15755–15757. https://doi.org/10.1073/pnas.242469699

Ghinea, M., Frunză, D., Chardonnet, J.-R., Merienne, F., & Kemeny, A. (2018). Perception of Absolute Distances Within Different Visualization Systems: HMD and CAVE. In L. T. De Paolis & P. Bourdot (Eds.), *Augmented Reality, Virtual Reality, and Computer Graphics* (Vol. 10850, pp. 148–161). Springer International Publishing. https://doi.org/10.1007/978-3-319-95270-3_10

Gil-Carvajal, J. C., Cubick, J., Santurette, S., & Dau, T. (2016). Spatial Hearing with Incongruent Visual or Auditory Room Cues. *Scientific Reports*, *6*(1), 37342. https://doi.org/10.1038/srep37342

Gillies, M., & Pan, X. (2018). Virtual reality for social skills training. *Proceedings of the Virtual and Augmented Reality to Enhance Learning and Teaching in Higher Education Conference 2018*, *8*, 83–92. https://research.gold.ac.uk/id/eprint/26891/

Givon-Benjio, N., & Okon-Singer, H. (2022). The role of fear in hierarchical processing of fear-related stimuli. *Journal of Psychopathology and Clinical Science*, *131*(7), 727.

Glennerster, A., Tcheang, L., Gilson, S. J., Fitzgibbon, A. W., & Parker, A. J. (2006). Humans Ignore Motion and Stereo Cues in Favor of a Fictional Stable World. *Current Biology*, *16*(4), 428–432. https://doi.org/10.1016/j.cub.2006.01.019

Goodman, W. K., Janson, J., & Wolf, J. M. (2017). Meta-analytical assessment of the effects of protocol variations on cortisol responses to the Trier Social Stress Test. *Psychoneuroendocrinology*, *80*, 26–35. https://doi.org/10.1016/j.psyneuen.2017.02.030

Gorini, A., Capideville, C. S., De Leo, G., Mantovani, F., & Riva, G. (2011). The Role of Immersion and Narrative in Mediated Presence: The Virtual Hospital Experience. *Cyberpsychology, Behavior, and Social Networking*, *14*(3), 99–105. https://doi.org/10.1089/cyber.2010.0100

Grace, C., Heinrichs, M., Koval, P., Gorelik, A., von Dawans, B., Terrett, G., Rendell, P., & Labuschagne, I. (2022). Concordance in salivary cortisol and subjective anxiety to the trier social stress test in social anxiety disorder. *Biological Psychology*, *175*, 108444. https://doi.org/10.1016/j.biopsycho.2022.108444

Grassini, S., & Laumann, K. (2020). Questionnaire Measures and Physiological Correlates of Presence: A Systematic Review. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00349

Green, E. J., & Barber, P. J. (1981). An auditory Stroop effect with judgments of speaker gender. *Perception & Psychophysics*, *30*(5), 459–466. https://doi.org/10.3758/BF03204842

Grillon, H., Riquier, F., Herbelin, B., & Thalmann, D. (2006). Virtual reality as a therapeutic tool in the confines of social anxiety disorder treatment. *International Journal on Disability and Human Development*, *5*(3). https://doi.org/10.1515/IJDHD.2006.5.3.243

Grumiaux, P.-A., Kitić, S., Girin, L., & Guérin, A. (2022). A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America*, *152*(1), 107–151. https://doi.org/10.1121/10.0011809

Guezenoc, C., & Seguier, R. (2018). HRTF Individualization: A Survey. *145th Audio Engineering Society Convention*. https://doi.org/10.17743/aesconv.2018.978-1-942220-25-1

Gunnar, M. R., Reid, B. M., Donzella, B., Miller, Z. R., Gardow, S., Tsakonas, N. C., Thomas, K. M., DeJoseph, M., & Bendezú, J. J. (2021). Validation of an online version of the Trier Social Stress Test in a study of adolescents. *Psychoneuroendocrinology*, *125*, 105111. https://doi.org/10.1016/j.psyneuen.2020.105111

Gustafson, K., & House, D. (2001). *Fun or boring? A web-based evaluation of expressive synthesis for children*. 565–568. https://doi.org/10.21437/Eurospeech.2001-151

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the Lowest of the Low: Neuroimaging Responses to Extreme Out-Groups. *Psychological Science*, *17*(10), 847–853. https://doi.org/10.1111/j.1467-9280.2006.01793.x

Hayes, A. T., Hughes, C. E., & Bailenson, J. (2022). Identifying and Coding Behavioral Indicators of Social Presence With a Social Presence Behavioral Coding System. *Frontiers in Virtual Reality*, *3*. https://doi.org/10.3389/frvir.2022.773448

Helminen, E. C., Morton, M. L., Wang, Q., & Felver, J. C. (2019). A meta-analysis of cortisol reactivity to the Trier Social Stress Test in virtual environments. *Psychoneuroendocrinology*, *110*, 104437. https://doi.org/10.1016/j.psyneuen.2019.104437

Hendrix, C., & Barfield, W. (1995). Presence in virtual environments as a function of visual and auditory cues. *Proceedings Virtual Reality Annual International Symposium '95*, 74–82. https://doi.org/10.1109/VRAIS.1995.512482

Herawati, D. N., Widajati, W., & Sartinah, E. P. (2022). The Role of Text To Speech Assistive Technology to Improve Reading Ability in E-Learning for ADHD Students. *Journal of ICSAR*, *6*(2), 169.

Het, S., Rohleder, N., Schoofs, D., Kirschbaum, C., & Wolf, O. T. (2009). Neuroendocrine and psychometric evaluation of a placebo version of the 'Trier Social Stress Test.' *Psychoneuroendocrinology*, *34*(7), 1075–1086. https://doi.org/10.1016/j.psyneuen.2009.02.008

Higashiyama, A. (1984). The effects of familiar size on judgments of size and distance: An interaction of viewing attitude with spatial cues. *Perception & Psychophysics*, *35*(4), 305–312. https://doi.org/10.3758/BF03206333

Hinkle, S., Soares, S., & Bohil, C. (2025). Finger Tapping as a Behavioral Indicator of Presence in VR. *PRESENCE: Virtual and Augmented Reality*, *34*(1), 27–42. https://doi.org/10.1162/PRES_a_00438

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Hornsey, R. L., & Hibbard, P. B. (2021). Contributions of pictorial and binocular cues to the perception of distance in virtual reality. *Virtual Reality*, *25*(4), 1087–1103. https://doi.org/10.1007/s10055-021-00500-x

Howard, M. C., & Gutworth, M. B. (2020). A meta-analysis of virtual reality training programs for social skill development. *Computers & Education*, *144*, 103707. https://doi.org/10.1016/j.compedu.2019.103707

Huang, Y.-H., Venkatakrishnan, R., Venkatakrishnan, R., Babu, S. V., & Lin, W.-C. (2021). Using Audio Reverberation to Compensate Distance Compression in Virtual Reality. *ACM Symposium on Applied Perception 2021*, 1–10. https://doi.org/10.1145/3474451.3476236

Huisman, T., Ahrens, A., & MacDonald, E. (2021). Ambisonics Sound Source Localization With Varying Amount of Visual Information in Virtual Reality. *Frontiers in Virtual Reality*, *2*, 722321. https://doi.org/10.3389/frvir.2021.722321

Hutmacher, F. (2019). Why Is There So Much More Research on Vision Than on Any Other Sensory Modality? *Frontiers in Psychology*, *10*, 2246. https://doi.org/10.3389/fpsyg.2019.02246

Immohr, F., Rendle, G., Kehling, C., Lammert, A., Göring, S., Froehlich, B., & Walessa, M. (2024). Subjective Evaluation of the Impact of Spatial Audio on Triadic Communication in Virtual Reality. *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*, 262–265. https://doi.org/10.1109/QoMEX61742.2024.10598292

Immohr, F., Rendle, G., Lammert, A., Neidhardt, A., Heyde, V. M. Z., Froehlich, B., & Raake, A. (2024). Evaluating the Effect of Binaural Auralization on Audiovisual Plausibility and Communication Behavior in Virtual Reality. *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 849–858. https://doi.org/10.1109/VR58804.2024.00104

Immohr, F., Rendle, G., Neidhardt, A., Göring, S., Ramachandra Rao, R. R., Arevalo Arboleda, S., Froehlich, B., & Raake, A. (2023). Proof-of-Concept Study to Evaluate the Impact of Spatial Audio on Social Presence and User Behavior in Multi-Modal VR Communication. *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, 209–215. https://doi.org/10.1145/3573381.3596458

Jack, C. E., & Thurlow, W. R. (1973). Effects of Degree of Visual Association and Angle of Displacement on the "Ventriloquism" Effect. *Perceptual and Motor Skills*, *37*(3), 967–979. https://doi.org/10.1177/003151257303700360

Jaeger, H., Bitzer, J., Simmer, U., & Blau, M. (2017). Echtzeitfähiges binaurales Rendering mit Bewegungssensoren von 3-D Brillen. *Proc. Fortschritte Der Akustik–DAGA*, 1130–1133.

Janke, W., & Erdmann, G. (2008). *Stressverarbeitungsfragebogen: SVF; Stress, Stressverarbeitung und ihre Erfassung durch ein mehrdimensionales Testsystem. Hogrefe.*

Javitt, D. C., & Sweet, R. A. (2015). Auditory dysfunction in schizophrenia: Integrating clinical and basic features. *Nature Reviews Neuroscience*, *16*(9), 535–550. https://doi.org/10.1038/nrn4002

Jeffreys, H. (1998). *The theory of probability*. OuP.

Jenny, C., & Reuter, C. (2020). Usability of Individualized Head-Related Transfer Functions in Virtual Reality: Empirical Study With Perceptual Attributes in Sagittal Plane Sound Localization. *JMIR Serious Games*, *8*(3), e17576. https://doi.org/10.2196/17576

Jensen, C., Farnham, S. D., Drucker, S. M., & Kollock, P. (2000). The effect of communication modality on cooperation in online environments. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 470–477. https://doi.org/10.1145/332040.332478

Jeunet, C., Albert, L., Argelaguet, F., & Lécuyer, A. (2018). "Do you feel in control?": Towards novel approaches to characterise, manipulate and measure the sense of agency in virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, *24*(4), 1486–1495.

Jicol, C., Wan, C. H., Doling, B., Illingworth, C. H., Yoon, J., Headey, C., Lutteroth, C., Proulx, M. J., Petrini, K., & O'Neill, E. (2021). Effects of Emotion and Agency on Presence in Virtual Reality. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445588

Kadunce, D., Vaughan, W., Wallace, M., & Stein, B. (2001). The influence of visual and auditory receptive field organization on multisensory integration in the superior colliculus. *Experimental Brain Research*, *139*(3), 303–310. https://doi.org/10.1007/s002210100772

Kaur, N., & Singh, P. (2023). Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review*, *56*(7), 5837–5880. https://doi.org/10.1007/s10462-022-10315-0

Kelly, J. W. (2023). Distance Perception in Virtual Reality: A Meta-Analysis of the Effect of Head-Mounted Display Characteristics. *IEEE Transactions on Visualization and Computer Graphics*, *29*(12), 4978–4989. IEEE Transactions on Visualization and Computer Graphics. https://doi.org/10.1109/TVCG.2022.3196606

Kemeny, A., Chardonnet, J.-R., & Colombet, F. (2020). Reducing Cybersickness. In A. Kemeny, J.-R. Chardonnet, & F. Colombet (Eds.), *Getting Rid of Cybersickness: In Virtual Reality, Augmented Reality, and Simulators* (pp. 93–132). Springer International Publishing. https://doi.org/10.1007/978-3-030-59342-1_4

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology*, *3*(3), 203–220. https://doi.org/10.1207/s15327108ijap0303_3

Kern, A. C., & Ellermeier, W. (2020). Audio in VR: Effects of a Soundscape and Movement-Triggered Step Sounds on Presence. *Frontiers in Robotics and AI*, *7*, 20. https://doi.org/10.3389/frobt.2020.00020

Khanam, F., Munmun, F. A., Ritu, N. A., Saha, A. K., & Firoz, M. (2022). Text to speech synthesis: A systematic review, deep learning based architecture and future research direction. *Journal of Advances in Information Technology*, *13*(5). https://www.academia.edu/download/95499988/20220831054604906.pdf

Kimbrel, N. A., Mitchell, J. T., & Nelson-Gray, R. O. (2010). An examination of the relationship between behavioral approach system (BAS) sensitivity and social interaction anxiety. *Journal of Anxiety Disorders*, *24*(3), 372–378.

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (2008). The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting. *Neuropsychobiology*, *28*(1–2), 76–81. https://doi.org/10.1159/000119004

Klumbies, E., Braeuer, D., Hoyer, J., & Kirschbaum, C. (2014). The Reaction to Social Stress in Social Phobia: Discordance between Physiological and Subjective Parameters. *PLoS ONE*, *9*(8), e105670. https://doi.org/10.1371/journal.pone.0105670

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Kolarik, A. J., Pardhan, S., Cirstea, S., & Moore, B. C. J. (2013). Using Acoustic Information to Perceive Room Size: Effects of Blindness, Room Reverberation Time, and Stimulus. *Perception*, *42*(9), 985–990. https://doi.org/10.1068/p7555

Kothgassner, O. D., & Felnhofer, A. (2020). Does virtual reality help to cut the Gordian knot between ecological validity and experimental control? *Annals of the International Communication Association*, *44*(3), 210–218. https://doi.org/10.1080/23808985.2020.1792790

Kraus, N., & White-Schwoch, T. (2015). Unraveling the Biology of Auditory Learning: A Cognitive–Sensorimotor–Reward Framework. *Trends in Cognitive Sciences*, *19*(11), 642–654. https://doi.org/10.1016/j.tics.2015.08.017

Kroczek, L. O. H., May, A., Hettenkofer, S., Ruider, A., Ludwig, B., & Mühlberger, A. (2025). The influence of persona and conversational task on social interactions with a LLM-controlled embodied conversational agent. *Computers in Human Behavior*, *172*, 108759. https://doi.org/10.1016/j.chb.2025.108759

Kroczek, L. O. H., & Mühlberger, A. (2023). Public speaking training in front of a supportive audience in Virtual Reality improves performance in real-life. *Scientific Reports*, *13*(1), 13968. https://doi.org/10.1038/s41598-023-41155-9

Kroczek, L. O. H., Roßkopf, S., Stärz, F., Blau, M., Van De Par, S., & Mühlberger, A. (2024). The influence of affective voice on sound distance perception. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(9), 918–933. https://doi.org/10.1037/xhp0001222

Kroczek, L. O. H., Roßkopf, S., Stärz, F., Ruider, A., Blau, M., Van de Par, S., & Mühlberger, A. (2023). *A room of one's own: A high-accuracy calibration procedure to align spatial dimensions between a virtual and a real room*. https://psyarxiv.com/erb49/

Kroczek, L. O. H., Rosskopf, S., Stärz, F., Van de Par, S., & Mühlberger, A. (2022). The influence of affective voice on sound distance perception. *Fortschritte Der Akustik*, 1163–1166.

Krohne, H. W., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1996). Untersuchungen mit einer deutschen version der" positive and negative affect schedule"(PANAS). *Diagnostica-Gottingen-*, *42*, 139–156.

Krumbholz, K. (2009). An Introduction to the Physiology of Hearing. *International Journal of Audiology*, *48*(1), 52–52. https://doi.org/10.1080/14992020802415603

Kupper, N., Jankovic, M., & Kop, W. J. (2021). Individual Differences in Cross-System Physiological Activity at Rest and in Response to Acute Social Stress. *Biopsychosocial Science and Medicine*, *83*(2), 138. https://doi.org/10.1097/PSY.0000000000000901

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Kwok, K. K. K., Ng, A. K. T., & Lau, H. Y. K. (2018). Effect of Navigation Speed and VR Devices on Cybersickness. *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 91–92. https://doi.org/10.1109/ISMAR-Adjunct.2018.00041

Kyrlitsias, C., & Michael-Grigoriou, D. (2022). Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality*, *2*, 786665.

Kytö, M., Kusumoto, K., & Oittinen, P. (2015). The Ventriloquist Effect in Augmented Reality. *2015 IEEE International Symposium on Mixed and Augmented Reality*, 49–53. https://doi.org/10.1109/ISMAR.2015.18

Larsson, P., Västfjäll, D., & Kleiner, M. (2008). Effects of auditory information consistency and room acoustic cues on presence in virtual environments. *Acoustical Science and Technology*, *29*(2), 191–194. https://doi.org/10.1250/ast.29.191

Latané, B., Liu, J. H., Nowak, A., Bonevento, M., & Zheng, L. (1995). Distance Matters: Physical Space and Social Impact. *Personality and Social Psychology Bulletin*, *21*(8), 795–805. https://doi.org/10.1177/0146167295218002

Latoschik, M. E., Roth, D., Gall, D., Achenbach, J., Waltemate, T., & Botsch, M. (2017). The effect of avatar realism in immersive social virtual realities. *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, 1–10. https://doi.org/10.1145/3139131.3139156

Latoschik, M. E., & Wienrich, C. (2022). Congruence and plausibility, not presence: Pivotal conditions for XR experiences and effects, a novel approach. *Frontiers in Virtual Reality*, *3*, 694433.

Lee, K. M. (2004). Presence, explicated. *Communication Theory*, *14*(1), 27–50.

Leiner, D. J. (2019). *SoSci Survey (Version 3.1. 06)[Computer software]. SoSci Survey GmbH*.

Leyrer, M., Linkenauger, S. A., Bülthoff, H. H., & Mohler, B. J. (2015). Eye Height Manipulations: A Possible Solution to Reduce Underestimation of Egocentric Distances in Head-Mounted Displays. *ACM Transactions on Applied Perception*, *12*(1), 1–23. https://doi.org/10.1145/2699254

Li, B., Zhang, R., Nordman, A., & Kuhl, S. A. (2015). The effects of minification and display field of view on distance judgments in real and HMD-based environments. *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, 55–58. https://doi.org/10.1145/2804408.2804427

Lindau, A., Maempel, H.-J., & Weinzierl, S. (2008). Minimum BRIR grid resolution for dynamic binaural synthesis. *Journal of the Acoustical Society of America*, *123*(5), 3498.

Lindau, A., & Weinzierl, S. (2009). On the spatial resolution of virtual acoustic environments for head movements in horizontal, vertical, and lateral direction. *Proc. of the EAA Symposium on Auralization, Espoo, Finnland*. https://core.ac.uk/reader/224999780

Lindau, A., & Weinzierl, S. (2012). Assessing the Plausibility of Virtual Acoustic Environments. *Acta Acustica United with Acustica*, *98*(5), 804–810. https://doi.org/10.3813/AAA.918562

Liou, W.-K., & Chang, C.-Y. (2018). Virtual reality classroom applied to science education. *2018 23rd International Scientific-Professional Conference on Information Technology (IT)*, 1–4. https://ieeexplore.ieee.org/abstract/document/8350861/

Liu, D., & Rau, P.-L. P. (2020). Spatially incongruent sounds affect visual localization in virtual environments. *Attention, Perception, & Psychophysics*, *82*(4), 2067–2075. https://doi.org/10.3758/s13414-019-01929-8

Liu, Q., & Zhang, W. (2020). Sex Differences in Stress Reactivity to the Trier Social Stress Test in Virtual Reality. *Psychology Research and Behavior Management*, *Volume 13*, 859–869. https://doi.org/10.2147/PRBM.S268039

Lladó, P., Mckenzie, T., Meyer-Kahlen, N., & Schlecht, S. J. (2022). Predicting Perceptual Transparency of Head-Worn Devices. *Journal of the Audio Engineering Society*, *70*(7/8), 585–600. https://doi.org/10.17743/jaes.2022.0024

Lloyd, D. M. (2009). The space between us: A neurophilosophical framework for the investigation of human interpersonal space. *Neuroscience & Biobehavioral Reviews*, *33*(3), 297–304. https://doi.org/10.1016/j.neubiorev.2008.09.007

Lønne, T. F., Karlsen, H. R., Langvik, E., & Saksvik-Lehouillier, I. (2023). The effect of immersion on sense of presence and affect when experiencing an educational scenario in virtual reality: A randomized controlled study. *Heliyon*, *9*(6).

Maddox, R. K., Pospisil, D. A., Stecker, G. C., & Lee, A. K. (2014). Directing eye gaze enhances auditory spatial cue discrimination. *Current Biology*, *24*(7), 748–752.

Mai, J. (2021). *Vorstellungsgespräch Fragen und Antworten: 100 Beispiele*. https://karrierebibel.de/vorstellungsgesprach-fragen/

Majdak, P., Goupell, M. J., & Laback, B. (2010). 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Attention, Perception, & Psychophysics*, *72*(2), 454–469. https://doi.org/10.3758/APP.72.2.454

Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior*, *72*, 276–285.

Mal, D., Wolf, E., Döllinger, N., Wienrich, C., & Latoschik, M. E. (2023). The Impact of Avatar and Environment Congruence on Plausibility, Embodiment, Presence, and the Proteus Effect in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, *29*(5), 2358–2368. https://doi.org/10.1109/TVCG.2023.3247089

Maruhn, P., Schneider, S., & Bengler, K. (2019). Measuring egocentric distance perception in virtual reality: Influence of methodologies, locomotion and translation gains. *PLOS ONE*, *14*(10), e0224651. https://doi.org/10.1371/journal.pone.0224651

Marx, E., Deutschländer, A., Stephan, T., Dieterich, M., Wiesmann, M., & Brandt, T. (2004). Eyes open and eyes closed as rest conditions: Impact on brain activation patterns. *NeuroImage*, *21*(4), 1818–1824. https://doi.org/10.1016/j.neuroimage.2003.12.026

MathWorks. (2022). *Audio toolbox user's guide* [Computer software].

Matthews, N., Todd, J., Mannion, D. J., Finnigan, S., Catts, S., & Michie, P. T. (2013). Impaired processing of binaural temporal cues to auditory scene analysis in schizophrenia. *Schizophrenia Research*, *146*(1–3), 344–348. https://doi.org/10.1016/j.schres.2013.02.013

McCall, C., & Blascovich, J. (2009). How, When, and Why to Use Digital Experimental Virtual Environments to Study Social Behavior. *Social and Personality Psychology Compass*, *3*(5), 744–758. https://doi.org/10.1111/j.1751-9004.2009.00195.x

Melo, M., Gonçalves, G., Monteiro, P., Coelho, H., Vasconcelos-Raposo, J., & Bessa, M. (2020). Do multisensory stimuli benefit the virtual reality experience? A systematic review. *IEEE Transactions on Visualization and Computer Graphics*, *28*(2), 1428–1442.

Méndez, Á., Navarro, M. D., Ferri, J., Noé, E., & Llorens, R. (2025). Influence of Demographic and Clinical Factors on Perceived Usability, Presence, Flow, Competence, Pleasant and Unpleasant Sensations, and Utility During Interaction with Virtual Reality Games for Motor and Cognitive Rehabilitation: An Observational Study in Patients with Stroke and Traumatic Brain Injury. *Games for Health Journal*. https://doi.org/10.1177/2161783X251370421

Mershon, D. H., Desaulniers, D. H., Amerson, T. L., & Kiefer, S. A. (1980). Visual capture in auditory distance perception: Proximity image effect reconsidered. *The Journal of Auditory Research*, *20*(2), 129–136.

Middlebrooks, J. C., & Onsan, Z. A. (2012). Stream segregation with high spatial acuity. *The Journal of the Acoustical Society of America*, *132*(6), 3896–3911.

Miller, R., Plessow, F., Kirschbaum, C., & Stalder, T. (2013). Classification Criteria for Distinguishing Cortisol Responders From Nonresponders to Psychosocial Stress: Evaluation of Salivary Cortisol Pulse Detection in Panel Designs. *Biopsychosocial Science and Medicine*, *75*(9), 832. https://doi.org/10.1097/PSY.0000000000000002

Minder, S., Notari, M., Schmitz, F., Hofer, R., & Woermann, U. (2012). Computer Generated Voice-Over in a Medical E-Learning Application: The Impact on Factual Learning Outcome. *J. Univers. Comput. Sci.*, *18*(3), 314–326.

Mirzaei, M., Kán, P., & Kaufmann, H. (2021). Effects of Using Vibrotactile Feedback on Sound Localization by Deaf and Hard-of-Hearing People in Virtual Environments. *Electronics*, *10*(22), 2794. https://doi.org/10.3390/electronics10222794

Mishev, K., Ristovska, A. K., Rashikj-Canevska, O., & Simjanoska, M. (2022). Assistive e-Learning Software Modules to Aid Education Process of Students with Visual and Hearing Impairment: A Case Study in North Macedonia. In L. Antovski & G. Armenski (Eds.), *ICT Innovations 2021. Digital Transformation* (Vol. 1521, pp. 145–159). Springer International Publishing. https://doi.org/10.1007/978-3-031-04206-5_11

Misu, T., Mizukami, E., Shiga, Y., Kawamoto, S., Kawai, H., & Nakamura, S. (2011). Analysis on Effects of Text-to-Speech and Avatar Agent in Evoking Users' Spontaneous Listener's Reactions. In R. L.-C. Delgado & T. Kobayashi (Eds.), *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop* (pp. 77–89). Springer New York. https://doi.org/10.1007/978-1-4614-1335-6_10

Mitchell, W. J., Szerszen, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A Mismatch in the Human Realism of Face and Voice Produces an Uncanny Valley. *I-Perception*, *2*(1), 10–12. https://doi.org/10.1068/i0415

Møller, H., Sørensen, M. F., Jensen, C. B., & Hammershøi, D. (1996). Binaural technique: Do we need individual recordings? *Journal of the Audio Engineering Society*, *44*(6), 451–469.

Morey, R. D., & Rouder, J. N. (2014). *BayesFactor version 0.9. 9: An R package for computing Bayes factor for a variety of psychological research designs*.

Nannipieri, O. (2022). Do Presence Questionnaires Actually Measure Presence? A Content Analysis of Presence Measurement Scales. In L. T. De Paolis, P. Arpaia, & M. Sacco (Eds.), *Extended Reality* (pp. 273–295). Springer International Publishing. https://doi.org/10.1007/978-3-031-15546-8_24

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3), 171.

Neidhardt, A., Ahrens, J., Pörschmann, C., Aspöck, L., Fels, J., Frank, M., Bertsch, M., Mühlberger, A., Kroczek, L. O. H., Roßkopf, S., Seeber, B., Llorca-Bofí, J., Vorlaender, M., Sladeczek, C., Bergner, J., Hock, K., Rodigast, R., Bös, J., Zotter, F., & Amengual Garí, S. (2024). *Anwendungen der Virtuellen Akustik*. *24*, 16–26.

Neidhardt, A., Schneiderwind, C., & Klein, F. (2022). Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms—Literature Review and Theoretical Framework. *Trends in Hearing*, *26*. https://doi.org/10.1177/23312165221092919

Newman, M., Gatersleben, B., Wyles, K. J., & Ratcliffe, E. (2022). The use of virtual reality in environment experiences and the importance of realism. *Journal of Environmental Psychology*, *79*, 101733. https://doi.org/10.1016/j.jenvp.2021.101733

Nicol, R., Gros, L., Colomes, C., Noisternig, M., Warusfel, O., Bahu, H., Katz, B. F., & Simon, L. S. (2014). *A roadmap for assessing the quality of experience of 3D audio binaural rendering*. https://www.academia.edu/download/87420148/index.pdf

Nowak, K., Tankelevitch, L., Tang, J., & Rintel, S. (2023). Hear We Are: Spatial Audio Benefits Perceptions of Turn-Taking and Social Presence in Video Meetings. *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, 1–10. https://doi.org/10.1145/3596671.3598578

Oh, C. S., Bailenson, J. N., & Welch, G. F. (2018). A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Frontiers in Robotics and AI*, *5*, 114. https://doi.org/10.3389/frobt.2018.00114

Oh, H. J., Kim, J., Chang, J. J., Park, N., & Lee, S. (2023). Social benefits of living in the metaverse: The relationships among social presence, supportive interaction, social self-efficacy, and feelings of loneliness. *Computers in Human Behavior*, *139*, 107498.

Ohayon, M. M., & Schatzberg, A. F. (2010). Social phobia and depression: Prevalence and comorbidity. *Journal of Psychosomatic Research*, *68*(3), 235–243. https://doi.org/10.1016/j.jpsychores.2009.07.018

Pan, X., & Hamilton, A. F. D. C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, *109*(3), 395–417. https://doi.org/10.1111/bjop.12290

Parmar, P., Ryu, J., Pandya, S., Sedoc, J., & Agarwal, S. (2022). Health-focused conversational agents in person-centered care: A review of apps. *Npj Digital Medicine*, *5*(1), 1–9. https://doi.org/10.1038/s41746-022-00560-6

Parsons, T. D. (2011). Neuropsychological Assessment Using Virtual Environments: Enhanced Assessment Technology for Improved Ecological Validity. In S. Brahnam & L. C. Jain (Eds.), *Advanced Computational Intelligence Paradigms in Healthcare 6.*

Virtual Reality in Psychotherapy, Rehabilitation, and Assessment (Vol. 337, pp. 271–289). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-17824-5_13

Peña, J., Craig, M., & Baumhardt, H. (2024). The effects of avatar customization and virtual human mind perception: A test using Milgram's paradigm. *New Media & Society*, *26*(8), 4730–4749. https://doi.org/10.1177/14614448221127258

Peperkorn, H. M., Diemer, J., & Mühlberger, A. (2015). Temporal dynamics in the relation between presence and fear in virtual reality. *Computers in Human Behavior*, *48*, 542–547. https://doi.org/10.1016/j.chb.2015.02.028

Pfaller, M., Kroczek, L. O. H., Lange, B., Fülöp, R., Müller, M., & Mühlberger, A. (2021). Social presence as a moderator of the effect of agent behavior on emotional experience in social interactions in virtual reality. *Frontiers in Virtual Reality*, *2*, 741138.

Poeschl, S., Wall, K., & Doering, N. (2013). Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. *2013 IEEE Virtual Reality (VR)*, 129–130.

Poirier-Quinot, D., & Lawless, M. S. (2023). Impact of wearing a head-mounted display on localization accuracy of real sound sources. *Acta Acustica*, *7*, 3. https://doi.org/10.1051/aacus/2022055

Pörschmann, C., & Arend, J. M. (2020). Analyzing the directivity patterns of human speakers. *Proceedings of the 46th DAGA*, 16–19.

Pribék, I. K., Szűcs, K. F., Süle, M., Grosz, G., Ducza, E., Vigh, D., Tóth, E., Janka, Z., Kálmán, J., Datki, Z. L., Gáspár, R., & Andó, B. (2021). Detection of acute stress by smooth muscle electromyography: A translational study on rat and human. *Life Sciences*, *277*, 119492. https://doi.org/10.1016/j.lfs.2021.119492

Prud'homme, L., & Lavandier, M. (2020). Do we need two ears to perceive the distance of a virtual frontal sound source? *The Journal of the Acoustical Society of America*, *148*(3), 1614–1623. https://doi.org/10.1121/10.0001954

Qiu, L., & Benbasat, I. (2005a). An investigation into the effects of Text-To-Speech voice and 3D avatars on the perception of presence and flow of live help in electronic commerce. *ACM Transactions on Computer-Human Interaction*, *12*(4), 329–355. https://doi.org/10.1145/1121112.1121113

Qiu, L., & Benbasat, I. (2005b). The effects of text-to-speech voice and 3D avatars on consumer trust in the design of live help interface of electronic commerce. *International Journal of Human–Computer Interaction*, *19 (1)*, 75–94.

Quadflieg, S., Mohr, A., Mentzel, H.-J., Miltner, W. H., & Straube, T. (2008). Modulation of the neural network involved in the processing of anger prosody: The role of task-relevance and social phobia. *Biological Psychology*, *78*(2), 129–137.

Rajguru, C., Brianza, G., & Memoli, G. (2022). Sound localization in web-based 3D environments. *Scientific Reports*, *12*(1), 12107. https://doi.org/10.1038/s41598-022-15931-y

Ramírez, M., Arend, J. M., Von Gablenz, P., Liesefeld, H. R., & Pörschmann, C. (2024). Toward Sound Localization Testing in Virtual Reality to Aid in the Screening of Auditory Processing Disorders. *Trends in Hearing*, *28*. https://doi.org/10.1177/23312165241235463

Randerath, J., Finkel, L., Shigaki, C., Burris, J., Nanda, A., Hwang, P., & Frey, S. H. (2021). Is This Within Reach? Left but Not Right Brain Damage Affects Affordance Judgment

Tendencies. *Frontiers in Human Neuroscience*, *14*. https://doi.org/10.3389/fnhum.2020.531893

Reichenberger, J., Pfaller, M., & Mühlberger, A. (2020). Gaze Behavior in Social Fear Conditioning: An Eye-Tracking Study in Virtual Reality. *Frontiers in Psychology*, *11*.

Reichenberger, J., Porsch, S., Wittmann, J., Zimmermann, V., & Shiban, Y. (2017). Social Fear Conditioning Paradigm in Virtual Reality: Social vs. Electrical Aversive Conditioning. *Frontiers in Psychology*, *8*.

Reichenberger, J., Wechsler, T. F., Diemer, J., Mühlberger, A., & Notzon, S. (2022). Fear, psychophysiological arousal, and cognitions during a virtual social skills training in social anxiety disorder while manipulating gaze duration. *Biological Psychology*, *175*, 108432. https://doi.org/10.1016/j.biopsycho.2022.108432

Renner, R. S., Velichkovsky, B. M., & Helmert, J. R. (2013). The perception of egocentric distances in virtual environments—A review. *ACM Computing Surveys*, *46*(2), 23:1-23:40. https://doi.org/10.1145/2543581.2543590

Rogers, S. L., Broadbent, R., Brown, J., Fraser, A., & Speelman, C. P. (2022). Realistic Motion Avatars are the Future for Social Interaction in Virtual Reality. *Frontiers in Virtual Reality*, *2*. https://doi.org/10.3389/frvir.2021.750729

Rojas, R., Geissner, E., & Hautzinger, M. (2022). *DAS-18. Dysfunctional Attitude Scale 18–deutsche Kurzfassung*. *Publisher: ZPID (Leibniz Institute for Psychology)–Open Test Archive*. https://doi.org/10.23668/psycharchives.5252

Roßkopf, S., Kroczek, L. O. H., Stärz, F., Blau, M., Van De Par, S., & Mühlberger, A. (2023). Comparable sound source localization of plausible auralizations and real sound sources evaluated in a naturalistic eye-tracking task in virtual reality. *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, 1485–1492. https://doi.org/10.61782/fa.2023.0377

Roßkopf, S., Kroczek, L. O. H., Stärz, F., Blau, M., Van de Par, S., & Mühlberger, A. (2023). The Effect of Audio-Visual Room Divergence on the Localization of Real Sound Sources in Virtual Reality. *Fortschritte Der Akustik*, 1431–1434.

Roßkopf, S., Kroczek, L. O. H., Stärz, F., Blau, M., Van de Par, S., & Mühlberger, A. (2024). Auditory Distance Perception in VR: The Influence of Visual Room Presentation and Estimation Method. *Fortschritte Der Akustik*. DAGA 2024, Hannover.

Roßkopf, S., Kroczek, L. O. H., Stärz, F., Blau, M., Van De Par, S., & Mühlberger, A. (2024). The impact of binaural auralizations on sound source localization and social presence in audiovisual virtual reality: Converging evidence from placement and eye-tracking paradigms. *Acta Acustica*, *8*, 72. https://doi.org/10.1051/aacus/2024064

Roßkopf, S., Kroczek, L. O. H., Stärz, F., Blau, M., van der Par, S., & Mühlberger, A. (2025). Effects of Individualizing Binaural Auralizations on Presence, Realism, and Affective Reactions in Stressful Social Virtual Interactions. *Fortschritte Der Akustik, DAGA 2025*.

Saini, N. (2023, February 4). *ChatGPT becomes fastest growing app in the world, records 100mn users in 2 month | Today News*. Mint. https://www.livemint.com/news/chatgpt-becomes-fastest-growing-app-in-the-world-records-100mn-users-in-2-month-11675484444142.html

Santl, J., Shiban, Y., Plab, A., Wüst, S., Kudielka, B. M., & Mühlberger, A. (2019). Gender differences in stress responses during a virtual reality trier social stress test. *International Journal of Virtual Reality*, *19*(2), 2–15.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, *10*(1), 24–30.

Schleicher, R., Spors, S., Jahn, D., & Walter, R. (2010). Gaze as a measure of sound source localization. *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*.

Schmidt-Peter, T., Wechsler, T. F., Kroczek, L. O. H., & Mühlberger, A. (2025). The effects of different variants of eye-tracking-based feedback of attentional processes during virtual social interactions. *Frontiers in Virtual Reality*, *6*, 1556898. https://doi.org/10.3389/frvir.2025.1556898

Schmitz, A., Kueppers, L., Klein, J., Frey, S., Karimzadeh, A., Neves, A. L., & Weltermann, B. (2025). Digital Health Applications (DiGA) for Treating Depression and Generalized Anxiety Disorder: Protocol for a Systematic Health App Review and Systematic Review of Published Evidence. *JMIR Research Protocols*, *14*(1), e63380.

Schneiderwind, C., Neidhardt, A., & Dominik, M. (2020). Comparing the effect of different open headphone models on the perception of a real sound source. *Journal of the Audio Engineering Society*, *10489*.

Schneiderwind, C., Richter, M., Merten, N., & Neidhardt, A. (2023). Effects of Modified Late Reverberation on Audio-Visual Plausibility and Externalization in AR. *2023 Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, 1–9. https://ieeexplore.ieee.org/abstract/document/10289186/

Schott, E., López García, I., Semple, L. A., & Froehlich, B. (2025). Estimating Detection Thresholds of Being Looked at in Virtual Reality for Avatar Redirection. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3706598.3714041

Schreiter, S., Mascarell-Maricic, L., Rakitzis, O., Volkmann, C., Kaminski, J., & Daniels, M. A. (2023). Digital Health Applications in the Area of Mental Health: A Scoping Review. *Deutsches Ärzteblatt International*, *120*(47), 797.

Schuller, D., & Schuller, B. W. (2018). The Age of Artificial Emotional Intelligence. *Computer*, *51*(9), 38–46. Computer. https://doi.org/10.1109/MC.2018.3620963

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. *The American Statistician*, *34*(4), 216–221. https://doi.org/10.1080/00031305.1980.10483031

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, *59*(20), 22–33.

Shiban, Y., Diemer, J., Brandl, S., Zack, R., Mühlberger, A., & Wüst, S. (2016). Trier Social Stress Test in vivo and in virtual reality: Dissociation of response domains. *International Journal of Psychophysiology*, *110*, 47–55.

Shiban, Y., Fruth, M. B., Pauli, P., Kinateder, M., Reichenberger, J., & Mühlberger, A. (2016). Treatment effect on biases in size estimation in spider phobia. *Biological Psychology*, *121*, 146–152.

Skalski, P., & Whitbred, R. (2010). Image versus sound: A comparison of formal feature effects on presence and video game enjoyment. *PsychNology Journal*, *8*(1), 67–84.

Skarbez, R., Brooks, Jr., F. P., & Whitton, M. C. (2018). A Survey of Presence and Related Concepts. *ACM Computing Surveys*, *50*(6), 1–39. https://doi.org/10.1145/3134301

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3549–3557. https://doi.org/10.1098/rstb.2009.0138

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British Journal of Psychology*, *109*(3), 431–433. https://doi.org/10.1111/bjop.12305

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., Pistrang, N., & Sanchez-Vives, M. V. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PloS One*, *1*(1), e39.

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *6*(6), 603–616.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport*, *12*(1), 7–10.

Solhjoo, S., Haigney, M. C., McBee, E., van Merrienboer, J. J. G., Schuwirth, L., Artino, A. R., Battista, A., Ratcliffe, T. A., Lee, H. D., & Durning, S. J. (2019). Heart Rate and Heart Rate Variability Correlate with Clinical Reasoning Performance and Self-Reported Measures of Cognitive Load. *Scientific Reports*, *9*(1), 14668. https://doi.org/10.1038/s41598-019-50280-3

Song, H., & Ma, K. (2022). Best Distance Perception in Virtual Audiovisual Environment. *Computational Intelligence and Neuroscience*, *6010667*(1).

Sorokowska, A., Sorokowski, P., Hilpert, P., Cantarero, K., Frackowiak, T., Ahmadi, K., Alghraibeh, A. M., Aryeetey, R., Bertoni, A., Bettache, K., Blumen, S., Błażejewska, M., Bortolini, T., Butovskaya, M., Castro, F. N., Cetinkaya, H., Cunha, D., David, D., David, O. A., … PierceJr., J. D. (2017). Preferred Interpersonal Distances: A Global Comparison. *Journal of Cross-Cultural Psychology*, *48*(4), 577–592. https://doi.org/10.1177/0022022117698039

Sosic, Z., Gieler, U., & Stangier, U. (2008). Screening for social phobia in medical in-and outpatients with the German version of the Social Phobia Inventory (SPIN). *Journal of Anxiety Disorders*, *22*(5), 849–859.

Spielberger, C. D., Gonzalez-Reigosa, F., Martinez-Urrutia, A., Natalicio, L. F., & Natalicio, D. S. (1971). The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican Journal of Psychology*, *5*(3 & 4).

Stärz, F., Kroczek, L. O. H., Rosskopf, S., Mühlberger, A., & Blau, M. (2022). Perceptual comparison between the real and the auralized room when being presented with congruent visual stimuli via a head-mounted display. *Proceedings of the 24th International Congress on Acoustics. International Commission for Acoustics (ICA)*.

Stärz, F., Kroczek, L. O. H., Roßkopf, S., Mühlberger, A., van de Par, S., & Blau, M. (2023). Mounting extra-aural headphones to a head-mounted display using a 3D-printed support. *Fortschritte Der Akustik*, 1636–1639.

Stärz, F., Kroczek, L. O. H., Roßkopf, S., Mühlberger, A., Van De Par, S., & Blau, M. (2024). Comparing Room Acoustical Ratings in an Interactive Virtual Environment to Those in the Real Room. *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, 5009–5016. https://doi.org/10.61782/fa.2023.0525

Stärz, F., Van De Par, S., Roßkopf, S., Kroczek, L. O. H., Mühlberger, A., & Blau, M. (2025). Comparison of binaural auralisations to a real loudspeaker in an audiovisual virtual classroom scenario: Effect of room acoustic simulation, HRTF dataset, and head-mounted display on room acoustic perception. *Acta Acustica*, *9*, 31.

Stauffert, J.-P., Niebling, F., & Latoschik, M. E. (2018). Effects of Latency Jitter on Simulator Sickness in a Search Task. *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 121–127. https://doi.org/10.1109/VR.2018.8446195

Steadman, M. A., Kim, C., Lestang, J.-H., Goodman, D. F. M., & Picinali, L. (2019). Short-term effects of sound localization training in virtual reality. *Scientific Reports*, *9*(1), 18284. https://doi.org/10.1038/s41598-019-54811-w

Stodt, B., Neudek, D., Getzmann, S., Wascher, E., & Martin, R. (2024). Comparing auditory distance perception in real and virtual environments and the role of the loudness cue: A study based on event-related potentials. *Hearing Research*, *444*, 108968. https://doi.org/10.1016/j.heares.2024.108968

Sunder, K., Tan, E.-L., & Gan, W.-S. (2014). Effect of headphone equalization on auditory distance perception. In *137th Audio Engineering Society Convention 2014*.

Taffou, M., & Viaud-Delmon, I. (2014). Cynophobic fear adaptively extends peri-personal space. *Frontiers in Psychiatry*, *5*, 122.

Triantafyllopoulos, A., Schuller, B. W., İymen, G., Sezgin, M., He, X., Yang, Z., Tzirakis, P., Liu, S., Mertes, S., André, E., Fu, R., & Tao, J. (2023). An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era. *Proceedings of the IEEE*, *111*(10), 1355–1381. Proceedings of the IEEE. https://doi.org/10.1109/JPROC.2023.3250266

van Dammen, L., Finseth, T. T., McCurdy, B. H., Barnett, N. P., Conrady, R. A., Leach, A. G., Deick, A. F., Van Steenis, A. L., Gardner, R., & Smith, B. L. (2022). Evoking stress reactivity in virtual reality: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *138*, 104709.

Västfjäll, D. (2003). The Subjective Sense of Presence, Emotion Recognition, and Experienced Emotions in Auditory Virtual Environments. *CyberPsychology & Behavior*, *6*(2), 181–188. https://doi.org/10.1089/109493103321640374

VIVE Pro Eye Support. (2024.). *Vive.com*. Retrieved July 15, 2024, from https://www.vive.com/us/support/vive-pro-eye/

Volkmann, T., Wessel, D., Franke, T., & Jochems, N. (2019). Testing the Social Presence Aspect of the Multimodal Presence Scale in a Virtual Reality Game. *Proceedings of Mensch Und Computer 2019*, 433–437. https://doi.org/10.1145/3340764.3344435

Volkmann, T., Wessel, D., Jochems, N., & Franke, T. (2018). German Translation of the Multimodal Presence Scale. *MuC*.

Vorländer, M. (2008). *Auralization: Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality* (1st ed). Springer.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wang, G., Obrenovic, B., Gu, X., & Godinic, D. (2025). Fear of the new technology: Investigating the factors that influence individual attitudes toward generative Artificial Intelligence (AI). *Current Psychology*, *44*(9), 8050–8067. https://doi.org/10.1007/s12144-025-07357-2

Wechsler, T. F., Kümpers, F., & Mühlberger, A. (2019). Inferiority or Even Superiority of Virtual Reality Exposure Therapy in Phobias?—A Systematic Review and Quantitative Meta-Analysis on Randomized Controlled Trials Specifically Comparing the Efficacy of Virtual Reality Exposure to Gold Standard in vivo Exposure in Agoraphobia, Specific Phobia, and Social Phobia. *Frontiers in Psychology*, *10*, 1758. https://doi.org/10.3389/fpsyg.2019.01758

Weinzierl, S., Lepa, S., & Ackermann, D. (2018). A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI). *The Journal of the Acoustical Society of America*, *144*(3), 1245–1257.

Wendt, J., Weyers, B., Stienen, J., Bönsch, A., Vorländer, M., & Kuhlen, T. W. (2019). Influence of Directivity on the Perception of Embodied Conversational Agents' Speech. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 130–132. https://doi.org/10.1145/3308532.3329434

Wendt, T., Van De Par, S., & Ewert, S. (2014). A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation. *Journal of the Audio Engineering Society*, *62*(11), 748–766. https://doi.org/10.17743/jaes.2014.0042

Wenger, E., Bronckers, M., Cianfarani, C., Cryan, J., Sha, A., Zheng, H., & Zhao, B. Y. (2021). "Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 235–251. https://doi.org/10.1145/3460120.3484742

Werner, S. (2018). Auditory illusion through headphones: History, Challenges and new Solutions. *Proceedings of Meetings on Acoustics*, *28*(050010), 1.

Werner, S., Klein, F., Mayenfels, T., & Brandenburg, K. (2016). A summary on acoustic room divergence and its effect on externalization of auditory events. *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6. https://doi.org/10.1109/QoMEX.2016.7498973

Westlund, J. M., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., DeSteno, D., & Breazeal, C. L. (2017). Flat vs. Expressive Storytelling: Young Children's Learning and Retention of a Social Robot's Narrative. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00295

Wiebe, A., Kannen, K., Selaskowski, B., Mehren, A., Thöne, A.-K., Pramme, L., Blumenthal, N., Li, M., Asché, L., Jonas, S., Bey, K., Schulze, M., Steffens, M., Pensel, M. C., Guth, M., Rohlfsen, F., Ekhlas, M., Lügering, H., Fileccia, H., … Braun, N. (2022). Virtual reality in the diagnostic and therapy for mental disorders: A systematic review. *Clinical Psychology Review*, *98*, 102213. https://doi.org/10.1016/j.cpr.2022.102213

Wiesing, M., Comadran, G., & Slater, M. (2025). Confusing virtual reality with reality – An experimental study. *iScience*, *28*(6), 112655. https://doi.org/10.1016/j.isci.2025.112655

Wijsman, J., Grundlehner, B., Penders, J., & Hermens, H. (2013). Trapezius muscle EMG as predictor of mental stress. *ACM Trans. Embed. Comput. Syst.*, *12*(4), 99:1-99:20. https://doi.org/10.1145/2485984.2485987

Williams, J., Rownicka, J., Oplustil, P., & King, S. (2020). *Comparison of Speech Representations for Automatic Quality Estimation in Multi-Speaker Text-to-Speech Synthesis* (No. arXiv:2002.12645). arXiv. https://doi.org/10.48550/arXiv.2002.12645

Wirler, S., meyer-Kahlen, N., & Schlecht, S. (2020). Towards transfer-plausibility for evaluating mixed reality audio in complex scenes. *Journal of the Audio Engineering Society*, *3–4*.

Wöstmann, M., Schmitt, L.-M., & Obleser, J. (2020). Does Closing the Eyes Enhance Auditory Attention? Eye Closure Increases Attentional Alpha-Power Modulation but Not Listening Performance. *Journal of Cognitive Neuroscience*, *32*(2), 212–225. https://doi.org/10.1162/jocn_a_01403

Xiao, X., Noh, H., Lefevre, A., Li, L., McKee, H., Algargoosh, A., & Ishii, H. (2025). ReMirrorFugue: Examining the Emotional Experience of Presence and (Illusory) Communications Across Time. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–26. https://doi.org/10.1145/3706598.3713328

Yamada, Y., Kawabe, T., & Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon. *Japanese Psychological Research*, *55*(1), 20–32. https://doi.org/10.1111/j.1468-5884.2012.00538.x

Yoon, B., Kim, H., Lee, G. A., Billinghurst, M., & Woo, W. (2019). The Effect of Avatar Appearance on Social Presence in an Augmented Reality Remote Collaboration. *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 547–556. https://doi.org/10.1109/VR.2019.8797719

Zahorik, P. (2001). Estimating sound source distance with and without vision. *Optometry and Vision Science*, *78*(5), 270–275.

Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *ACTA Acustica United with Acustica*, *91*(3), 409–420.

Zamoscik, V., Niemeyer, C., Gerchen, M. F., Fenske, S. C., Witthöft, M., & Kirsch, P. (2017). Sensorik Inventar (SI)–Selbstbeurteilung der sensorischen Sensitivität für Erwachsene und Jugendliche. *Fortschritte Der Neurologie· Psychiatrie*, *85*(09), 541–551.

Zhang, J., Li, S., Zhang, J.-Y., Du, F., Qi, Y., & Liu, X. (2020). A Literature Review of the Research on the Uncanny Valley. In P.-L. P. Rau (Ed.), *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments* (pp. 255–268). Springer International Publishing. https://doi.org/10.1007/978-3-030-49788-0_19

Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and ESD. *Speech Communication*, *137*, 1–18. https://doi.org/10.1016/j.specom.2021.11.006

Zimmer, P., Buttlar, B., Halbeisen, G., Walther, E., & Domes, G. (2019). Virtually stressed? A refined virtual reality adaptation of the Trier Social Stress Test (TSST) induces robust endocrine responses. *Psychoneuroendocrinology*, *101*, 186–192.

Zorn, J. V., Schür, R. R., Boks, M. P., Kahn, R. S., Joëls, M., & Vinkers, C. H. (2017). Cortisol stress reactivity across psychiatric disorders: A systematic review and meta-analysis. *Psychoneuroendocrinology*, *77*, 25–36.

10 Appendix

10.1 Supplementary tables

Table 9: Job interview questions

N°	Original (German)	English Translation
1	Was sind Ihre Stärken und Schwächen?	What are your strengths and weaknesses?
2	Wie gehen Sie damit um, wenn Sie einen Fehler gemacht haben?	How do you handle it when you make a mistake?
3	Wo sehen Sie sich in 5 Jahren?	Where do you see yourself in 5 years?
4	Was war der größte Misserfolg in Ihrem Leben und wie sind Sie damit umgegangen?	What was the biggest failure in your life and how did you deal with it?
5	Welches Verhalten einer anderen Person würde Sie auf 180 bringen?	What behavior from another person would make you really angry?
6	Wann haben Sie das letzte Mal eine Regel missachtet – und warum?	When was the last time you broke a rule – and why?
7	Was haben Sie letzte Woche gelernt?	What did you learn last week?
8	Wie stehen Sie zur Legalisierung von Cannabis?	What is your opinion on the legalization of cannabis?
9	Warum sind Sie besser als andere?	Why are you better than others?
10	Wie sieht Ihr Traumberuf aus?	What does your dream job look like?
11	Was müsste passieren, damit Sie den Schritt zu uns bereuen?	What would have to happen for you to regret joining us?
12	Was sollte ich unbedingt über Sie wissen?	What should I absolutely know about you?
13	Welche 3 positiven Charaktereigenschaften fehlen Ihnen?	Which 3 positive character traits do you lack?
14	Wie würden Sie Ihren Arbeitsstil beschreiben?	How would you describe your working style?
15	Was tun Sie, wenn Sie merken, die Tagesaufgaben unmöglich zu schaffen?	What do you do when you realize the day's tasks are impossible to complete?
16	Wann und wie haben Sie das letzte Mal einen Kollegen kritisiert?	When and how did you last criticize a colleague?
17	Welche Aufgabe war für Sie zu schwer – und wie haben Sie das Problem gelöst?	Which task was too difficult for you – and how did you solve the problem?
18	Was sind die zentralen Eigenschaften einer guten Führungskraft?	What are the key traits of a good leader?
19	Und was sind die zentralen Eigenschaften einer schlechten Führungskraft?	And what are the key traits of a bad leader?
20	Was ist Ihr Vorbild?	Who is your role model?
21	Haben Sie heute einen schlechten Tag oder treten Sie immer so auf?	Are you having a bad day today or do you always come across like this?
22	Erzählen Sie mir etwas von sich, das nicht im Lebenslauf steht.	Tell me something about yourself that’s not on your CV.
23	Wie mache ich mich in Ihren Augen als Interviewer?	How am I doing as an interviewer in your eyes?
24	Was war Ihr bisher schwächster Teil in diesem Vorstellungsgespräch?	What has been your weakest part in this interview so far?
25	Welche Frage möchten Sie nicht gestellt bekommen?	Which question would you prefer not to be asked?
26	Was führte bisher zu Problemen im Team?	What has previously led to problems in a team?
27	Wie haben Sie sich gefühlt, als Sie für Ihre Arbeit kritisiert worden sind?	How did you feel when your work was criticized?
28	Was ist Ihr größter Erfolg, der nichts mit Ihrem Beruf zu tun hat?	What is your greatest success that has nothing to do with your profession?
29	Wovor haben Sie Angst?	What are you afraid of?
30	Wie schaffen Sie so schnell wie möglich eine Vertrauensbasis in einem neuen Team?	How do you quickly build a foundation of trust in a new team?

Table 10: Ratings

°	Rating Item	Question
Stress	Stress	How stressed do you feel at the moment?
	Stress – Committee Introduction	How stressed did you feel in front of the committee?
	Stress – post VST	How stressed did you feel during the job interview?
VR Experience	Presence	How much do you currently experience virtual reality as if you were really “there”?
	Social Presence	How much did you just feel like you were with other people (regarding the committee)?
	Realism	How realistic did you find the virtual environment?
	Social Realism	How realistic did you find the virtual agents?
Subjective Audio Quality	Externalization	Did you hear the audio in your head or outside in the room?
	Acoustic Presence	Did the virtual committee sound as if people present had spoken to you?
	Acoustic Realism	The sound of the speech was like in a real room.
	Audio Liking	How much did you like the sound experience?
	Speech Intelligibility	How did you find the speech intelligibility? (0: hard – 100: effortless)
	Tone Richness	How did you find the tone richness? (0: low – 100: high)

10.2 Supplementary figures

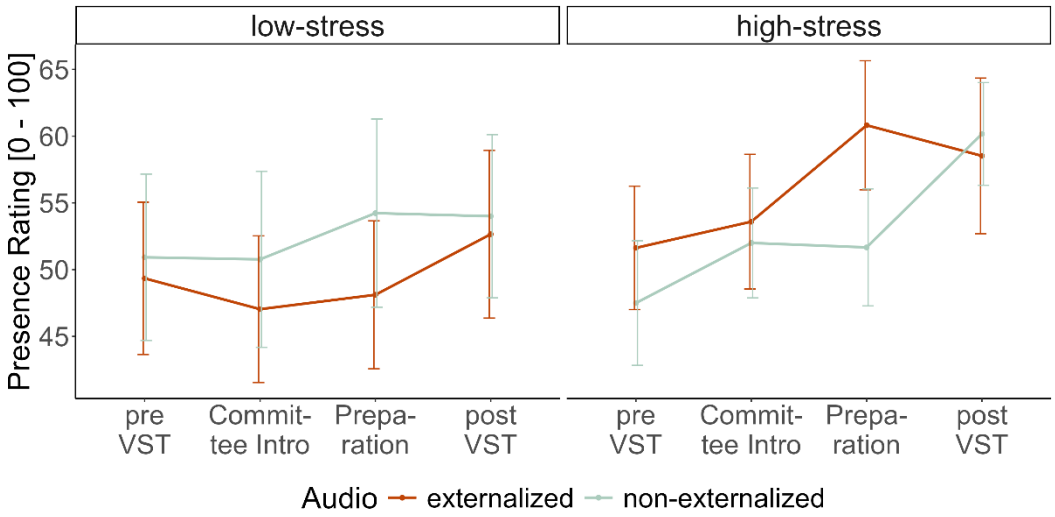

Figure 32: Presence ratings as a function of stress and audio manipulation at four different measurement time points. Error bars indicate the standard error.

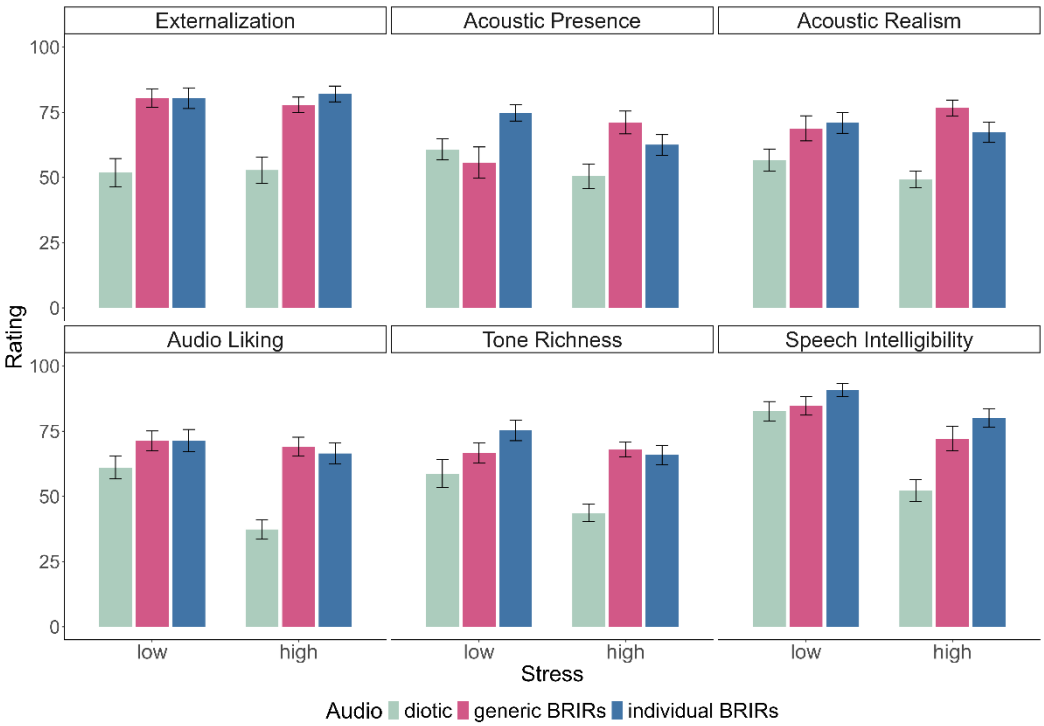

Figure 33: Subjective audio quality ratings. Error bars indicate the standard error.

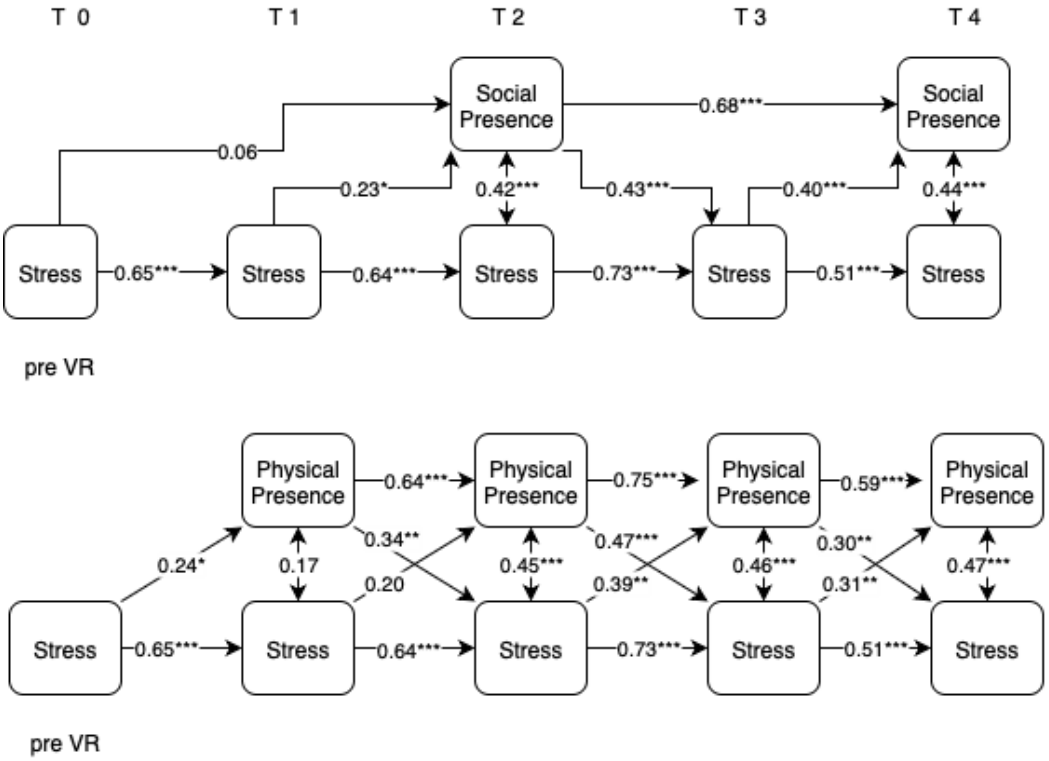

Figure 34: Time-lagged correlations between presence and stress ratings.
* p < .05, ** p < .01, *** p < .001

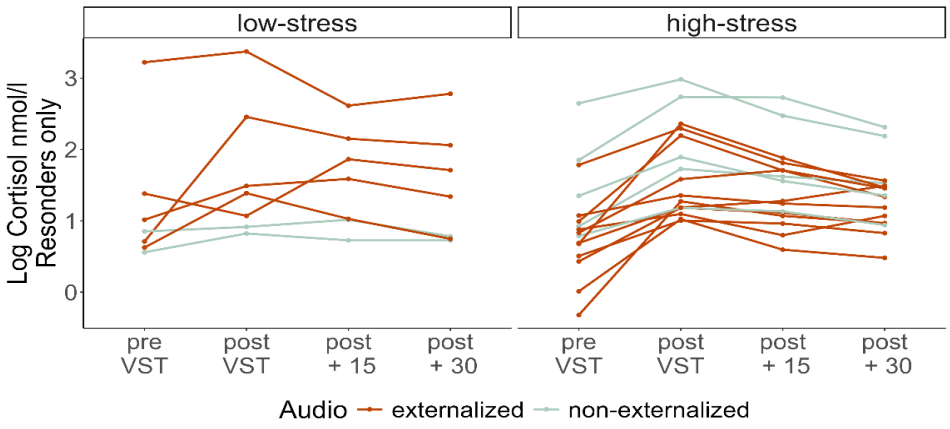

Figure 35: Individual log-transformed salivary cortisol levels in nmol/l for responders at a function of audio condition and stress for four measurement points.

10.1 Supplementary analysis

Presence, realism, and subjective audio quality

An exploratory mixed ANOVA was computed to investigate the effects of *Audio, Time, and Stress* on physical presence. A significant main effect of time on physical presence was found, $F(2.59, 191.82) = 3.03$, $p = .038$, $\eta_p^2 = 0.05$, but no other significant influence.

We exploratorily analyzed the relationship between (social) presence and subjective audio quality. Physical presence measured during the VR scenario positively correlated with acoustic realism at all time points (mean rating) r=0.40, $p < .001$, and also measured via the subscale of the MPS, $r = 0.33$, $p = .003$. While the correlation between acoustic realism and social presence measured with the MPS subscale was not significant ($r = 0.21$, $p = .069$), it reached significance for the mean social presence rating during VR, $r = 0.23$, $p = .041$. Though both social presence measurements correlated with acoustic presence (MPS, $r = .41$, $p < .001$; mean, $r = 0.47$, $p < .001$). The externalization rating only correlated with audio liking, $r = .48$, $p < .001$, and tone richness, $r = 0.34$, $p = .003$. We furthermore conducted mixed ANOVAs to investigate effects of *Stress x Audio* on subjective audio quality ratings. A main effect of Audio was found on acoustic realism, $F(1, 74) = 11.66$, $p = .002$, $\eta_p^2 = 0.14$; audio liking, $F(1, 74) = 15.16$, $p < .001$, $\eta_p^2 = 0.17$; tone richness, $F(1, 71) = 11.24$, $p = .001$, $\eta_p^2 = 0.14$; and speech intelligibility, $F(1, 74) = 8.38$, $p = .005$, $\eta_p^2 = 0.10$. Stress also had a main effect on audio liking, $F(1, 74) = 6.86$, $p = .011$, $\eta_p^2 = 0.08$; and speech intelligibility, $F(1, 74) = 17.28$, $p < .001$, $\eta_p^2 = 0.19$. Marginal significant interaction effects of Audio by Stress were found on speech intelligibility ($p = .060$) and audio liking ($p = .058$). Last, we examined whether the auralization used had an effect on anticipatory anxiety, which was not the case.

Effects of social anxiety

We explored differential effects of audio externalization and social anxiety on all outcome variables. Therefore, participants were categorized as either low- or high in social anxiety based on their SPIN scores using a median split (median = 28). A mixed ANOVA with the factors *Audio* and *Social Anxiety* (high vs. low social anxiety) revealed an effect of *Audio by Social Anxiety* on social presence, both measured with the during-VR rating, $F(1, 74) = 5.70$, $p = .020$, $\eta_p^2 = 0.07$, and after VR with the subscale of the MPS, $F(1, 74) = 4.09$, $p = .047$, $\eta_p^2 = 0.05$, see 10.2, Figure 31. Follow-up *t*-tests revealed significantly higher social presence (MPS) in the high-social anxious ($M = 2.99$, $SD = 0.96$) compared to the low-social anxious participants ($M = 2.45$, $SD = 0.81$), which was only found in the externalized auralization condition, $t(20.23) = -1.05$, $p = .032$. Similar results were found for the mean social presence rating during VR, with the only significant difference found within the externalized auralization. Again, the high-social anxiety group had higher levels of social presence ($M = 56.05$, $SD = 21.1$) compared to the low-social anxiety group ($M = 41.38$, $SD = 16.7$), $t(50.34) = 2.82$, $p = .007$. However, no interaction effects were found for several stress indicators and acoustic quality ratings, all $ps > .05$. Furthermore, we explored the potential influence of social anxiety on stress response. A repeated measures ANOVA with the within-subject factor *Time* and the between-subject factor *Social Anxiety* (median split via SPIN mean: high vs. low social anxiety) revealed no significant effect of *Time*, *Stress*, or *Time by Stress* on salivary cortisol levels. Also, a linear regression model using the raw SPIN value as a predictor was not significant. Concerning heart rate increase, a main effect of *Social Anxiety*, $F(1, 132) = 8.27$, $p = .005$, $\eta_p^2 = 0.06$, and of *Time*, $F(1, 132) = 37.58$, $p < .001$, $\eta_p^2 = 0.22$ was found, but no interaction effect. Similar effects were found for the dependent variable subjective stress, which was also significantly affected by *Time*, $F(1, 152) = 20.41$, $p < .001$, $\eta_p^2 = 0.12$, and *Social Anxiety*, $F(1, 152) = 56.04$, $p < .001$, $\eta_p^2 = 0.27$, but again no interaction effect was found. Also, visual attention was affected by social anxiety, but not in interaction with the audio condition. Participants with high social anxiety had a significantly shorter latency ($M = 992$, $SD = 214$) until the first fixation on the speaking agents than participants with high social anxiety ($M = 1184$, $SD = 262$), $t(63.53) = -3.43$, $p = .001$, $d = 0.73$. Concerning a possible relationship between social anxiety and presence, the SPIN neither correlated significantly with the social presence rating, $r = 0.21$, $p = .063$, nor with the physical or social presence subscale of the MPS, $r = 0.16$, $p = .156$, $r = 0.20$, $p = .084$, respectively. Also, no significant differences between high- and low-social anxious participants concerning presence outcomes were found.

11 Acknowledgements

I want to express my sincere gratitude to everyone who has supported and enriched me on my path to a doctorate and contributed to the completion of this work. First and foremost, I am deeply grateful to Prof. Dr. Andreas Mühlberger for his valuable supervision throughout my doctorate. I greatly appreciate the framework he provided, which enabled me to develop and pursue my own scientific interests. My heartfelt thanks also go to Dr. Leon Kroczek for his continuous support and motivation during this time. Thanks for always being available with an open ear and a helping hand.

I am grateful to our cooperation partners in Oldenburg, Prof. Dr. Matthias Blau and Prof. Dr. Steven van de Par, for their valuable advice and feedback on my work. In particular, I would like to thank Felix Stärz, my fellow doctoral student in Oldenburg, for his endless patience and guidance on technical issues, as well as his general support during my postgraduate studies.

I want to thank the Deutsche Forschungsgemeinschaft (DFG) for funding the interdisciplinary priority program AUDICTIVE – Auditory Cognition in Interactive Virtual Environments (SPP 2236, see https://gepris.dfg.de/gepris/projekt/422686707), as well as for funding our subproject under project ID 444832396. Everyone I have met through Audictive has contributed to an inspiring and enriching experience. The project has provided countless stimulating discussions, open-minded exchanges, and unwavering support during challenging times. I want to express my special thanks to the women of Audictive for their encouragement and solidarity.

In addition, I would like to thank all my former and current colleagues in our department for their support and for making this time enriching. I would especially like to thank Blerta Miftari, Clara Prager, Marieke Bruckmann, Nora Schmid, Franziska Woltz, Jana Gast, and Angela Reitinger for their help with data acquisition. Without Nora in particular, this project would undoubtedly have been more troublesome. Special thanks go to Andreas Ruider and Alexander May for their technical support and to Teresa Schmidt-Peter and Benedikt Schröder for the wonderful hours spent together in the office.

My gratitude also extends to the young acousticians, Benedikt Bugl, Anna Rieger, and Viola Schneider at OTH Regensburg for their prompt assistance and motivating exchanges.

Finally, and most importantly, I would like to thank my husband, Paul, for his endless patience, support, and love, as well as my family and friends for being the fundamental backbone of my life.