Enhancing Perceived Social Presence of Embodied Conversational Agents: A Multimodal Approach to Natural Communication

Von der Fakultät für Informatik der RWTH Aachen University
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
genehmigte Dissertation

vorgelegt von

Jonathan Ehret, geb. Wendt, Master of Science

Berichter:
Prof. Dr. Torsten W. Kuhlen
Prof. Dr. Janina Fels

Tag der mündlichen Prüfung: 08.12.2025

Abstract

Embedding virtual anthropomorphic characters as embodied conversational agents (ECAs) in virtual reality (VR) applications offers benefits across various domains, including training environments and therapeutic settings, by simulating face-to-face interaction partners for human users. However, ensuring that ECAs are perceived as authentic and human-like remains a challenge, requiring the integration of multiple modalities to create convincing virtual humans.

This thesis explores key modalities that contribute to the believability of ECAs, which act inherently multimodal when delivering speech acts encompassing verbal as well as co-verbal behavior—such as gaze direction and gestures. We investigate the impact of voice and prosody, making suggestions on how to balance technical effort with their effects on perceived social presence. Additionally, we examine auralization, evaluating whether simulating natural sound directionality enhances conversational realism and determining the necessary level of technical fidelity. Furthermore, we delve into some communicative functions conveyed by co-verbal behavior. Specifically, we examine turn-taking, which governs speaker transitions in multi-party interactions, and back-channeling, which conveys agreement or understanding.

Throughout this work we put a particular focus on perceived social presence, the extent to which users feel they are interacting with a real person. To this end, we first review existing subjective and objective metrics for measuring social presence, identifying a gap in objective evaluation methods. Therefore, we assess objective metrics for social presence by leveraging the heard text recall (HTR) paradigm, developed in collaboration with psychology researchers. By systematically degrading ECA performance quality using aforementioned co-verbal components, we investigate HTR as a potential proxy for measuring cognitive load and social presence more rigorously.

To facilitate VR-based user studies, we introduce and assess the *StudyFramework*, a newly developed tool that streamlines factorial-design experiments and includes a system for rendering participant avatars to enhance immersion. Additionally, we explore methods for generating and capturing gestures using off-the-shelf VR hardware and analyze their influence on perceived social presence.

In summary, this research advances the understanding of ECA behavior in verbal communication, providing insights into key modalities that enhance natural and immersive interactions in VR.

Zusammenfassung

Die Einbettung virtueller anthropomorpher Charaktere als Embodied Conversational Agents (ECAs) in Anwendungen der virtuellen Realität (VR) bietet Vorteile in Bereichen wie Trainingsumgebungen und therapeutischen Settings, in denen sie persönliche Interaktionspartner simulieren. Eine zentrale Herausforderung besteht darin, ECAs als authentisch und menschenähnlich wahrnehmbar zu machen, wofür die Integration mehrerer Modalitäten erforderlich ist.

Diese Arbeit untersucht die wichtigsten Modalitäten, die zur Glaubwürdigkeit von ECAs beitragen. ECAs agieren multimodal und ihr Verhalten umfasst verbale sowie co-verbale Aspekte wie Blickrichtung und Gestik. Wir analysieren den Einfluss von Stimme und Prosodie und diskutieren, wie sich technischer Aufwand und wahrgenommene soziale Präsenz in Einklang bringen lassen. Zudem untersuchen wir die Auralisierung, indem wir bewerten, ob die Simulation natürlicher Klangausbreitung die Realitätsnähe von Gesprächen erhöht und welche technische Präzision erforderlich ist. Ein weiterer Schwerpunkt liegt auf kommunikativen Funktionen des co-verbalen Verhaltens, insbesondere Turn-Taking, das den Sprecherwechsel in Mehrparteien-Interaktionen regelt, und Back-Channeling, das Zustimmung oder Verständnis signalisiert.

Ein zentrales Thema dieser Arbeit ist die wahrgenommene soziale Präsenz – das Ausmaß, in dem Nutzer das Gefühl haben, mit einer realen Person zu interagieren. Wir analysieren bestehende subjektive und objektive Metriken und identifizieren eine Lücke bei objektiven Bewertungsmethoden. Zu diesem Zweck untersuchen wir das in Zusammenarbeit mit Psychologieforschern entwickelten HTR-Paradigma (Hearing Text Recall). Durch die gezielte Variation der ECA-Leistungsqualität anhand co-verbaler Komponenten evaluieren wir HTR als potenziellen Proxy für die Messung kognitiver Belastung und sozialer Präsenz.

Zur Unterstützung VR-basierter Nutzerstudien stellen wir das *StudyFramework* vor – ein Tool zur Vereinfachung experimenteller Designs, das zudem ein System zur Avatar-Darstellung zur Erhöhung der Immersion enthält. Des Weiteren untersuchen wir Methoden zur Generierung und Erfassung von Gesten mit handelsüblicher VR-Hardware und analysieren deren Einfluss auf die soziale Präsenz.

Zusammenfassend trägt diese Forschung zum Verständnis des ECA-Verhaltens in der verbalen Kommunikation bei und liefert Einblicke in Schlüsselmodalitäten, die natürliche und immersive Interaktionen in VR verbessern.

Acknowledgments

The process of conducting and writing of this thesis would not have been possible without the help and support of numerous people. While this list will most certainly not be complete, I nevertheless want to explicitly express my gratitude to a few people.

First, I want to thank my scientific advisor, Torsten W. Kuhlen, for his support, trust and guidance in pursuing this research endeavour. I also want to formulate my great gratitude to Andrea Bönsch. Thank you for guiding me through this entire process, always being there to discuss ideas, giving invaluable feedback, and foremost being a friend. I am also grateful for the scientific guidance I haven gotten from Benjamin Weyers and Tom Vierjahn when beginning this research journey. I also want to thank my other dear colleagues from the Virtual Reality and Immersive Visualization Group who I add the pleasure to work with throughout the years: Michael Anhuth, Martin Bellgardt, Aliki Charalabidou, Jan Delember, Ali Can Demiralp, Sevinc Eroglu, Sebastian Freitag, Sascha Gebhardt, Tim Gerrits, David Gilbert, Dirk Helmrich, Bernd Hentschel, Joachim Herber, Claudia Hänel, Kris Tabea Helwig, Jens Koenen, Marcel Krüger, Yuen Law, Fabian Lennartz, Aleksandra Lukic, Sebastian Menne, Jan Frieder Milke, Jan Müller, Christian Nowke, Simon Oehrl, Heiko Overath, Sebastian Pape, Till Petersen-Krauß, Sebastian Pick, Faysal Qurabi, Dominik Rausch, Timon Römer, Daniel Rupp, Patric Schmitz, Andrea Schnorr, Lukas Schröder, Benedikt Thelen, Tim Weißker, Viktor Wolf, and Daniel Zielasko. Thank you for all the fruitful discussion, upraising during breaks together, but most of all fun during these years. I also want to express my gratitude to the student workers, who supported me in making this work a success, namely: Marcel Czaplinski, Malte Kögel, Marcel Krüger, Denys Kuznietsov, Patrick Nossol, Azur Ponjavic, Marius Schmeling, and Ying Zhou.

Furthermore, I want to explicitly thank the project partners from other discipline. From the Institute for Hearing Technology and Acoustics these are especially: Lukas Aspöck, Carolin Breuer, Cosima Ermert, Philipp Schäfer, Jonas Stienen, and Janina Fels who I am especially grateful to for being the second examiner for this work. From the department for Work and Engineering Psychology I want to especially mention Chinthusa Mohanathasan, Isabel Schiller, and Sabine Schlittmeier.

Foremost, however, I want to thank my family for carrying me through this sometimes nerve-racking time. Thanks to my parents for always supporting me and believing in me. Thanks to my wonderful sons who always kept me grounded and showed me what matters most in live. But mostly I want to express my gratitude to my wife, who always had my back and supported me on so many levels. Thank you.

Eidesstattliche Erklärung
Declaration of Authorship

I, Jonathan Ehret

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Hiermit erkläre ich an Eides statt / I do solemnly swear that:

1. This work was done wholly or mainly while in candidature for the doctoral degree at this faculty and university;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others or myself, this is always clearly attributed;

4. Where I have quoted from the work of others or myself, the source is always given. This thesis is entirely my own work, with the exception of such quotations;

5. I have acknowledged all major sources of assistance;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. Parts of this work have been published before as listed below.

Aachen 18.12.2025

Jonathan Ehret

Pre-Released Publications

1. **J. Wendt**, B. Weyers, A. Bönsch, J. Stienen, T. Vierjahn, M. Vorländer, and T. W. Kuhlen. Does the Directivity of a Virtual Agent's Speech Influence the Perceived Social Presence? In: *IEEE VR Workshop on Virtual Humans and Crowds for Immersive Environments (VHCIE)*, 2018
2. **J. Wendt**, B. Weyers, J. Stienen, A. Bönsch, M. Vorländer, and T. W. Kuhlen. Influence of Directivity on the Perception of Embodied Conversational Agents' Speech. In: *19th ACM International Conference on Intelligent Virtual Agents*, 2019, doi:10.1145/3308532.3329434
3. **J. Ehret**, J. Stienen, C. Brozdowski, A. Bönsch, I. Mittelberg, M. Vorländer, and T. W. Kuhlen. Evaluating the Influence of Phoneme-Dependent Dynamic Speaker Directivity of Embodied Conversational Agents' Speech. In: *20th ACM International Conference on Intelligent Virtual Agents*, 2020, doi:10.1145/3383652.3423863
4. **J. Ehret**, A. Bönsch, L. Aspöck, C. T. Röhr, S. Baumann, M. Grice, J. Fels, and T. W. Kuhlen. Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents' Speech? In: *Transactions on Applied Perception*, 2021, doi:10.1145/3486580
5. **J. Ehret**, A. Bönsch, P. Nossol, C. A. Ermert, C. Mohanathasan, S. J. Schlittmeier, J. Fels, and T. W. Kuhlen. Who's next? Integrating Non-Verbal Turn-Taking Cues for Embodied Conversational Agents. In: *ACM International Conference on Intelligent Virtual Agents*, 2023, doi:10.1145/3570945.3607312
6. **J. Ehret**, A. Bönsch, J. Fels, S. J. Schlittmeier, and T. W. Kuhlen. StudyFramework: Comfortably Setting up and Conducting Factorial-Design Studies Using the Unreal Engine. In: *IEEE VR Workshop on Open Access Tools (OAT) and Libraries for Virtual Reality*, 2024, doi:10.1109/VRW62533.2024.00087
7. **J. Ehret**, A. Bönsch, I. S. Schiller, C. Breuer, L. Aspöck, J. Fels, S. J. Schlittmeier, and T. W. Kuhlen, Audiovisual Coherence: Is Embodiment of Background Noise Sources a Necessity? In: *IEEE VR Workshop on Virtual Humans and Crowds for Immersive Environments (VHCIE)*, 2024, doi:10.1109/VRW62533.2024.00017
8. **J. Ehret**, V. Dasbach, J.-N. Hartmann, J. Fels, T. W. Kuhlen, and A. Bönsch. Exploring Gaze Dynamics: Initial Findings on the Role of Listening Bystanders in Conversational Interactions. In: *TIEEE VR Workshop on Virtual Humans and Crowds for Immersive Environments (VHCIE)*, 2025, doi:10.1109/VRW66409.2025.00151
9. **J. Ehret**, J. Schüppen, C. Mohanathasan, C. A. Ermert, J. Fels, S. J. Schlittmeier, T. W. Kuhlen, and A. Bönsch. Objectifying Social Presence: Evaluating Degraded Speech Performance in ECAs Using the Heard Text Recall Paradigm. In: *Transactions on Visualization and Computer Graphics (TVCG)*, 2025, doi:10.1109/TVCG.2025.3636079

For my contribution to each of these papers, see Appendix C.1.

Contents

List of Acronyms

Acronym	Description
ANOVA	Analysis of Variance
AR	Augmented Reality
ART	Aligned Rank Transform
ASA	Artificial Social Agent
AU	Action Unit
CAVE	Cave Automatic Virtual Environment
DV	Dependent Variable
ECA	Embodied Conversational Agent
FACS	Facial Action Coding System
GLMM	Generalized Linear Mixed-Effect Model
GUI	Graphical User Interface
HMD	Head-Mounted Display
HRTF	Head-Related Transfer Function
HTR	Heard Text Recall
IK	Inverse Kinematics
IMU	Inertia Measurement Unit
IPA	International Phonetic Alphabet
IV	Independent Variable
IVE	Immersive Virtual Environment
LE	Listening Effort
LMM	Linear Mixed-Effect Model
M	Mean
MPS	Multimodal Presence Scale
RNMQ	Revised Network Minds Questionnaire
RO	Research Objective
SD	Standard Deviation
SPS	Social Presence Survey
SUS	System Usability Scale
TTS	Text-to-Speech
UE	Unreal Engine
VA	Virtual Agent
VAS	Visual Analog Scale
VR	Virtual Reality

Chapter 1
Introduction

In recent years virtual reality (VR) has seen a large gain in attention due to its versatile application in very different fields and decreasing costs. Embedding virtual humans into those VR applications can fulfill diverse needs. Virtual humans thereby can mean both virtual representations of the user (so called avatars) and also anthropomorphically embodied virtual agents (VAs), which operate autonomously through algorithms rather than direct user control. When these agents possess conversational capabilities, they are referred to as embodied conversational agents (ECAs) [Cassell, 2000]. These capabilities allow the agents to participate in conversations using natural language, thereby enabling meaningful interactions encompassing spoken communication with VR users. In the remainder of this work, we will primarily use the term ECA to emphasize the verbal communication of the agents.

ECAs have a variety of applications within immersive virtual environments (IVEs). For instance, they can enliven these environments as background characters [Trescak et al., 2014; Antunes and Correia, 2022], expose human users to social situations to treat social phobia [Wechsler et al., 2019], or facilitate public speaking training [Chollet et al., 2015; Kang et al., 2016]. Additionally, ECAs can serve as training partners [Chan et al., 2011; Anderson et al., 2013; Gratch et al., 2016], tutors [Cook et al., 2017; Grivokostopoulou et al., 2020; Mostajeran et al., 2020; Oker et al., 2020], therapists [Devault et al., 2014], or even product recommenders [Cassell et al., 1999; Qiu and Benbasat, 2010].

As VR emerges as a powerful platform for psychological research [Pan and de C. Hamilton, 2018], integrating authentic and life-like ECAs becomes crucial for enhancing user engagement and ensuring the reliability of studies (e.g., [Keller et al., 2024]). By embedding these realistic agents as social stimuli within research contexts, findings gain high ecological validity, making them applicable and meaningful in real-world scenarios beyond IVEs. To this end, it is essential to design effective ECAs that engage users in a meaningful way. First of all it has to be considered how these ECAs look, ranging from abstract (e.g., robotic or cartoonish) representation to photo-realistic renderings, but also comprises different clothing, gender, age etc. Key aspects elaborated in this work include verbal communication capabilities——encompassing both the production and rendering of speech——non-verbal behaviors, which include general non-verbal cues such as facial expressions and gestures, as well as co-verbal behaviors that occur in conjunction with speech. Additionally, fidelity in movement and expression is crucial for creating believable interactions. Each of these elements must be addressed individually while also considering their interactions with one another to create a cohesive user experience. In the following paragraphs, we will explore these aspects in greater detail, highlighting their importance and how they contribute to the overall effectiveness of ECAs.

Speech Production According to Johar [2016], verbal aspects of human communication that extend beyond the content of speech are referred to as paralanguage. This includes prosody——encompassing the rhythm of speech, stresses, pauses, pitch, and loudness——as well as vocal fillers (e.g., “um” or “ooh”). These vocal characteristics enable listeners to infer a speaker’s age, gender, geographic origin, or emotional state [Johar, 2016]. Seaborn et al. [2021] identified anthropomorphism as another important vocal aspect in ECAs. This refers to how human-like a voice is perceived. Typically, recorded human voices are compared with synthetic voices generated employing text-to-speech (TTS) synthesis. TTS offers advantages such as reduced technical effort compared to recording human speakers and increased flexibility as new speech acts can be generated in real-time if scenarios extend beyond pre-scripted conversations. However, while TTS systems provide these benefits, they often produce less natural-sounding speech and frequently generate inadequate prosody during synthesis [Kühne et al., 2020]. This limitation presents a significant research gap regarding its impact on user perception of ECAs. To this end, our research will specifically examine voice anthropomorphism as one influencing factor in this thesis to understand its contribution to bridging this gap and enhancing the overall effectiveness of ECAs. We will further call this factor **voice** in contrast to anthropomorphism, to better distinguish it from anthropomorphism consideration of other behavioral modalities of ECAs.

Speech Rendering In addition to these verbal aspects, another critical consideration in effective communication is auralization. Auralization [Vorländer, 2008] describes the technical way in which sound, or more particular here speech material, is presented to a human listener, for example, including artificial reverberation. Auralization techniques can significantly influence how users perceive spoken interactions within virtual environments [Dicke et al., 2010]. Therefore, this thesis will explore various **auralization** methods that account for directional propagation effects of speech, where the sound reaching a listener depends on the orientation of the speaking ECA to them.

Non-Verbal Behavior The effectiveness of ECAs is enhanced by leveraging their non-verbal capabilities, such as gaze and gesture behavior. These features can significantly increase user engagement [Oertel et al., 2020; Robb et al., 2023] and foster trust [Luo et al., 2023; Etienne et al., 2024]. Moreover, they contribute to greater enjoyment [Lee et al., 2006] and satisfaction [Biocca, 2001; Tu and McIsaac, 2002] during interactions while also improving the effectiveness of training simulations (e.g., [Strojny et al., 2020]) and making communication more efficient, for example, by streamlining the conversational flow (see, e.g., [López et al., 2008]). Wang and Ruiz [2021] identify the following forms of non-verbal behavior to be most prevalent (behaviors not particularly addressed in this thesis are grayed out):

- Facial Expression
- Gaze
- Gesture
- Posture
- Proxemics
- Behavioral Mimicry
- Complex Non-verbal Behavior

Thereby **facial expressions** can be used to transport emotions and articulation movement is also needed to realistically present speech with an ECA. **Gaze** and **gestures** not only accompany verbal communication but also provide additional information during pauses between speech acts. In this thesis, we will focus primarily on behaviors that occur in parallel to speaking. Therefore, we will refer to these as co-verbal behaviors rather than non-verbal behaviors. **Behavioral mimicry** describes the mirroring of a human interactant's behavior. While humans often do this unconsciously——such as following gaze direction or mirroring posture (e.g., crossing one's arms)——it can also be an effective technique to convey active listening [Maatman et al., 2005]. Proxemics, the distance an ECA keeps from a human user (see, e.g., [Bönsch, 2024]), and the general posture of the ECA are also important aspects which, however, will not be a focus of this work as this work is primarily concerned with face-to-face interactions in stationary contexts. Additionally, all these non-verbal behaviors together can convey complex information known as dialogue functions [Bente et al., 2008], which differ from discourse functions, which are closely related to speech production and understanding, and socio-emotional functions, which convey emotional states and interpersonal relationships. These dialogue functions encompass mechanisms such as listener **back-channels** that signal understanding or agreement [Bevacqua et al., 2010], enhancing the overall communicative experience. In this thesis, we will primarily focus on **turn-taking** cues as dialogue functions, which manage who is speaking next. It is essential for an ECA to effectively signal turn-taking so that users can easily follow the flow of conversation. By examining which specific cues are most impactful, we aim to gain a deeper understanding of how different non-verbal modalities interact within conversational dynamics.

Fidelity in Movement and Expression Furthermore, the fidelity of the ECAs' movements, including gazing, gestures, posture, facial expressions, locomotion, and subtle cues like breathing movement, have to be catered for when designing a believable ECA. For example, Ferstl et al. [2021] showed that low-fidelity movement can significantly shape how we perceive and like ECAs during an interaction.

Challenges in ECA Evaluation Evaluating the speech performance of ECAs, which encompasses both verbal and non-verbal behaviors, presents significant challenges [Sterna and Zibrek, 2021]. Many existing evaluations are often conducted in highly abstract settings, limiting their applicability to real-world scenarios [Sterna et al., 2023]. To address this issue, we decided to conduct our evaluations within more realistic contexts using the Unreal Engine (UE), as it allows for the creation of realistic IVEs that enhance the authenticity of ECAs as social stimuli. To achieve this, we developed a plugin, as described in Sec. 3, to equip researchers with the essential software tools for efficiently setting up and conducting VR user studies in UE, addressing the lack of a dedicated framework for UE.

In addition to this study framework, we created a modular ECA plugin specifically designed for behavior control, which incorporates functionality relevant to ECA behavior described in the following chapters (see App. C.3). This plugin enables researchers to customize and extend ECA functionalities within the UE environment, facilitating more nuanced interactions and enhancing the overall user experience. While our earlier research utilized the *SmartBody* toolkit [Thiebaux et al., 2008], which is capable of interpreting Behavior Markup Language (BML) [Kopp et al., 2006], we aimed to leverage the latest character animation capabilities offered by UE while maintaining flexibility for extending functionality. Consequently, we developed our own solution, and its components will be detailed throughout this thesis.

Social Presence When interacting with ECAs, one important concept is social presence [Short et al., 1976], which plays a crucial role in evaluating their speech performance and overall perception by human interactants. In essence this described the feeling of being in the presence of and interacting with a real person [Biocca et al., 2001; Oh et al., 2018]. Higher levels of social presence facilitate smoother interactions because humans can best utilize their communicative skills learned through countless prior interactions with other humans. However, evaluating social presence subjectively through questionnaires often results in significant variance due to the challenges participants face when reflecting on their experiences retrospectively. To address this issue, we investigate objective measures of social presence (see [Oh et al., 2018]), which will be detailed further in the following chapter. We will evaluate whether the measured cognitive spare capacity can serve as a proxy for social presence, based on the premise that unexpected ECA behavior consumes cognitive resources that would otherwise be available for performance in the listening task. One promising new approach is to employ the Heard Text Recall (HTR) paradigm [Schlittmeier et al., 2023], originally designed to assess memory performance in listening tasks, in more plausible scenarios. In this paradigm, participants listen to family stories, which can be presented as a dialogue between two speakers, and are then required to answer questions about the content they heard [Fintor et al., 2021; Mohanathasan et al., 2024] (see App. A.1 for an example text).

1.1. Contribution

Designing effective ECA behavior is complex and non-trivial, as it involves multiple interacting modalities that significantly impact user experience. Thereby, aiming for a high perceived social presence is desirable, as it can lead to more engaging and natural interactions. Thus, understanding which aspects of this multimodal behavior should be prioritized is crucial for enhancing user-ECA interactions. Consequently, the primary research objective **RO1** of this thesis is to evaluate different aspects of ECA behavior and their influence on perceived social presence. To this end, this thesis adopts a comprehensive approach by examining the breadth of ECA behavior rather than focusing exclusively on a single aspect, describing all modalities necessary for stationary interactions with ECAs. Therefore, locomotion and proxemics are specifically excluded from this analysis. We assess the verbal components of voice and prosody (Sec. 4.1) alongside advanced auralization techniques that utilize directional rendering of ECA speech (Sec. 4.2).

As a prerequisite to evaluating fully animated ECAs we present our solution to full-body motion capturing using standard VR hardware. As this simple tracking produces considerable tracking errors, we evaluate what effects various improvement techniques have on the perceived naturalness (which has a positive correlation with perceived social presence [Xenakis et al., 2023]) and preference of recorded co-verbal gestures (Sec. 5.3). We then use these recorded co-verbal gestures to address the research objective **RO2** of how non-verbal turn-taking cues contribute to the clearness of turn-taking and perceived social presence (Sec. 5.5). This is important because the ability to easily assess conversational flow enhances users' feelings of connection and engagement. Such clarity fosters a more open and natural dialogue with the ECA, promoting effective information exchange and increasing users' willingness to interact with the agent which is an essential factor for various applications, such as ECAs embedded as therapists or tutors.

As ECAs are not always the direct interactants of a human user, but can also be additional listeners or background characters, we address the research objective of how well these have to be represented by means of behavioral realism (**RO3**). We evaluate the effect of the presence and behavior of additional listeners on perceived clearness of turn-taking cues in Sec. 5.5.5 and the required audio-visual coherence of background characters in Sec. 6.2.

Building on the previous discussions regarding user engagement and interaction quality, the concept of social presence will be central to our considerations throughout this work. Thus, one significant contribution of this research is to provide a comprehensive overview of the concept of social presence and its subjective and objective measurements. However, measuring social presence presents various challenges due to the lack of reliable assessment methods. Consequently, we aim to explore alternative measurement techniques by focusing on cognitive load as an indirect indicator of social presence, which can provide insights into communication effectiveness and overall interaction quality. We anticipate a negative correlation between cognitive load and social presence, indicating that when users engage with ECAs without feeling overwhelmed or irritated——resulting in low cognitive load——they are likely to perceive a greater sense of connection and engagement, thereby experiencing higher levels of social presence. In light of this, one further research objective (**RO4**) explores whether the HTR paradigm can be employed to assess memory performance or cognitive spare capacity as a proxy for social presence, potentially offering a straightforward method for evaluating various ECA performance modifications (Sec. 6.1). Thereby, we further investigate how different ECA speech performance degraders impact perceived social presence, aiming to provide guidance for developers on prioritizing limited resources again addressing our primary research objective.

Lastly, to facilitate the user studies necessary for addressing the four presented research objectives and enhance the methodological rigor of these studies, an additional contribution of this thesis is the careful design, development, and evaluation of a comprehensive study framework tailored for factorial-design studies (see Sec. 3.2 and Sec. 3.3). Additionally, we contribute an avatar plugin (see Sec. 3.4) allowing users a synchronized representations of themselves via diverse sparse marker sets, enhancing self-awareness and realism during user-ECA interactions.

As this thesis prioritizes evaluating multimodal ECA behavior in terms of voice, speech auralization, and non-verbal behavior, we consciously do not alter visual fidelity aspects such as rendering quality or representation style (e.g., realistic, cartoonish, or robotic representations as detailed by [Zibrek et al., 2018] and [Zibrek and McDonnell, 2019]). Instead, we consistently employ state-of-the-art realistic visual representations to ensure a high-quality user experience. This approach aligns with our goal of integrating authentic and life-like ECAs as social stimuli, facilitating generalization to real-world scenarios beyond immersive settings, as required by psychological research. To further enhance our evaluations, we ensure that all modalities not in focus for a certain study are executed as effectively as possible based on state-of-the-art approaches. This strategy was employed to minimizes the risk of unintended side effects caused by inconsistencies in behavioral realism across different modalities [Kätsyri et al., 2015]. Additionally, while we do not delve into the cognitive aspects of ECAs, we maintain authentic interactions by adopting a Wizard-of-Oz (WoZ) approach [Devault et al., 2015], allowing an experimenter to remotely control the ECA's responses using their own social intelligence. This ensures that interactions remain grounded in realistic social cues and responses, enhancing the validity of our findings while avoiding complexities introduced by decision-making driven by artificial intelligence.

As this work reproduces several publications that were previously published and to value the contribution of all co-authors, the author will use the pronoun "we" instead of "I" throughout this work. This is also meant to pay tribute to the invaluable feedback and inspiring discussions the author had with them throughout the process of this work. For the publications reproduced in this thesis the individual contributions can be found in App. C.1.

1.2. Outline

This thesis is organized as follows: Following this introduction, we thoroughly introduce the concept of social presence in Chapter 2 and explain why we place it at the core of our evaluation while also giving an overview of potential subjective and objective metrics measuring it. After that, we will introduce the tools we developed to facilitate VR research, namely the *StudyFramework* as well as our implementation of body-avatars in Chapter 3. This is followed by overviews of verbal (Chapter 4) and co-verbal behavior (Chapter 5). During these chapters we will discuss our research on various specific aspects, as detailed before. Concluding these individual assessments we will take a more holistic view in Chapter 6, where we will look at the influence of different degraders on ECA performances and evaluate the HTR task as a proxy for measuring social presence, before reflection audio-visual coherence in more detail. This is followed up by a general discussion in Chapter 7 before concluding this work in Chapter 8.

Chapter 2

Evaluating Social Presence

Evaluating how authentic or natural an Embodied Conversational Agent (ECA) is perceived by human users is crucial for creating effective and engaging interactions that foster meaningful communication. For example, aspects such as behavior, speech, and appearance can be assessed to better understand how these contribute to enhancing the design of an ECA. Thereby eliciting higher social presence has multiple advantages. For example, it correlates with increased trust (see, e.g., [Luo et al., 2023; Etienne et al., 2024]), and enjoyment ([Lee et al., 2006]) and satisfaction ([Biocca, 2001; Tu and McIsaac, 2002]) of the interaction. Furthermore, it can also change the outcome of, for example, a training simulation. Strojny et al. [2020] found virtual bystanders to only have an influence on the performance of fire fighters being trained, if they rated the social presence of these bystanders as high. We further believe that social presence is superior to aforementioned concepts influencing social presence, like naturalness and realism, since it goes beyond mere biologically appropriate presentation and looks also into cognitive processes happening in the observers and being projected onto the ECAs.

Biocca and Harms [2002] proposed the notion of *social presence* for ECAs, originally introduced by Short et al. [1976]. It was meant as a metric "to compare the relative effectiveness of various mediated technologies, interface features, or agents" [Biocca et al., 2001]. Social presence describes the feeling of being in the presence of another human being – which is also often described as *co-presence* – and having "access, and connection to the intentional, cognitive, or affective states of the other" [Biocca et al., 2001]. Biocca and Harms [2002] further describe that the aspects going beyond co-presence can be further subdivided into the subjective level (experiencing an accessibility of the other) and an inter-subjective level (mutual social presence, so feeling also experienced as socially present by the other). While in some literature co-presence and social presence are used interchangeably (e.g., [Oh et al., 2018; Sterna and Zibrek, 2021]), we follow the definition in [Biocca and Harms, 2002] which defines co-presence only as the "being together" aspect of social presence. Following Bente et al. [2008], "higher levels of social presence include a sense of behavioral engagement, which is expected to lead to actions that are perceived as linked, reactive, and interdependent.” Sterna and Zibrek [2021] state that social presence can be seen as a needed baseline for research on social behavior in VR, since social reactions to ECAs are tied to an enhanced perceived social presence of those, not only experiencing them as mere computer programs. An in-depth historical analysis of the concept can be found in [Kreijns et al., 2022]. The concept of social presence is inspired by the concept of *presence*, which describes the sensation of feeling to “be there” as part of a virtual environment rather than observing it from the outside and is, for example, influenced by the immersion provided by the used hardware (see [Heldal et al., 2005; Grassini and Laumann, 2020]). In a differently pivoted approach Slater [2009] uses the dimensions *Place Illusion (PI)* and *Plausibility Illusion (Psi)* to describe presence, where the first describes the illusion that one feels to be at that virtual place while the later is characterized by the illusion that events occurring within the virtual space are actually plausible for the depicted scenario (so a dragon can be well plausible in a virtual fantasy world albeit not plausible in a simulation depicting reality). Skarbez et al. [2017] accompany these two illusion by the co-presence illusion, which again described the feeling of togetherness. Following them, the co-presence illusion together with Psi creates a *Social Presence Illusion*. These concepts are still in open discussion, see for example [Latoschik and Wienrich, 2022], where plausibility and congruence are used to replace the illusions described by Slater [2009] and further refined in [Slater et al., 2022]. Kreijns et al. [2022] addressed social presence from a distance education perception and found that in the literature there exists a jingle-fallacy for social presence, which means that the term social presence is sometimes used with very different meanings and therefore special care has to be taken when using this term. Furthermore, they give their own interpretation of social presence, more focused on social interaction during remote teaching and state that social group and sociability should be regarded separately, which we would also see under the umbrella of social presence as conceptualized in [Biocca and Harms, 2002]. A further framework for rating social presence in the context of joint music making is given by Kerrebroeck et al. [2021]. In this work, we will, however, use social presence as introduced by [Biocca et al., 2001], since it captures more aspects than mere co-presence, that help to understand how an ECA is perceived by a human user.

The remainder of this chapter is structured as follows. First we will give a brief overview about other concepts that are strongly related and intertwined with social presence in Sec. 2.1. After that we will introduce subjective (Sec. 2.2) and objective (Sec. 2.3) measurements before presenting influencing factors of social presence in Sec. 2.4. We will end this chapter by discussing shortcomings and potential research directions of the aforementioned in Sec. 2.5.

2.1. Related Concepts

There is a multitude of concepts that either describe similar aspects as social presence or can be seen as predictors of social presence. A large group of those centers around the idea of *naturalness*, which, based on Xenakis et al. [2023], is fundamental to the occurrence of social presence. Naturalness itself is often expressed through various related terms such as *realism*, *human-likeness*, or *fidelity*. All of these describe the idea of ECAs being presented in a specific way, i.e., emulating interactions and human behavioral experiences from the real world as closely as possible [Alexander et al., 2005]. An overview of how to rate realism in IVEs in general can be found in [Goncalves et al., 2021]. However, the application context always needs to be taken into account as argued by Xenakis et al. [2023]: the perceived naturalness of, e.g., given co-verbal cues, depends strongly on the specific scenario and is not sufficient to elicit high social presence but overall sensible behavior is required as well.

This can potentially be described by the terms *authenticity* and *believability*, which seem to go further the pure naturalness by also considering a simulated consciousness of those ECAs [Kope et al., 2013], which can only be experienced in more complex interactions. Another related concept put forward by Burgoon et al. [2016] is the violation of expectation, which can be especially detrimental for very realistic ECAs which cannot live up to the formed expectations. However, violations of expectations can also be positive, if low expectations are positively violated by behavior going beyond the expected, which can be beneficial for the outcome of an interaction [Burgoon et al., 2016].

Recently, this range of concepts was further broadened by the introduction of the *Interaction Fidelity Model (IntFi)* [Bonfert et al., 2024], which tries to categorize fidelity over all levels of a VR application with a focus on the interaction. Of the eight introduced fidelity aspects, *simulation fidelity* and *rendering fidelity* come closest to the previously discussed concepts and could be applied to ECAs. A narrative survey about the realism of ECAs and the interplay with other concepts—with a special focus on animation—can be found in [Rekik et al., 2024], which also relates these aspects to the uncanny valley theory.

The *uncanny valley* goes back to Mori [1970] It describes the effect that when the human-likeness of an ECA is increased, the familiarity/affinity increases, before sharply dropping for more human-like but not quite realistic characters, before then again climbing farther than before after this uncanny valley. This effect was first reported by Mori [1970] for wooden puppets and only later on applied to anthropomorphic virtual characters, where it was believed that ECAs looking somewhat eerie and corpse-like, especially when moving in not quite human ways, cause this effect and should therefore be carefully avoided. However, there exists an open debate whether the uncanny valley effect in this simple form does actually depict the effect in its entirety. More recently, while still commonly using the same name, the concept has shifted to a perceptual mismatch hypothesis [de Borst and de Gelder, 2015]. The core idea for this hypothesis is that if single modalities are more human-like then others this causes repulsion and eeriness in human observers. They state, for example that “not the most realistic looking virtual characters evoke an eerie feeling, but rather those on the border between non-human and human categories, especially if they are combined with human-like motion” [de Borst and de Gelder, 2015]. So this effect should also be carefully considered when looking into social presence of ECAs.

2.2. Subjective Measurements

There are many questionnaires available to rate social presence or related concepts. One of the earliest was the *Semantic Differences Survey (SDS)* [Short et al., 1976] (see App. B.1) which used semantic differentials to rate ECAs between two bipolar attributes like “cold” and “warm”. After that many questionnaires were developed using Likert scales and statements to be rated, like “I perceive that I am in the presence of another person in the room with me.” from the *Social Presence Survey (SPS)* [Bailenson et al., 2001] (see App. B.2). While most of them were specifically designed for virtual agents, like ECAs, some were also made to rate robots, for example, the *Godspeed Questionnaire* [Bartneck et al., 2009] (see App. B.7). Fitrianie et al. [2019] conducted a survey on available questionnaires for virtual agents and robots alike and then conceived a questionnaire, namely the *Artificial Social Agents (ASA) Questionnaire* [Fitrianie et al., 2022] (see App. B.11). For this very questionnaire we were also involved in formally deriving a German translation (see [Albers et al., 2024]) of which parts were used in some of the studies described in this work.

In App. B, we reproduced the questionnaires most commonly used in our research domain. We will therefore not present all of them here. Instead, we have curated a list of recommended questionnaire constructs derived from various sources for use in user studies involving ECAs. This list aims to provide readers with an overview of existing constructs while acknowledging that it only reflects our perspective on the current state of the art and may require re-evaluation as new questionnaires are published.

The primary goal of this curated list is to measure whether interactions with one or multiple humanoid virtual agents in an IVE felt realistic and whether these agents were perceived as socially present. Furthermore, the list should be as concise as possible, while compromising all relevant aspects. We intentionally do not focus on improving perceived personality or emotions of the agents. Instead, our emphasis is on (non-)verbal behavior improvements, which are central to this work. Explicit manipulations of emotions fall outside the scope of this thesis. Additionally, we do not evaluate agent appearance since we consistently use high-quality virtual human models and do not consider appearance an independent variable relevant to our research focus. To this end, we, for instance, chose not to include the *Human-Like Appearance* construct from the *ASA Questionnaire* (see App. B.11).

In our early user studies, we utilized the SPS [Bailenson et al., 2001] (see App. B.2), particularly in studies presented in Sec. 4.2. However, feedback indicated that participants found it challenging to respond effectively. For example, one participant complained that rating a virtual agent to be “only a computerized image” felt odd. We hypothesize that this was one potential reason why this questionnaire failed in our studies to reveal significant effects, as it measured not exactly the construct necessary there (which probably would have been more concerned with naturalness). To this end, we excluded the SPS from our curated list.

We also did not include, e.g., the *Social Presence* dimension of the questionnaire by Nowak and Biocca [2003] (see App. B.4), since it was too closely tailored for a face-to-face interaction between two interactants and not applicable to a broader set of scenarios. Furthermore, we did not consider constructs from the original *Networked Minds Questionnaire* [Biocca, 2001] (see App. B.3) since it was later on revised by the same authors into the *Revised Network Minds Questionnaire (RNMQ)* [Harms and Biocca, 2004] (see App. B.3.1) which is now more actively used and items of which were considered for this combined questionnaire (see below).

For the curated list we extracted the *Social Presence (SP)* construct (see Tab. 2.1) from the *Multimodal Presence Scale for Virtual Reality Environments (MPS)* [Makransky et al., 2017] (see App. B.8). It closely resembles the aforementioned SPS [Bailenson et al., 2001]. However, feedback from our user studies suggests that participants found it more straightforward to respond to. For example, the second question was reformulated from "I feel that the person is watching me and is aware of my presence" (SPS) to "I felt that the people in the virtual environment were aware of my presence" (MPS). The more general phrasing in the MPS is preferable because it allows for a broader interpretation of awareness, whereas the SPS question specifically ties awareness to gazing behavior. This distinction is important, as ECAs can convey a sense of awareness even without continuously looking at the participant.

Social Presence	
SP1	I felt like I was in the presence of [another person] in the virtual environment.
SP2	I felt that the [person] in the virtual environment was aware of my presence.
SP3	The [person] in the virtual environment appeared to be sentient (conscious and alive) to me.
SP4	During the simulation there were times where the computer interface seemed to disappear, and I felt like I was working directly with [another person].
SP5	I had a sense that I was interacting with [another person] in the virtual environment, rather than a computer simulation.

Table 2.1.: The social presence dimension of the *Multimodal Presence Scale for Virtual Reality Environments (MPS)* [Makransky et al., 2017] rated on a 5-point Likert scale between "completely disagree" (1) , "disagree" (2), "neither disagree nor agree" (3), "agree" (4), and "strongly agree" (5). The full questionnaire can be found in App. B.8.

From the *Temple Presence Inventory (TPI)* [Lombard et al., 2009] (see App. B.6) we considered the construct *Social presence - Active interpersonal* to measure spontaneous user reactions towards the ECA as an indicator of the ECA's plausibility. This kind of reactions can be an indicator that users pereived the ECA as a social entity and therefore response adhering to learned social norms, potentially without consciously deciding to do (cp. [Sheridan, 1992; Hoffmann et al., 2009]). To this end, we also incorporated one item from the emphSocial Presence - Actor within Medium (Parasocial Interaction) dimension as *SR4* (see Tab. 2.2), as it aligns well with the spontaneous social reactions. The other items from this dimension were, however, excluded as their content is already well covered by the MPS construct above. Although the items of the *Social Realism* dimension seemed helpful for measuring naturalness on a first glance, we did not include them. Our concern was that participants might interpret these items too broadly and rate aspects of the virtual environment beyond just the ECAs. For instance, statements introduced with “The events I saw/heard ...” could lead to varied interpretations among different participants, introducing inconsistency and ambiguity in the data collected. Instead, we believe that the item will be more effectively assessed content-wise through the *ASA Questionnaire*’s *Human-like Behavior* dimension (see below).

Spontaneous Reactions	
SR1	How often did you make a sound out loud (e.g. laugh or speak) in response to someone you saw/heard in the media environment?
SR2	How often did you smile in response to someone you saw/heard in the media environment?
SR3	How often did you want to or did you speak to a [person] you saw/heard in the media environment?
SR4	How often did you want to or did make eye-contact with someone you saw/heard?

Table 2.2.: The “Social presence - Active interpersonal) dimension of the *Temple Presence Inventory (TPI)* [Lombard et al., 2009] (SR1 - SR3) enhanced by one question from the *Social Presence - Actor within Medium (Parasocial Interaction)* dimension (SR4). All items are rated on a 7-point Likert scale between “Never” (1) and “Always” (7). The full questionnaire can be found in App. B.6.

We further extracted the dimensions *1.2 Human-Like Behavior* and *13 Agent’s Coherence* from the *ASA Questionnaire* [Fitrianie et al., 2022] (see App. B.11) as they nicely cover naturalness aspects of the evaluated agents (see Tab. 2.3). Both constructs focus on key aspects that define how convincingly an ECA can simulate human behavior by directly evaluating the degree to which the ECA’s behavior is seen as human-like and assessing the consistency and rationality of the ECA’s actions.

While coherence and rationality are essential for assessing the ECA’s naturalness, understanding users’ perceptions of its intentionality reveals how purposeful they find its behavior. This perception is crucial as it enhances user engagement and trust, making interactions feel more natural. When users see the ECA as acting with clear intentions, they can better predict its actions, leading to smoother interactions and increased reliance on its decisions. To this end, we also included the ASA Questionnaire’s construct 14 on *Agent’s Intentionality* (see Tab. 2.3). Additionally, we incorporated an item from *5. Agent’s Sociability*——specifically, “[The person] interacts socially with [me]"—to evaluate how effectively the ECA engages in social interactions, which is integral to enhancing user experience. In this thesis we focus on stationary interactions. For dynamic scenarios where social formations and movement is involved (see, e.g., [Bönsch, 2024]), however also the first two items AS1 and AS2 can be of relevance.

Human-Like Behavior		
HLB1	A human would behave like [the person]	
HLB2	[The person]'s manners is consistent with that of people	
HLB3	[The person] behavior makes me think of human behavior	
HLB4	[The person] behaves like a real person	
HLB5	[The person] has a human-like manner	
Agent's Coherence		
AC1	[The person]'s behavior does not make sense	inv
AC2	[The person]'s behavior is irrational	inv
AC3	[The person] is inconsistent	inv
AC4	[The person] appears confused	inv
Agent's Sociability		
AS1	[The person] can easily mix socially	
AS2	It is easy to mingle with [the person]	
AS3	[The person] interacts socially with [me]	
Agent's Intentionality		
AI1	[The person] acts intentionally	
AI2	[The person] knows what it is doing	
AI3	[The person] has no clue of what it is doing	inv
AI4	[The person] can make its own decision	

Table 2.3.: The "Human-like Behavior", "Agent's Coherence", "Agent's Sociability", and "Agent's Intentionality" dimensions from the *ASA Questionnaire* [Fitrianie et al., 2022] rated on a 7-point Likert scale with labels: "disagree" (-3), "neither agree nor disagree" (0), and "agree" (3). "inv" means that the item's rating is inverted before computing the mean of the individual items. The full questionnaire can be found in App. B.11.

Furthermore, we suggest to utilise the *Perceived Behavioral Interdependence* construct from the *Revised Networked Minds Questionnaire* [Harms and Biocca, 2004] (see App. B.3.1). This construct (see Tab. 2.4) introduces the reciprocal dimension of social presence, where perceived social presence also depends on whether the user feels perceived as socially present by the ECA. We prefer this one over the *19 User-Agent Interplay* dimension of the *ASA Questionnaire* since we found those items more intuitive to answer and less focused on emotional responses like, for example, the item "[My / The user's] emotions influence the mood of the interaction" which is part of the ASA Questionnaire.

Perceived Behavioral Interdependence	
PBI1	My behavior was often in direct response to [the other person]'s behavior.
PBI2	The behavior of [the other person] was often in direct response to my behavior.
PBI3	I reciprocated [the other person]'s actions.
PBI4	[The other person] reciprocated my actions.
PBI5	[The other person]'s behavior was closely tied to my behavior.
PBI6	My behavior was closely tied to [the other person]'s behavior.

Table 2.4.: The "Perceived Behavioral Interdependence" dimension from the *Revised Networked Minds Questionnaire* [Harms and Biocca, 2004]. As in the original publications not concrete rating scale is given, we propose for consistency to rate it on a 7-point Likert scale with labels: "disagree" (-3), "neither agree nor disagree" (0), and "agree" (3). The full questionnaire can be found in App. B.3.1

Lastly, we consider the semantic differentials described in the first dimension, namely *Anthropomorphism*, of the *Godspeed* Questionnaire [Bartneck et al., 2009] (see App. B.7) relevant, which are rated on a 5-point bipolar scale. Participants have to rate the ECAs on the scale between these two opposing terms (see Tab. 2.5), and thereby this instrument gives a very explicit and conscious way of rating the ECAs compared to the rated statements of the constructs before, albeit asking for similar concepts.

From the second dimension (*Animacy*) we considered the pairs "Inert - Interactive" and "Apathetic - Responsive" interesting, as they have a stronger focus on the interactive capacities of the ECAs as the aforementioned semantic differentials. However, we found them already well-enough covered by the *Perceived Behavioral Interdependence* construct from the *Revised Networked Minds Questionnaire* [Harms and Biocca, 2004], and tried to reduce the cherry-picking of construct items to a minimum.

Anthropomorphism		
Please rate your impression of [the other person] on these scales:		
ANT1	Fake	Natural
ANT2	Machinelike	Humanlike
ANT3	Unconscious	Conscious
ANT4	Artificial	Lifelike
ANT5	Moving rigidly	Moving elegantly

Table 2.5.: The “Anthropomorphism” dimension from the *Godspeed* Questionnaire [Bartneck et al., 2009]. Items are rated on a 5-point bipolar scale between the given two anchor labels. The full questionnaire can be found in App. B.7

The questionnaire items presented above were partially modified for consistency to ensure they can be used together in a single questionnaire. Specifically, we standardized the term “person” throughout, replacing various terms such as “agent”, “interaction partner”, or “virtual character” found in the original questionnaires. However, terms in brackets (like “[person]”) are intended to be replaced by the experimenter with

We found our curated list of constructs to well capture the general perception in an interaction with an ECA, without any particular focus. To this end, depending on the research, the questionnaire needs to be extended. For example, if verbal behavior is tested, both intelligibility of the ECA’s speech but also perceived message understanding of the ECA might be of interest. For the latter, for example, the *RNMQ - Perceived Message Understanding* can be used as addition.

Interested readers are referred to [Oh et al., 2018] or [Fitrianie et al., 2019] and App. B for further information on available questionnaires. However, there exists also a general criticism towards the potential of questionnaires to measure social presence, since they always require a subjective rating to questions after the exposure, which might be hard to do and potentially subject to biases. Slater et al. [2010] for example proposes the usage of the *Configuration Transition* method where participants are allowed to change aspects of the scene until they deem it good enough, potentially incentivized to change as little as possible. Beyond that Slater et al. [2022] also proposes that participants write short essays about their experience during a study condition and then *Sentiment Analysis* is used to derive how positively or negatively the virtual simulation was perceived. Along these lines Wolfert et al. [2024] found, while comparing a newly developed questionnaire and direct comparison of co-verbal gestures, that the direct comparison where participants had to repeatedly identify the better stimulus from a pair of stimuli, yielded more powerful results.

While there exist many questionnaire-based instruments to measure social presence (with only the most prominent being reproduced above and in App. B), these self-report based measures often do not yield significant results for subtle changes in the ECAs, e.g., Harms and Biocca [2004] were not able to significantly discriminate video and text-based interactions using the above mentioned scale. Therefore, we will next look at objective ways to quantify social presence in the following section.

2.3. Objective Measurements

One possibility for objective measurements of social presence could be proxemics, the distance a user keeps from an ECA [Bönsch et al., 2018a], which was found to enlarge with increasing social presence [Bailenson et al., 2001, 2004]. Furthermore, the Ash conformity test [Kyrlitsias and Michael-Grigoriou, 2018] can be used, which examines whether a human user adapts to the exhibited behavior of an ECA. Similarly, socially conditioned behavior can be provoked [Sheridan, 1992], like grasping for an offered object or responding to a sneezing or waving ECA. One specific aspect of this is mimicry/alignment, copying non-verbal behavior of ones interactant, which was found to be an indicator of higher social presence [Hasler et al., 2017], or the adaptation of verbal behavior to that of the ECA (see, e.g., [Pütten et al., 2010; Ochs et al., 2017, 2022]). Interestingly, Bergmann et al. [2015] found that lexical alignment, i.e., whether similar words are used, actually decreases with higher social presence. However, this is also the case when comparing interactions with a computer and with another human, where lexical alignment is much greater for the computer, so these results illustrate well that social presence increases when the interactant is experienced to be more human-like. Hayes et al. [2022] developed the *Social Presence Behavioral Coding System* to formalize analysis of user behavior, looking for mimicry but among other things also for emotional engagement, socially engaging, or self-disclosing during an interaction with an ECA.

With the advent of head-mounted displays (HMDs) eye tracking also became more accessible, since trackers are already integrated in some consumer headset. This allowed the analysis of gaze data, which showed that humans tend to look more in the eyes of ECAs than they do when they expect to virtually interact with a real human and adhere to social norms (see, e.g., [Rehm and André, 2005; Cañigueral and Hamilton, 2019]). Along these lines, Holleman et al. [2020] found participants to gaze more towards the eye of another person if the presentation was given asynchronously and their gaze wasn't perceived by the presenting person, which was further backed by Cañigueral et al. [2021]. An in-depth description of how to analyze gaze tracking was published by Lamb et al. [2022].

Beyond these behavioral measures, also physiological measures are put forward. For example, Sterna et al. [2023] tried to employ heart rate and skin conductance to measure social presence. In addition to this, Kock [2005] links media naturalness to cognitive effort, seeing a higher cognitive effort if the degree of realism is too low. Further, Bailenson et al. [2005] found memory performance to be influenced by social presence, albeit not conclusively.

In the remainder of this thesis we will elaborate on the usage of objective measures for social presence by utilizing them in various user studies side-by-side with traditional questionnaires. Thereby we plan to evaluate their applicability in potentially replacing such questionnaires. Lastly in Sec. 6.1, we will evaluate a specific task developed by psychologists for measuring memory performance and cognitive spare capacity as potential proxy for objectively measuring social presence.

2.4. Influencing Factors of Social Presence

While there are many ways to measure social presence as discussed previously, numerous factors can impact the perceived social presence of an ECA. There exists a large body of work evaluating different aspects of ECAs and their influence on social presence. For example, Nowak and Biocca [2003] found that a higher social presence was perceived if anthropomorphic **representation** were shown and Zibrek et al. [2017] showed that more realistically rendered ECAs can improve the perceived social presence. This was also confirmed in [Zibrek and McDonnell, 2019], however, Arboleda et al. [2024] were not able to reproduce this effect between "realistic" and "cartoonish" representations. Further, Wang et al. [2019b] found that participants preferred miniature ECAs over full-size ones in augmented reality (AR). These described effects could origin from the interdependence of behavioral and visual realism on social presence described by Bailenson et al. [2005] and Sterna et al. [2023], where social presence decreases if behavioral and visual realism do not match, which was the case for most studies described above if only the visual representation was manipulated.

Oh et al. [2018] state that in general **behavioral realism** is a powerful predictor for perceived social presence. For example, Kim et al. [2018] found a significant influence of appropriate movement on social presence. Kimmel et al. [2023] found facial expression and facial animation in general to elicit higher degrees of social presence, while Luo et al. [2023] were not able to show this influence on social presence. Another important aspect of behavior are so-called backchannel movements, which is one key aspect discussed in this thesis and will be elaborated in Sec. 5.4. Poppe et al. [2011] and Pütten et al. [2010] found that they have a positive influence on the perceived social presence. Furthermore, Ferstl et al. [2021] evaluated that degraded, robot-like gestures are perceived as less human-like, unfortunately not directly measuring social presence. An in-depth literature review of the influence of non-verbal behavior on (social) presence can be found in [Xenakis et al., 2023], also putting forward the possibility of having "super-natural" behavior, like a single ECA making eye-contact with multiple users at the same time.

In general, Sterna et al. [2023] found **interactability** to have a strong positive influence on social presence, albeit only when the behavioral realism was in general high. This is supported by research conducted by Skalski and Tamborini [2007] and Garau et al. [2005], where the latter explicitly stated that participants felt "ghost-like" or even voyeuristic if the ECAs did not appropriately react to their presence, for example, by acknowledging them by means of gazing at them.

Another important aspect of ECAs is how they sound. This is also an integral part of this work, addressed in Sec. 4.2. Higgins et al. [2022] showed that synthetic **voices** used for ECAs, rather than recording a human speaker, has a detrimental effect on social presence and Miniota et al. [2023] found that one deficiency of current speech synthesizers is that the speech does not sound truly spontaneous which negatively impacts naturalness. Beyond that, Lam et al. [2023] found that human observers have a clear expectations of which voices fit which visual ECA representation, which should be carefully considered to increase believability. Similarly, Lee and Nass [2003] found higher social presence rating for voices matching the portrayed personality of the ECA, in their case comparing introvert and extrovert portrayals.

Regarding **personality**, Lee et al. [2006] found that opposite personalities to that of the participants provoked higher social presence. In this regards, Cheng and Wang [2024] researched the influence that an ECA's clothing can have on its perceived personality showing influences of, for example, the color or the style (e.g., professional compared with casual etc.) of outfits. Allmendinger [2010] even constitute that the personality, in her case shyness, of the person interacting with an ECA, has an influence on whether higher or lower social presence is perceived as more pleasant by the person. Further information about the influence of personality and trustworthiness of ECAs can be found in [Etienne et al., 2024], which is however beyond the scope of this thesis, focusing on social presence.

Furthermore, the used **display system**, for example whether a desktop monitor or an HMD is used, can have an influence on the perception of social presence (see, e.g., [Guimarães et al., 2020]). As the used medium for mediated-communication was the original purpose of the conception of social presence, there exists a large body of work, mainly finding higher degrees of social presence for more immersive technology like HMDs (see, e.g., [Heldal et al., 2005]). However, other researchers, for example, Bente et al. [2008] did not find a strong influence of the medium, only showing increased social presence of the tested degrees of immersion against pure text-based communication. Another potentially beneficial capability of such a system is haptics. Sallnäs [2010] found that adding haptics when passing objects in mediated collaboration did significantly increase social presence.

Other influencing factors can be the perceived **agency** (see, e.g., [Kyrlitsias and Michael-Grigoriou, 2022]), so whether participants believed to interact with a real human embedded as avatar or a computer-controlled ECA. Oh et al. [2018] report, that there is some evidence that believing to interact with a real human increases the social presence, independent of whether the interaction actually is done with a human or a computer-controlled ECA. Beyond looking at individual aspects of ECAs, Qiu and Benbasat [Qiu and Benbasat, 2010] report that in general higher social presence is reported for ECAs that match the ethnicity of participants.

For further reading a narrative review of how different ECA design decision influence emotional experience, psychological discomfort, presence, engagement, and social presence in interactions with ECAs in VR can be found in [Mulvaney et al., 2024]. Other interesting reviews can be found in [Yassien et al., 2020] and [Oh et al., 2018], while Norouzi et al. [2020] looked specifically at interactions with ECAs in AR.

2.5. Summary

The presented literature shows that there are many factors influencing social presence. Thereby eliciting higher social presence has multiple advantages, as mentioned before, like increased trust (see, e.g., [Luo et al., 2023; Etienne et al., 2024] but also [Alimardani et al., 2024] who were unable to reproduce this), and enjoyment ([Lee et al., 2006]) and satisfaction ([Biocca, 2001; Tu and McIsaac, 2002]) of the interaction. For this work it is important to understand that there are many influencing factors that potentially even influence each other. We believe that social presence is a more comprehensive concept than the previously mentioned factors influencing it, such as naturalness and realism. Unlike these factors, which focus primarily on biologically appropriate presentation, social presence also encompasses the cognitive processes occurring in observers and how these processes are projected onto the ECAs.

We presented the most commonly used questionnaires for quantifying social presence above (and in App. B). However, since they often fall short to measure social presence reliably, there are constantly new questionnaires developed and published. Furthermore, often objective measures are also used. For example, Sterna and Zibrek [2021] constitute that there is a shortage of validated measures and a clear and generally-accepted definition of social presence still does not exist. Nevertheless, Sterna and Zibrek [2021] further propose the usage of indirect, objective metrics alongside subjective questionnaires.

In this work we will therefore look into other objective metrics in Sec. 6.1. Before that, however, will will shed more light onto the different modalities used for ECAs and their influence on social presence, presenting original research alongside general conceptualizations. However, we will now first introduce some tool which were implemented to facilitate this research process.

Chapter 3

Essential Tooling: A Factorial-Design Study Framework with Body Avatar Integration

The contents of this section are based on and taken in part from work previously published in [Ehret et al., 2024a].

To understand the influence that different changes for example on ECA behavior have on the perceived social presence as outlined in the previous chapter, formal evaluation is necessary. Thereby it is required to generate a deep understanding of human behavior in interactions with these ECAs, to be able to incrementally enhance the effectiveness of those. Controlled user studies are a key method to assessing new or refined techniques, to be able to further guide development into the most promising directions and confirm that meaningful enhancement over existing methods were taken. User studies can reveal how users interact with ECAs and identify factors in the multimodal ECAs' design that enhance realistic interactions and perceived social presence. Furthermore, VR is an emerging tool in, for example, psychological research where it can help to conduct controlled experiments in settings closer to real life than classical lab experiments. However, setting up those studies is time-consuming and holds a lot of potential caveats with respect to experimental design and data management, such as counterbalancing randomized condition orders and thoroughly storing all relevant data. If setup errors go unnoticed until the study conductance or data evaluation, there is a substantial risk of data corruption, making parts or all of the data unusable.

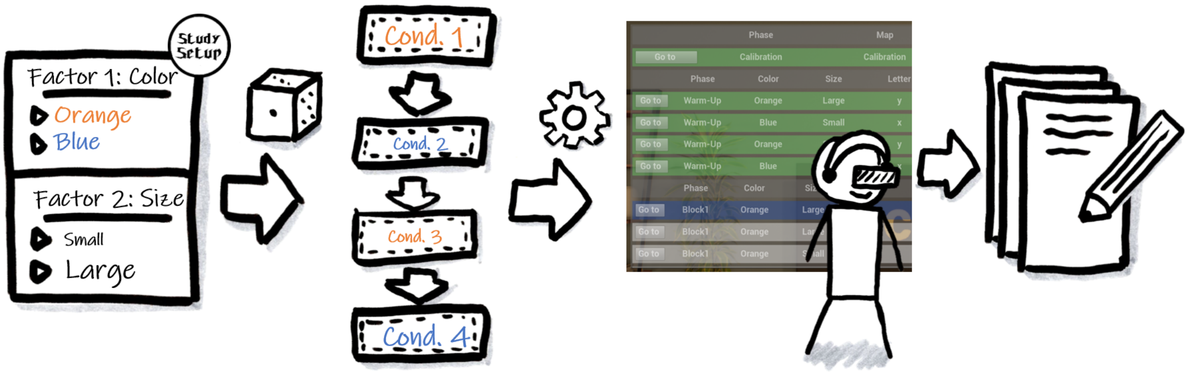


Figure 3.1.: The pipeline of the *StudyFramework*: The factorial-design setup is randomized into an ordered list of conditions. This list can then be executed in VR, providing an helpful control interface to the experimenter. Finally, all gathered data is carefully logged.

Modern game engines like the *Unreal Engine*[1] or *Unity*[2] are readily available and are increasingly often used as they provide a lot of helpful tools to simplify setting up IVEs and also to implement complex interactions. However, their complex internal architectures can introduce numerous challenges when trying to implement a robust study that, for example, should include different virtual scenes, especially for engine novices. On top of that, the required code for setting up such studies is often very similar. Consequently, identical code has to be reproduced over and over again, potentially introducing unwanted side effects.

To facilitate the development and execution of factorial-design studies using the Unreal Engine, we developed the *StudyFramework*. A factor in this context can be, for example, the degree of usage of co-verbal gestures being varied between the levels: no gestures, non-fitting gestures, fitting gestures. This factor can then potentially be paired with another factor like the visual representation varying between the levels realistic and cartoon-like.

Next, we explore related approaches in Sec. 3.1. We then define our contributions by outlining our goals and detail the core components of our framework, employing an illustrative 2-factorial study, where the visibility of displayed letters is judged based on color and size (see Fig. 3.1), in Sec. 3.2. We chose a simple example here, that is not directly linked to ECA evaluation, as it was easier to visualize and therefore allowed for better clarity in the presentation. In Sec. 3.3, we present an evaluation based on feedback from 11 independent study developers — computer science students or researchers in acoustics or psychology — who have used our framework. Rather than conducting an artificial study to compare our framework with existing ones or none at all, we deliberately chose to gather real-world experiences and insights from those actively using our framework. This approach allowed us to enhance our solution based on practical usage scenarios. After that we will also introduce one aspect specific to VR user studies conducted with HMDs, namely the need for a body-avatar (Sec. 3.4. We briefly present

[1] https://www.unrealengine.com
[2] https://www.unity.com

a solution that was also implemented to easily be integrated in studies set up and conducted using the StudyFramework.

3.1. Related Work

Several frameworks and tools exist that should support researchers from diverse research areas to implement, set up, and conduct experiments in VR using various rendering engines. Most of them are built on top of Unity. One of these is the *Unity Experiment Framework (UXF)*[3] [Brookes et al., 2020], which allows to specify experimental orders a priori or progressively using a session-block-trial model and also supports remote experiments outside the lab. Another possibility is *BMLtux*[4] [Bebko and Troje, 2020], which allows to implement and conduct factorial-design experiments, aiding experimenters in visually keeping track of the progress of each session. Further frameworks are the *Virtual Reality Scientific Toolkit (VRSTK)*[5] [Wölfel et al., 2021] and the *Unified Suite for Experiments (USE)*[6] [Watson et al., 2019], which both provide more integrated sensing capabilities, e.g., for brain activity. USE does that specifically by introducing a hardware device (*USE SyncBox*) to integrate measurements of electrophysiological recording devices with high-precision timing. VRSTK offers advanced features enabling experimenters to virtually immerse themselves within the IVE for enhanced interaction. Furthermore, it facilitates session replay and analysis. A recent framework that enables running distributed experiments with the potential for multiple remote participants is *Ubiq-Exp* [Steed et al., 2022]. It also supports conducting experiments both with and without an experimenter overseeing the process. Finally, *EVE*[7] [Grübel et al., 2017] allows the integration of a commercial plugin (*MiddleVR*) so that experiments can be run in CAVE systems, as Unity natively only supports VR using HMD. As this concludes our discussion on Unity frameworks, interested readers can find a detailed feature comparison of Unity frameworks for user study design and execution in [Wölfel et al., 2021].

Considering other frameworks besides Unity, *vexptoolbox* [Schuetz et al., 2023], provides more experimental control to the commercial VR platform *Vizard* (World-Viz, Santa Barbara, CA, USA) or a commercial solution for conducting VR experiments and exposition therapy: *CyberSession*[8]. Lastly, *R2VR* [Vercelloni et al., 2021] should also be mentioned, which allows to build simple VR experiments directly in *R*, a well-known statistical software environment. For experiments not requiring VR, often python-based frameworks are used, like *PyEPL* [Geller et al., 2007] or, even more often, *PsychoPy* [Peirce et al., 2022], which provides a versatile graphical user interface and is especially prominent for low-latency stimuli presentation and measurements. The latter was confirmed by [Bridges et al., 2020], performing a large-scale study comparing stimuli timing and latency for desktop-based experimental frameworks.

[3] `https://github.com/immersivecognition/unity-experiment-framework`

[4] `https://github.com/BioMotionLab/TUX`

[5] `https://github.com/ixperience-lab/VRSTK`

[6] `https://github.com/att-circ-contrl/use`

[7] `https://cog-ethz.github.io/EVE/`

[8] `https://www.cybersession.info/`

While the aforementioned frameworks support implementing and conducting an experiment to various degrees, there are also frameworks tailored to specific research domains, requiring minimal customization only. For example, *VREX*[9] [Vasser et al., 2017] aids in setting up experiments in the field of experimental psychology and neuroscience in complex virtual indoor scenes. There are several toolkits to build navigational studies, e.g., *PandaEPL* [Solway et al., 2013], *Landmarks* [Starrett et al., 2021], *NavWell* [Commins et al., 2020], or *DeFINE* [Tiwari et al., 2021], which require little to no coding. *VREVAL* [Bailey et al., 2022], an Unreal-Engine-based tool, facilitates the efficient setup and execution of studies aimed at evaluating architectural models. Another Unreal tool is *DomeVR* [Shapcott et al., 2022], which was specifically designed to run experiments with rodents but also humans in a dome-shaped display device. For acoustical research, *Oticon Medical Virtual Reality (OMVR)* [Pedersen et al., 2023] was developed and provides a variety of valuable virtual scenes.

To the best of our knowledge, there existed no general, open-source framework to design and conduct factorial-design studies using the Unreal Engine, e.g., comparable to BMLtux. Additionally, many of the mentioned tools, due to their specificity limiting adaptability, lack the modularity required for an easy and seamless integration into a research prototype for evaluation. Finally, Aguilar et al. convincingly argue that the auditability and reproducibility of studies are paramount, emphasizing the current limitation of many tools and frameworks in effectively addressing these crucial aspects [Aguilar et al., 2024].

As stated by Cunningham and Wallraven [2012], another important aspect when designing studies in VR is how to present the stimuli that should be evaluated and how participants are asked to rate those. Here, Robotham et al. [2022] compare different ways, in this case to rate audio stimuli, either by rating multiple stimuli side-by-side or having pairwise comparisons between the individual stimuli. However, in VR studies often questionnaires are used. Wagener et al. [2020] make a point for posing these questionnaires directly in VR instead of having participants fill them out afterwards using pen and paper. On the other hand, Graf and Schwind [2020] found questionnaires being filled out after the VR exposure to better differentiate the experienced VR scenarios. For Unity, Feick et al. [2020] developed a toolkit specifically designed for embedding questionnaires into VR, for the Unreal Engine we are not aware of an easily available solution.

3.2. StudyFramework Implementation

Addressing the aforementioned gap, we introduce a new framework, called *StudyFramework*[10], based on the Unreal Engine (developed for version 4.26, 4.27, and 5.3). The core idea is to provide a light-weight solution that supports developing and conducting user studies with the following main aspects:

[9] `https://vrex.mozello.com`

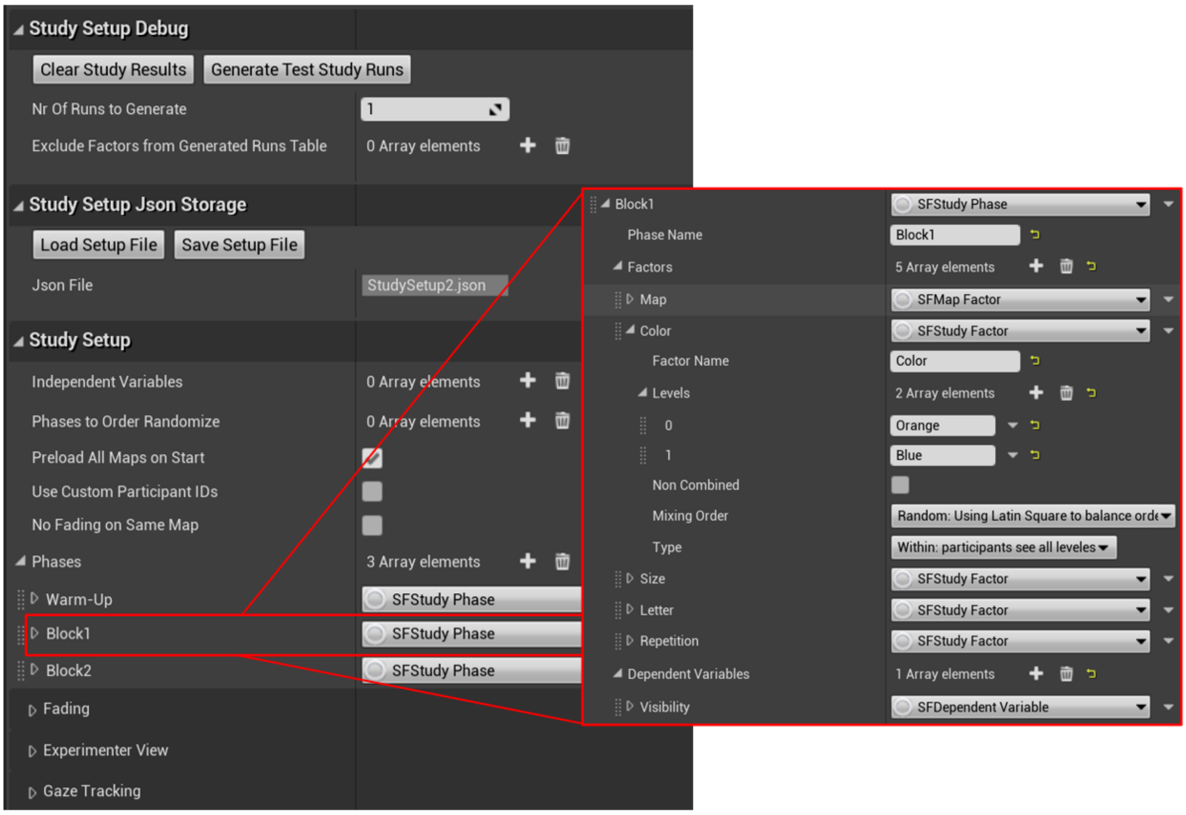
[10] `https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/unreal-study-framework`

- easy setup of factorial-design studies
- out-of-the-box solutions for randomization/counterbalancing
- thoroughly tested and redundant data logging
- simple graphical interface facilitating monitoring and controlling user studies
- focus on VR but also a possibility for desktop studies
- support for multiple VR platforms, like HMD and CAVE

In academia, a lot of studies are implemented and conducted by students evaluating their thesis projects, for example, having implemented a new interaction metaphor. Thereby they often lack both experience in the engine used and ample time to implement and thoroughly test their user study. Additionally, as mentioned above, researchers from other domains, like psychology, can benefit greatly from conducting VR user studies using game engines, while they potentially do not have a software development background. Consequently, one primary goal is to create a framework that is particularly user-friendly for individuals new to game engines, specifically within the context of Unreal Engine, as well as study design. While it is comparably simple to setup small interactive scenes in the Unreal Engine, mastering all intricacies to reliably conduct experiments (e.g., fading between scenes, having full control especially when something unexpected happens, reliably logging data, or counterbalancing orders) adds an extra burden onto novice developers. To this end, we provide the aforementioned functionality and made it accessible from both code (in this case *C++*) and also the visual scripting language provided by Unreal (called *blueprints*). The framework has been implemented as an Unreal plugin, ensuring an easy and seamless integration into any Unreal project. Another objective is to keep the framework lightweight, emphasizing its core purpose of facilitating the creation and execution of factorial-design studies. As a deliberate choice, we thus did not incorporate features unrelated to the study design or data management, such as immersive questionnaires. Users seeking this functionality can seamlessly integrate them through other plugins, like [Feick et al., 2020], or embed web questionnaires into the IVEs. We also developed a similar Unreal plugin to easily integrate Likert-scale questionnaires[11], which was however deliberately not integrated into the StudyFramework plugin directly.

These mentioned aspect separate this approach from most of the existing tools mentioned above. Grübel [2023] argues in a recent paper that there are already sufficient experimental frameworks. However, we are convinced that our framework can contribute to a crucial gap. First, due to its modular design, we avoid the issue of overly specialized and overloaded frameworks, allowing novice as well as experienced developers to use it. Second, to the best of our knowledge, it is the first of its kind for the Unreal Engine, setting it apart from the aforementioned plethora of Unity-based solutions. This is significant, because Unreal, in contrast to Unity, enables VR application for CAVEs [Cruz-Neira et al., 1992] through its native *nDisplay* plugin, thus enabling a larger range of VR display settings for the studies. This is especially important for us, as we run a CAVE with a 49-node cluster, called AixCAVE [Kuhlen and Hentschel, 2014]. Third, Unreal is open-source and its user-friendly visual-scripting blueprints provide an accessible programming interface, particularly suited for novice developers.

[11] `https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/likert-scale-plugin`



Study Setup Debug

Clear Study Results | Generate Test Study Runs

Nr Of Runs to Generate	1
Exclude Factors from Generated Runs Table	0 Array elements

Study Setup Json Storage

Load Setup File | Save Setup File

Json File	StudySetup2.json

Study Setup

Independent Variables	0 Array elements
Phases to Order Randomize	0 Array elements
Preload All Maps on Start	☑
Use Custom Participant IDs	☐
No Fading on Same Map	☐
Phases	3 Array elements
Warm-Up	SFStudy Phase
Block1	SFStudy Phase
Block2	SFStudy Phase
Fading	
Experimenter View	
Gaze Tracking	

Block1	SFStudy Phase
Phase Name	Block1
Factors	5 Array elements
Map	SFMap Factor
Color	SFStudy Factor
Factor Name	Color
Levels	2 Array elements
0	Orange
1	Blue
Non Combined	☐
Mixing Order	Random: Using Latin Square to balance orde
Type	Within: participants see all leveles
Size	SFStudy Factor
Letter	SFStudy Factor
Repetition	SFStudy Factor
Dependent Variables	1 Array elements
Visibility	SFDependent Variable

Figure 3.2.: The *Details* section of a *StudySetupActor*. In the third section, multiple phases are specified. The specifics of one of the phases is expanded on the right, specifically also expanding one of the factors (*Color*) details, so its levels can be seen.

3.2.1. Components

The core idea of this framework is to develop **factorial-design** studies (similar to [Bebko and Troje, 2020]). Factorial design is an experimental setup that consists of two or more independent variables also known as factors, with each factor having multiple levels. With a full factorial design, all possible combinations of the levels of a factor can be studied against all possible levels of other factors (in contrast to fractional factorial designs, systematically only showing a fraction of these conditions to each participant, but which are not directly covered here). Per condition (combination of specific levels per factor), data for one or multiple dependent variables (DVs) is gathered. Similar to blocks in the session-block-trial model (see [Brookes et al., 2020]), we structure the experiment in different phases for which specific factors and DVs are defined individually (see Fig. 3.2). Sometimes a task is repeated multiple times in the same condition, for example, if multiple selection tasks are performed consecutively, and data beyond a mean performance score should be collected. Therefore, in addition to the previously specified DVs collecting one value per condition, we introduced special multiple-trial DVs. While in the original design of the framework only one value per condition was collected for each DV, we extended this by special DVs, for which multiple trials per condition could be collected (called multi-trial-DV). This can be helpful if a task is repeated multiple times in one condition and data beyond the mean should be collected, like multiple trials of a reaction task that is performed in parallel to the main task. Lastly, also independent variables (IVs), going beyond the aforementioned factors, can be specified for which the data is collected at the very beginning of the experiment by means of input prompts shown on the experimenter screen (multiple-choice or text, for example, participant-specific data that should be reacted on). Similarly, participant IDs can be set at the start, but for counterbalancing a sequential number is also always stored internally.

Since one key idea of the *StudyFramework* is to require as little programming background as possible for study developers, we tried to utilize graphical user interface (GUI) elements of the Unreal Editor. To create a new study, developers first drag-and-drop a *StudySetupActor* (actor is the Unreal term for any object within a map) into an empty map. Then the details panel (see Fig. 3.2) is used to configure all crucial design aspects of the experiment, which will be described below. Thereby, the configuration is directly stored and updated into a configuration file in human-readable *json* format. Developers can switch those files, for example, to assess different study configurations during development (see the second section in Fig. 3.2). When starting the study, the setup information is parsed and one singleton object (derived from the *GameInstance* class of the Unreal architecture) is created which holds all interfaces, e.g., for logging or controlling the study. A reference to this object is easily accessible both in C++ and blueprints.

However, before a session for a specific participant is started, the order of the individual conditions has to be **randomized and balanced** according to the setup. The framework supports both within- and between-subjects factors (see [Cunningham and Wallraven, 2012]). When multiple within-subject factors are specified for a phase, the conditions are created by combining each level of one factor with all the levels of the other factors. This process results in the Cartesian product of all factors (see Fig. 3.1). The default case would be to balance the order of these generated conditions using Balanced Latin Squares [Edwards, 1951] so that the position at which each condition is presented and the condition after which each condition is presented is counter-balanced over participants to avoid potential position and order effects. To achieve this, we use the sequential number of the participants to pick an appropriate row from the Balanced Latin Square. Additionally, we shift the picked row of the Latin Square by the phase ID to avoid potential identical randomization in two repeated phases with the same factors. Moreover, our framework allows factors to always be presented *in order* (i.e., first all conditions with the first level and so on) or at least such that the same levels of one factor are presented *en bloc* For instance, if the virtual scene is varied, all conditions within the same scene can be shown back-to-back, to minimize frequent scene transitions. Obviously, only one factor per phase can be specified as either of the two. This option can also be used to implement repetitions by defining a repetition factor and setting it to "*in order*" so that all first repetitions are finished before the second repetitions start. Furthermore, sometimes there should be balancing, e.g., of a task, which is not a factor to be examined. This is possible by defining *non-combined* factors, which do not contribute to the Cartesian product and are potentially randomly mapped to the aforementioned conditions. As a last resort, there is also a specific callback function that can be implemented by developers and gives the possibility to reorganize or filter generated conditions. Further documentation and examples can be found in the project Wiki[12]. As a tool for checking the setup balancing for correctness, we added the possibility to generate the condition lists of an arbitrary number of participants and store them into a single text file for further inspection (see “Generate Test Study Runs” button at top of Fig. 3.2). In the shown example, *Size* and *Color* are 2-level factors, while *Letter* is a non-combined, randomly assigned factor with different levels in the *Warm-Up* and the *Block1* phase. Additionally, *Block1* has a third factor *Repetition* with two levels so that each combination of the first two factors is presented twice.

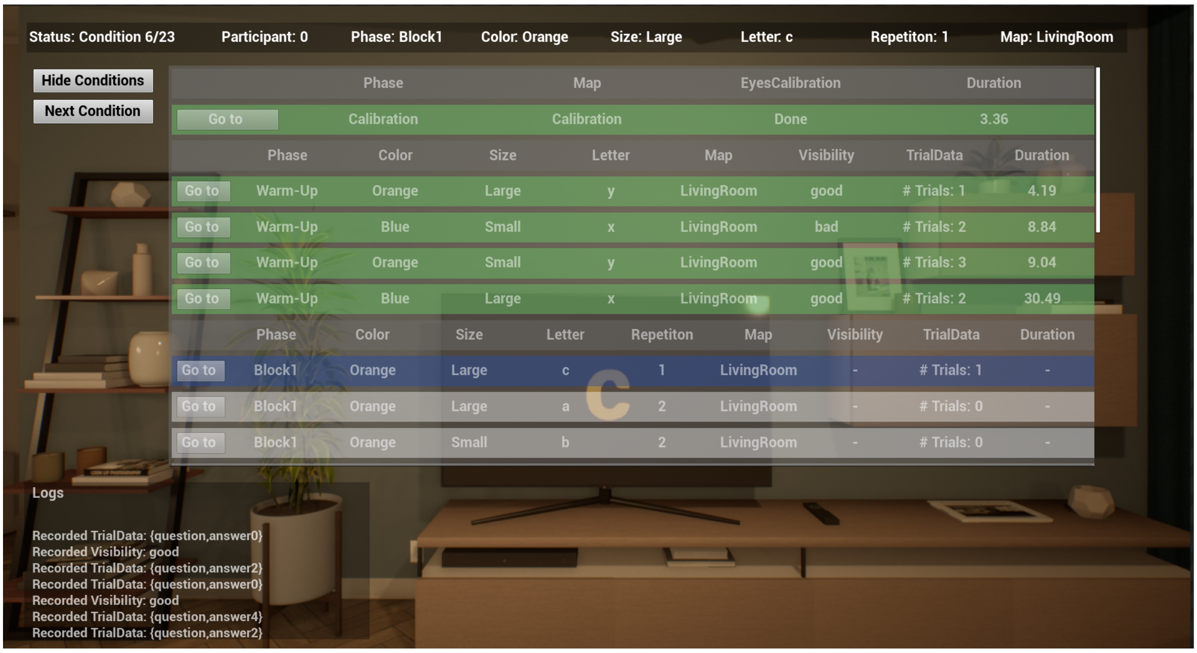
The framework also includes **logging** for positional data, data gathered for DVs, and potential events, all with timestamps. An event can be, for example, a participant interacting with a specific object, for which developers can log an arbitrary text, such as *“Object A picked up”*, using provided interfaces. Adding new actors of which position and orientation should be logged at each frame of the application, or less often if specified, is straight-forward by just adding a special logging component to these actors. Additionally, this component also allows to log custom data frame-wise, like the status of the actor or whatever is required in the specific use case. For each study phase a table in csv-format (comma separated values) is created, holding data collected for the DVs as well as the duration of each condition for all participants. These tables are created in long format, holding one line per condition and participant, as opposed to one line per participant, so that they can be easily loaded into statistics tools like R (see Tab. 3.1). The aforementioned multiple-trial DVs, do not fit into this format and therefore create one csv-file per variable with a line per recorded data point, so potentially multiple lines per condition and participant. Furthermore, all data is also logged redundantly per participant and session into a separate text file. In general special care was taken that all data is stored safely to avoid potential data losses.

ID	Gender	Phase	Color	Size	Let.	Map	Visib.	Time
0	male	Warm-Up	Orange	Large	y	LivingRoom	good	4.19
0	male	Warm-Up	Blue	Small	x	LivingRoom	bad	8.84
0	male	Warm-Up	Orange	Small	y	LivingRoom	good	9.04
0	male	Warm-Up	Blue	Large	x	LivingRoom	good	30.49
1	male	Warm-Up	Orange	Large	y	LivingRoom	bad	10.26
1	male	Warm-Up	Blue	Small	x	LivingRoom	good	7.44
				...				

Table 3.1.: Excerpt of an example phase log file, here for the example study setup used in Fig. 3.2 and Fig. 3.3. The csv file format is split into columns here for visibility. It contains the participant ID, independent variables (here: `Gender`), the phase name, factor levels (here: `Color, Size, Letter, Map`), dependent variables (here: `Visibility`) and the duration of the condition.

For an immersive study, a virtual scene is mandatory and is sometimes also varied as part of the study. It is therefore also formalized as a factor in our framework. When switching scenes,

[12] StudyFramework Wiki: `https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/unreal-study-framework/-/wikis/Randomization/Examples`



Status: Condition 6/23 Participant: 0 Phase: Block1 Color: Orange Size: Large Letter: c Repetiton: 1 Map: LivingRoom

	Phase	Map	EyesCalibration	Duration
Go to	Calibration	Calibration	Done	3.36

	Phase	Color	Size	Letter	Map	Visibility	TrialData	Duration
Go to	Warm-Up	Orange	Large	y	LivingRoom	good	# Trials: 1	4.19
Go to	Warm-Up	Blue	Small	x	LivingRoom	bad	# Trials: 2	8.84
Go to	Warm-Up	Orange	Small	y	LivingRoom	good	# Trials: 3	9.04
Go to	Warm-Up	Blue	Large	x	LivingRoom	good	# Trials: 2	30.49

	Phase	Color	Size	Letter	Repetiton	Map	Visibility	TrialData	Duration
Go to	Block1	Orange	Large	c	1	LivingRoom	-	# Trials: 1	-
Go to	Block1	Orange	Large	a	2	LivingRoom	-	# Trials: 0	-
Go to	Block1	Orange	Small	b	2	LivingRoom	-	# Trials: 0	-

Logs

```
Recorded TrialData: {question,answer0}
Recorded Visibility: good
Recorded TrialData: {question,answer2}
Recorded TrialData: {question,answer0}
Recorded Visibility: good
Recorded TrialData: {question,answer4}
Recorded TrialData: {question,answer2}
```

Figure 3.3.: The experimenter view overlay displayed over a demo study scene. Additionally to the status bar (top) and log section (bottom left), the condition list is currently shown (can be de-/activate by the top button on the left). In the list completed conditions are marked in green and the current condition is highlighted in blue. Already gathered data is displayed.

we recommend to use **fading**, i.e. transitioning to a predefined color and after a while back to the new scene, to not confuse participants by immediately changing their entire surroundings and potential lags due to loading. To this end a configurable fading is implemented that works in VR and desktop mode and also defines callbacks that allow developers to react to a new level being loaded or having faded in, e.g., by starting a task only once the new scene is faded in.

To support experimenters, we added an **experimenter view**. It contains a status bar (see Fig. 3.3), that always shows what status the application is in and which condition is currently presented, and a log section showing the latest log messages. Developers can decide for logged messages specifically whether they should also be shown in the log section, to give the experimenter all relevant information and simultaneously not overload this log so that relevant information might be missed. On clicking the “Show Conditions” button in this experimenter view a scrollable condition list can be displayed (see Fig. 3.3). There, on top of seeing which condition is currently active and which are already finished, recorded data of DVs is shown. Furthermore, to simplify experiment development and debugging, specific conditions can directly be started there without jumping through previous conditions. This functionality can also be used during study execution to restart a specific condition, e.g., if the participant was distracted by something else happening and missed the start of a condition or wants to repeat a familiarization phase. This restarting is obviously also noted in the participant’s log file, but should be used very carefully during execution, potentially having confounding effects. The experimenter view can be shown as an overlay on what the participant sees in the HMD or in a separate window potentially on a second screen in desktop mode. This study control functionality is expanded by the possibility of recovering failed study sessions. If on starting the study an unfinished previous study run is detected, the experimenter can choose whether to continue the study from the last unfinished condition or start with a new participant from the beginning. This is helpful for quick and clean recovery if the software crashes unexpectedly during study execution.

Beyond the scope of [Ehret et al., 2024a] and slightly contradicting our statement mentioned at the beginning, that the framework should only focus on the very core feature, we deliberately added functionality for **gaze tracking**. Since many studies, given an eye tracker is present, might want to evaluate participants' gaze behavior, we wanted to make this easily available for novice users without having to go through all the intricacies of integrating it themselves. As the gaze tracking does not interfere with any of the other functionality, it can also be easily ignored if it is not needed. There exist tools for analyzing gaze data, but they often focus on easier to track 360° videos or images, which are easier to analyze as gaze data and fixation points can be easily visualized on top of the stimuli themselves (see, e.g., [David et al., 2024]). In our case, however, gazes to potentially moving 3D objects should be tracked, while the user is also able to physically and virtually move through the 3D space. Currently, eye tracking is easily possible using the *HTC Vive Pro Eye* (support for other HMD using the eye tracking extension of *OpenXR* is planned). When activated the gaze tracker can continuously log the gaze direction, eye openness, and pupil dilation. Additionally, actors can be flagged as gaze targets, and by means of line traces it is checked whether one gaze target is currently gazed at. If no eye-tracking sensor is available, head orientation can also be used as a coarse proxy for gaze direction. However, this should be used with care since there is evidence that this approximation is problematic [Sidenmark and Gellersen, 2019].

3.3. Evaluation of the StudyFramework

The *StudyFramework* was already successfully used in 14 experiments related to human-ECA interaction (e.g., [Ehret et al., 2023, 2024b; Ermert et al., 2023; Bönsch et al., 2023a,b]) and successively updated and improved in that process. All developers of these studies (if not the first author of this paper) were asked to fill out a short subjective evaluation questionnaire after conducting their respective study. This questionnaire contained general questions with regard to the experience of the developers (see App. A.3), specific questions for features of the framework, and the System Usability Scale (SUS) questionnaire [Brooke, 1995].

In total we received filled-out questionnaires from 11 different study developers over the course of one and a half years of which one had to be excluded due to incompleteness. Of the remaining $n = 10$ projects, three were bachelor thesis and four master thesis projects. The remaining three were in the context of different PhD projects in the realm of acoustic, psychology, and VR research. Answers to the questions regarding prior experience, ease of development, and confidence during study execution, which were rated on a 5-point Likert-scale between 1 (*Strongly*

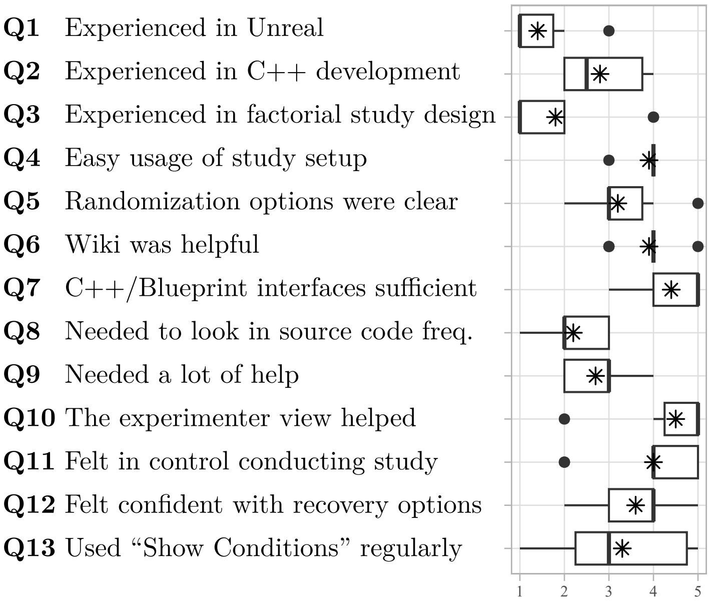


Figure 3.4.: The answers by $n = 10$ developers given to the individual statements as box plots on a scale from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). The stars indicates the mean while the boxes show the quartiles with whiskers extending to the full range and outliers being displayed as dots. Statements are given here as shortened versions, full-length statements can be found in App. A.3

Disagree) and 5 (*Strongly Agree*), can be found in Fig. 3.4. Additionally, when asked what the most helpful feature was, five study developers (50%) stated the status bar, three (30%) the "Next Condition" button, and one each stated the output log (10%) and "all of them" (10%). When asked for the least helpful feature seven participants (70%) answered "None", while the "Next Condition" button, the "Show Conditions" button, and the output log where picked by one developer (10%) each.

Evaluating the results of the SUS yielded a mean score of 76.5 ($SD = 16.6$) on a scale of 0 to 100, which is considered a "good" usability score. When looking closer at the individual scores, in five cases the framework was rated above 80, which constitutes an excellent score, however, in one case it was even rated as low as 40 (while all other ratings stayed above 65, which is the average SUS rating). Following up on the free field comments in this particular case did not yield any insights on a specific shortcoming.

However, when looking at the free comments of all developers regarding implementation ease, there are a few inconveniences and feature requests mentioned. Some of them (like logging custom data or debugging functionality requested) were already solved during further development of the framework, but two still persist. One study developer stated that *"The nested design of the study setup (phases etc.) was at times hard to digest, partly also because of the small fonts"* (see Fig. 3.2). Another comment hinted at the missing possibility to dynamically insert, e.g., *"a phase only between specific conditions of another phase"*, which goes beyond standard factorial design and was thus consciously not added. However, also several comments stated, e.g., *"the framework makes creating a study very easy, especially if you are new to Unreal"* or that *"it was easy to use for a beginner in study design and actually helped to understand the structure of studies with (in)dependent variables, factors and conditions"* and that the StudyFramework's *"blueprints were really nice and easy to use."* Looking into the comments regarding conducting the actual study, the only negative comments regarded shortcomings that were already fixed in the meantime or feature requests that would go beyond the targeted scope of this framework. One developer, for example, requested the possibility to reset the participant's position which is very specific and should therefore rather be implemented for a study individually. Apart from that, there were positive comments like *"the possibility to restart the study where the participant left of as well as the good logging came in very handy"* and *"during execution of the study, it was pretty good. I could not think of something to immediately improve."*

3.3.1. Discussion

When looking at the results of questions **Q1 - Q3**, we can conclude that we had rather novice users: While some of them were more proficient in C++ development, nearly all of them were very new to Unreal and study design. Although the framework should also benefit more proficient developers in not having to write study code from scratch, the plugin was especially designed with Unreal novice users in mind. Therefore, it is very encouraging that they found "the usage of the StudySetupActor clear and easy" (**Q4**, $M = 3.9$, $SD = 0.3$). The "different randomization and ordering options" in this setup, however, were apparently experienced as less clear (**Q5**, $M = 3.2$, $SD = 0.9$). This is also illustrated by the comment regarding the complexity of the nested design, probably referring to the nested setup of phases, factors, and levels. Partly, this might already have been improved, since we simplified the condition creation and balancing during further development, e.g., removing dedicated repetition functionality which could just as well be implemented with an additional factor. At the same time also more documentation and several examples were added to the Wiki of the respective git-project[12]. However, there might also be room for improvement with regard to the clearness of the randomization and balancing. The Wiki in general and the provided interfaces were rated well above average (**Q6** & **Q7**). This can also be seen in the low ratings to the statement "I had to look into the source code frequently to understand what was going on" (**Q8**, $M = 2.2$, $SD = 0.8$). However, looking at the answers to "I needed a lot of help to develop the study" (**Q9**, $M = 2.7$, $SD = 0.7$) reveals that still some additional help was required, albeit that these ratings drop over time (the only "4"-rating was at the beginning of the evaluation, while the last ratings were "2"). That potentially hints that the quality improved over time, but could also be caused by other factors. Furthermore, the statement does not clearly differentiate between needing help to understand the *StudyFramework* or with study design in general.

These observations together with the relatively high system usability score bring us to the conclusion that the developed framework is well usable by novices to set up factorial-design user studies, which was one of our main goals. The pursuit of simplicity inevitably limits the ability to accommodate more intricate configurations, such as those requested by a developer to dynamically adjust conditional orders based on participants' performance or choices. But overall the provided functionality seems sufficient and user-friendly.

Another important aspect of the framework is the support for experimenters during the execution of a study by means of a GUI. Developers were very grateful for the experimenter view and agreed that it was helpful (**Q10**, $M = 4.5$, $SD = 1.0$) and greatly agreed with the statement "I felt in full control over the study" (**Q11**, $M = 4.0$, $SD = 1.2$). Only one developer, using an early version of the framework, rated both with 2. Reported problems were subsequently fixed. Another participant asked for the possibility to restart an already started condition in the conditions list (using the "Go to" button, see Fig. 3.3), which was initially deactivated during study runs but consequently provided to grant experimenters full control. A similarly enabling functionality is the recovery on crashed study runs (**Q12**, $M = 3.6$, $SD = 1.1$). However here, experimenters did not feel just as confident. This might have come, because they did not test it thoroughly before starting the study and did not implement it themselves, so they were not entirely sure what would happen. We, however, implemented this feature in a way that no data loss can appear, because potentially removed data from an already started, but not finished, condition would be backed up before removal. Generally, the features within our experimenter view often go unimplemented in many studies that are developed from scratch. Due to time constraints, essential functions are prioritized over convenience features, relegating these valuable additions to the bottom of the priority list. This is again a strong argument for using our proposed framework, as feeling in control while conducting a user study is a very reassuring feeling for the experimenter and potentially also increases the number of valid and useful data sets gathered. Which particular feature was useful for which study had a larger spread, this can be seen in the answers to **Q13** but also in the answers to the most and least helpful feature. While the status bar was liked most by a majority of experimenters, there is no clear preference for a least helpful feature that potentially could be removed, hinting that all implemented features are well integrated and supportive.

However, the evaluation is based on a limited number of subjective responses ($N = 10$) and can therefore only be generalized with caution. Furthermore, as stated, the framework was developed continuously during data gathering so that some features might have become more helpful in the process. Nevertheless, we are confident that the presented *StudyFramework* provided value to the surveyed developers and facilitated their development as it will for future study developers.

For future developments of our framework, our focus will be on enhancing user-friendliness, enabling new as well as experienced developers to quickly and effortlessly set up new studies, minimizing the likelihood of unforeseen issues arising during or after study execution. Thus, our strategy revolves around optimizing and streamlining our current functionality based on incoming developer feedback and future Unreal versions, while we will only incorporate new features that align with our vision of a light-weight, general framework in response to user requests. Additionally, we appreciate the initiative of Aguilar et al. [2024] to increase the reproducibility of experiments. While our *StudyFramework* already contributes to this goal, we believe that the use of widely accepted common formal descriptions could further enhance reproducibility. Consequently, we are open to engaging in discussions to determine the appropriateness of our current *json* file implementation or explore whether other data formats might offer superior compatibility and reproducibility benefits.

From the reported studies only approximately half used gaze tracking, and were happy with the availability, especially since an efficient implementation is not straightforward and should run asynchronously to the main thread. After the publication of the StudyFramework the eye tracking was extended beyond the HTC Vive Pro Eye headset to include all eye tracking capable devices using the eye tracking extension of *OpenXR*, which was already used in [Ehret et al., 2025b]. To date, the framework has been utilized for 17 VR-based user studies.

3.4. Avatar Plugin

Due to the broad adaptation of HMDs in recent years, another aspect to consider when conducting human subject studies are body avatars. This need arises, because the view to the participants' own body is blocked, in contrasts to projection-based VR as in CAVEs. These body avatars are supposed to evoke a *body ownership illusion*, so the feeling of actually accepting the virtual body as one's own, and the sense of *agency*, i.e., feeling in control of the virtual body [Spanlang et al., 2014]. It was found that avatars with high motion fidelity are beneficial [Cao et al., 2023], especially when engaging in social scenarios [Smith and Neff, 2018], which is the main focus in this work.

Virtual representations can vary from representing just the hands, over additionally visualizing a floating upper body to full-body visualizations [Lugrin et al., 2018]. Thereby the visual representations can be custom-made to fit the appearance of the user as closely as possible or rather generic (see, e.g., [Waltemate et al., 2018; Hepperle et al., 2022]). However, when integrating avatars it should also be carefully considered that those can also influence behavior of the users. For example, users potentially involuntarily adapt to incongruent avatar movement [Boban et al., 2023] or perceive time differently [Unruh et al., 2023]. Another way in which avatars can alter user behavior is based on their virtual representation, the so called Proteus Effect [Yee and Bailenson, 2007]. Yee and Bailenson [2007] found, for example, that participants acted more confident when being embodied in a taller and more attractive self-representation, but similar effects were also shown for other avatar qualities. A literature review for the effects of synchrony, realism and body completeness on self-embodiment can be found in [Yassien et al., 2020], which states that full-body representations significantly enhance social presence and are therefore crucial in the scope of this work.

However, the virtual body and the real body are in most cases required to match in size and proportions, to accurately steer the virtual representation with one's own movements. The most straight-forward solution is to measure the height of a user (for example by using the HMD and an upright-standing calibration pose) and then uniformly scaling the entire model to this height. However, this does not take into consideration differences in body proportions. When using an avatar, this becomes, for example, particularly obvious if the physical arms are fully extended but the virtual arms are still bend. A solution to this is presented by Pujades

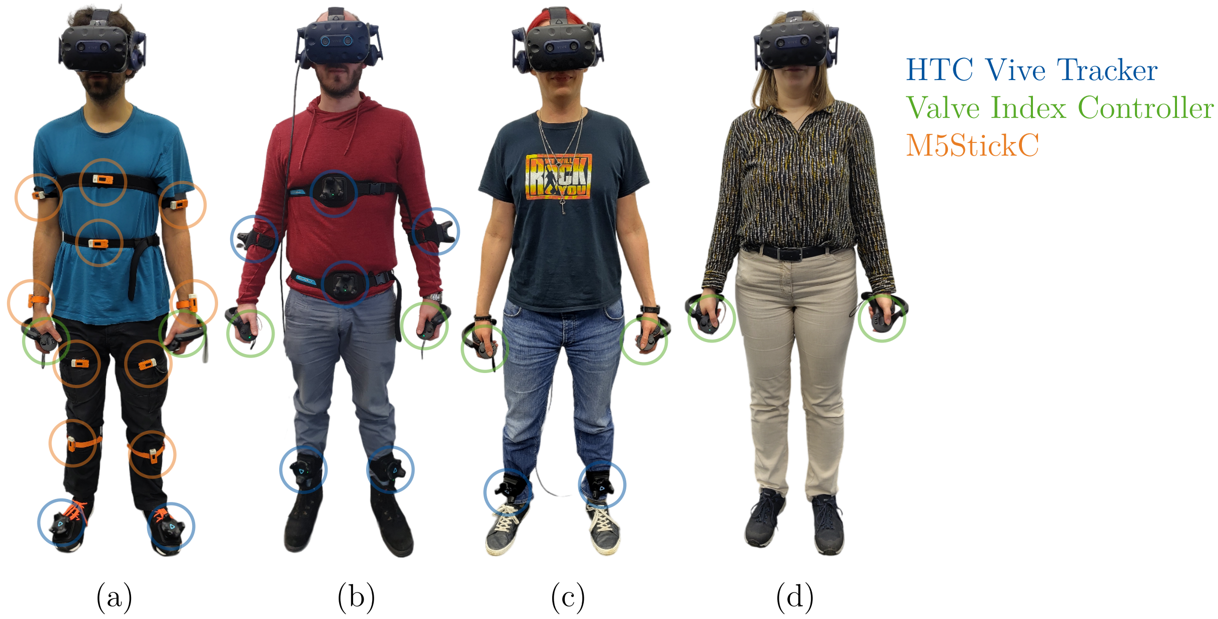


Figure 3.5.: Different tracking setups used. 10 IMUs with 2 additional Vive Trackers, Valve Index Controller and HMD (a); Full tracking setup with 6 Vive Trackers, HMD, and Valve Index Controllers (b); Reduced setup using only the HMD, foot trackers and controllers (c); Only HMD and controllers (d).

et al. [2019] with their *Virtual Caliper* system. In this system several predefined poses have to be made with the controllers to measure respective body proportions (see Fig. 3.7(left)). If the avatar should also visually resemble the user, it is possible to scan the actual body of the person (see, e.g., [Waltemate et al., 2018]). This requires, however, a dedicated, potentially expensive scanning setup and can be very time consuming. The process can be made simpler with recent scanning applications which only use a smartphone (e.g., [Menzel et al., 2024]). These are, however, still error-prone and results are often not as good as specifically designed 3D models. However, time-efficiency calibration is in many applications and especially user studies crucial, so size-only calibrations of existing 3D models are often preferred.

Valvoda et al. [2007] showed that embedded avatars should consistently follow the movements of the user's physical body. While they did their study using exocentric avatar, this also transfers to egocentric avatars as shown They further found that one key feature that users appreciate most is tracking the feet and applying their motion to the avatar. The highest fidelity can be achieved with optical tracking using a room-mounted tracking system with multiple camera and tracking markers attached to the user (see, e.g., [Spanlang et al., 2014; Chan et al., 2011]). These systems, however, are expensive and require a lot of space and time to set up. The most common alternative is to utilize the hardware users use anyways for interaction with the VR application, namely the HMD and tracked controllers. While these can be used to track head and hands the remainder of the data has to be approximated (see, e.g., [Winkler et al., 2022; Fletcher et al., 2023]). However, additional connected trackers (using the same tracking technology as HMD and controllers) can be used to actually measure this data. Another option are inertia measurement units (IMUs) which can be used to measure the movement of other body parts (see, e.g., [Yi et al., 2021; Cha et al., 2021; Roetenberg et al., 2013]). These IMUs, however, cannot measure their positions directly but can only be used to estimate position and orientation. More specifically, they measure orientation change and linear acceleration and by integrating these measurements compute orientation and positional offset. Due to this integration, however, measurement inaccuracies can lead to drifts in the estimated values over time. Recently also tracking using a single camera gained attention and produces surprisingly accurate results (see, e.g, [Güler et al., 2018; István et al., 2023]). However, these tracking results often still contain temporal noise and discontinuities, rendering them inferior to the previously discussed techniques for tracking avatars. Furthermmore, an aspect to consider with all of these tracking setup is latency, since high latency can have detrimental effects [Waltemate et al., 2016]. A thorough review of different tracking technologies can be found in [Zhou and Hu, 2008].

3.4.1. Avatar Animation

For our own implementation we used an HMD that uses *SteamVR* lighthouses for tracking, for example the *HTC Vive Pro Eye*, which additionally provides eye tracking which can be used. Adkins [2022] found that articulated hands improve user comfort and immersion, therefore rudimentary finger tracking was implemented using the *Valve Index Controllers* which can be secured to users' hands with straps and can measure the distance for each finger from the controller. They provide for each finger a scalar value describing the curl and the spread between each neighbouring pair of fingers. This data can directly be applied to the skeleton joints in the hands of the avatar model. Additionally to this we used in most setups *Vive Trackers* [Borges et al., 2018], to provide tracking anchors for the feet and potentially other body parts with high fidelity, as the use the same inside-out tracking technology as the HMD. A base setup with only two trackers attached to the feet can be seen in Fig. 3.5(c).

For high fidelity tracking of body parts in between those anchors, we initially evaluated the use of IMUs due to their cost-effectiveness and the potential for deploying multiple units to enhance overall accuracy. We utilized *M5StickC* micro controllers, which are comparatively cheap, can be equipped with an IMU, and provide wireless networking capabilities and a display and buttons for easy control (see Fig. 3.5(a)). We were, however, not able to mitigate the aforementioned problem of drifting, also using more advanced data integration, like Mahony filters [Mahony et al., 2008]. In our implementation the drift was on average $0.75°/s$, which rendered this method not usable for us, since for example the elbow would have rotated by 45° after one minute for a user standing perfectly still. Consequently, we abandoned the idea of using IMUs and instead used additional trackers using the lighthouses for tracking. As trade-off between tracking accuracy and setup time and cost, we opted for four additional trackers. These trackers were strategically placed to capture common movements that cannot be directly inferred from head, hand and foot positions. Therefore, pelvis and chest trackers enabled the capture of more complex torso movements such as twisting and bending, which would otherwise need to be estimated from other sensors. Thereby movement fidelity is gained, particularly when distinguishing between back bending and downward pelvic movement or differentiating torso rotation from head turns. Additionally, lower arm trackers facilitated the disambiguation of arm positions when hands remained fixed, allowing for free movement of the elbows with arms not fully extended. This setup can be seen in Fig. 3.5(b) and is further called 6-tracker setup, as in total 6 Vive Trackers are used additionally to HMD and controllers.

We use the full-body inverse kinematics (IK) solver integrated into Unreal Engine to apply the tracked data to the body avatar. Since this solver operates on the virtual skeleton within the 3D mesh of the avatar model, it is necessary to transform the position of the trackers attached to the user's body into the reference frame of the skeleton used for the body avatar. To this end, users have to take up a reference pose and align their physical limbs as closely as possible with the virtual body and confirm the alignment by pressing and holding a button on the controller for one second (to avoid accidentally confirming too early). Initially, we employed a T-pose as the reference pose, where the user has to extend both arms to the sides. However, we transitioned to a pose with both arms extended forward (see Fig. 3.6(right))), allowing users to simultaneously see both virtual representations of the hand/arm tracking positions and the virtual arms to be matched,. This increased the accuracy of the alignment procedure, since the feet do naturally not moved once placed at the right virtual positions, but the arms tend to move, especially when turning the head to look at the other hand. This alignment procedure relies on the user's perception only, so it can be repeated if users afterwards experience their virtual body to not move in sync with their physical one. The rotational and positional offset of the trackers and their associated positions on the skeleton during the reference pose are captured and stored in a file to be further used throughout the same session even when having to restart the Unreal application, for example due to other technical problems. With these offsets and the current tracker positions and rotations the IK solver can be used to position the virtual avatar model as closely as possible to the physical body of the users. To ensure physiological plausibility and enhance the realism of the avatar's motion while reducing artifacts that may arise from inaccurate tracking, certain constraints are applied to condition the IK solver. For example, knee rotation is limited to one axis only.

In practice this 6-tracker setup still required a few minutes when setting up individual participants during user studies. Anecdotal observations during study sessions have indicated that participants move less and avoid movements such as co-verbal gestures, when having all the trackers attached to their bodies, potentially resulting in less natural behavior. Therefore, we introduced a reduced set of trackers using only two trackers at the ankles for foot tracking additionally to the HMD and controllers (see Fig. 3.5(c)). For a faithful tracking, we had to do some more adaptations before running the IK algorithm. First of all the pelvis has to be estimated from head and feet positions. Thereby the orientation of the pelvis is always estimated from the feet positions, such that the forward direction of the pelvis is perpendicular to the connection line between the feet (see Fig. 3.6(left)), while always staying upright. The position of the pelvis ($\mathbf{p}_{\text{pelvis}}$) is based on that of the feet ($\mathbf{p}_{\text{left/right}}$) and head ($\mathbf{p}_{\text{head}}$). The horizontal position is computed based on [Roth et al., 2016] as:

$$\mathbf{p}_{\text{pelvis}} = 0.4 \cdot \mathbf{p}_{\text{left}} + 0.4 \cdot \mathbf{p}_{\text{right}} + 0.2 \cdot \mathbf{p}_{\text{head}}$$

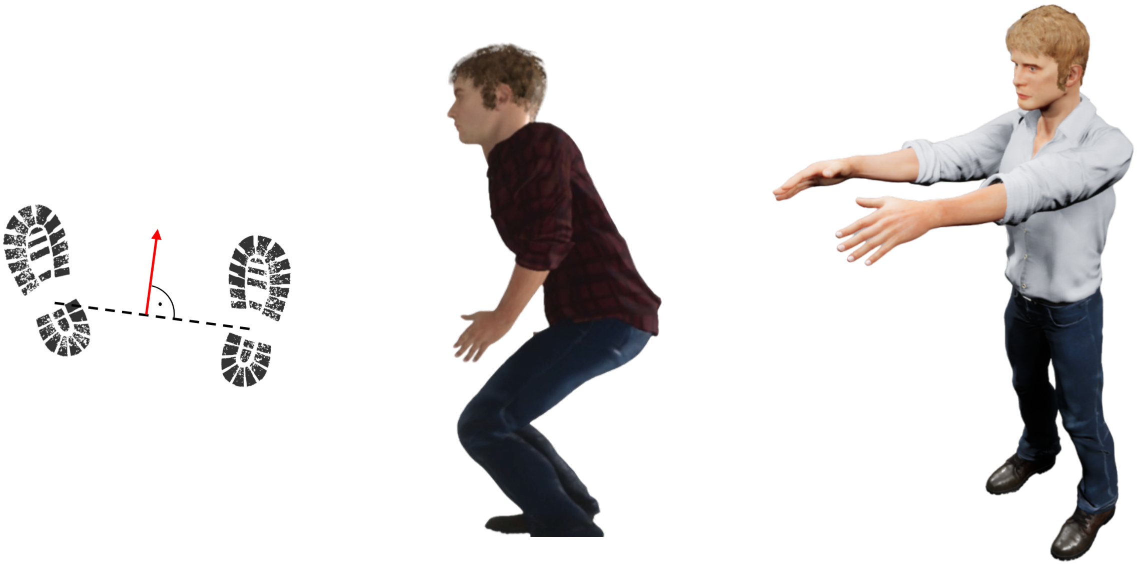


Figure 3.6.: The orientation of the pelvis (red arrow) based on the foot positions (left). A side view of the crouching movement, especially showing the backward movement of the pelvis (middle) and the pose used to align the virtual and the physical body (right).

The height (h_x denoting the height of x) of its position is set using the quotient $q_\text{pelvis} = \frac{h_\text{pelvis,cal}}{h_\text{head,cal}}$ of the used skeleton when standing upright in the calibration pose (cal) to

$$h_\text{pelvis} = h_\text{lowerFoot} + q_\text{pelvis} \cdot (h_\text{head} - h_\text{lowerFoot})$$

Additionally to the implementation in [Roth et al., 2016], we also moved the pelvis farther backwards the lower the head goes along the forward ($\vec{v}_\text{forward}$) vector estimated by the foot positions by

$$\Delta\mathbf{p}_\text{pelvis} = \frac{1}{4} \cdot (h_\text{head} - h_\text{head,cal}) \cdot \vec{v}_\text{forward}$$

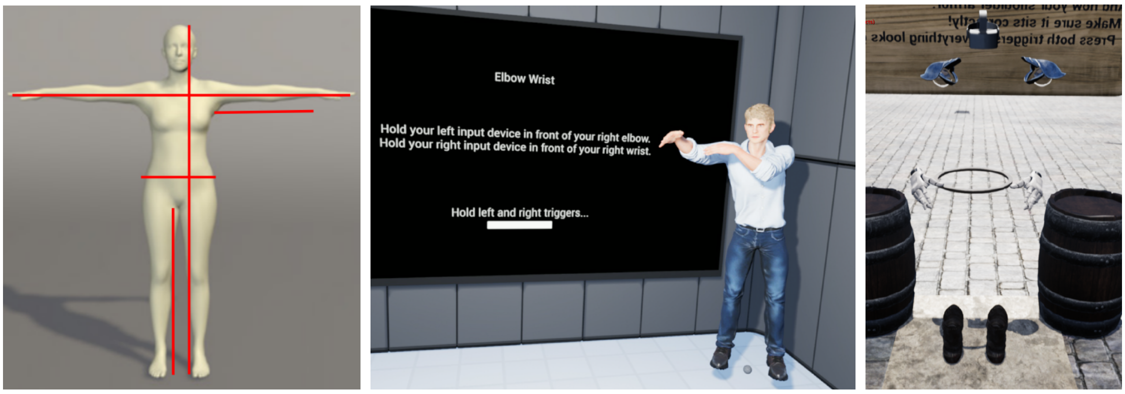
This factor proved to produce good looking movements during extensive testing. Furthermore, some constraints are used to condition the IK solver to make the backbone to move strongly only if the movement cannot be performed by the limbs, and keep the shoulders at reasonable positions.

As Debarba et al. [2022] stated accurately tracking the feet is important. In our studies we, however, often dealt with static scenarios which do not justify the effort of attaching foot trackers. For these scenarios we also implemented the flexibility of omitting foot trackers (see Fig. 3.5(d)). In this case the animation system always keeps the virtual feet on the floor below the head moved 15 cm to each side and facing in the direction of the head. Since foot positions should not be corrected on every movement of the head, we added thresholds for position deviations (Θ_pos =20 cm) and the rotation (Θ_rot = 45°). If one foot's divergence form the optimal placement in either rotation or position is above the respective threshold, the foot is re-positioned. For this, the position is linearly interpolated while the foot is smoothly raised to 6 cm height and set down again along this movement. While one foot is moving the other cannot be moved. Just once the first foot is set down the second can be moved if it violates one of the thresholds. Otherwise the feet are kept in place again until another threshold is passed. This procedural animation was particularly easy to implement due to the employed IK system. An alternative to this procedural approach could be to use motion matching [Fletcher et al., 2023] or entirely learned models [Winkler et al., 2022] for the lower body. *QuestSim* [Winkler et al., 2022], for example, implemented a data-driven model that is able to produce convincing full-body movement only from head and hand positions, taking also light body sways into account to synthesize the leg movement, alleviating the burden of attaching additional tracking hardware. Another alternative to using IK altogether would be physics-based animations using reinforcement learning (see, e.g., [Llobera and Charbonnier, 2022]).

We used this plugin in several studies, using the 6-tracker, the 2-tracker and the 0-tracker setup (see Fig. 3.5). As stated before, we qualitatively observed during multiple studies that participants tended to move less articulated wearing the 6-tracker setup, while needing more time to be equipped with the trackers. When only using the 2-tracker setup in another study there were no remarks by participants regarding missing fidelity of the tracking. On the contrary, participants were still very positive about the avatar moving with their physical body and did not complain about unexpected movement artifacts. Based on this observation, we hypothesize that the 2-tracker setup is sufficient for many applications, particularly when incorporating interactions that do not require nuanced body language. We even found, that in scenarios where participants are asked to remain at designated virtual positions (e.g., [Ehret et al., 2025b], further explained in Sec. 6.1) the 0-tracker setup sufficed since participants typically did not move their feet significantly. However, in use-cases such as collaborative mediated interactions with other humans or ECAs, where conveying subtle movements such as stance shifts becomes crucial for effective communication and shared understanding, we recommend using the 6-tracker configuration as it potentially better conveys this information. Additionally, for situational contexts where participants move we would strongly recommend to use foot trackers, as internal testing showed that agency is degraded if the feet are algorithmically moved, since often the wrong foot is taken and the movement in general does not fully resemble the actual movement. If a study in a non-stationary context without foot trackers should be performed, we would advice to take a closer look at the solution proposed by Winkler et al. [2022]. In conclusion, we recommend careful consideration of the tracking configuration based on the situational context of the interaction. For our upcoming settings, however, the 0-tracker and 2-tracker configurations are sufficient and thus used where applicable.

3.4.2. Avatar Calibration

In the previous section we already introduced the alignment process to attach the virtual to the physical body. However, as explained before it also important to have the virtual body's proportions to map those of the physical body to produce movements with high fidelity. The most straight-forward approach to this is simply scaling the virtual body to match the height



Elbow Wrist

Hold your left input device in front of your right elbow.
Hold your right input device in front of your right wrist.

Hold left and right triggers...

Figure 3.7.: Some body measurements used by [Pujades et al., 2019] exemplary visualized (left). An ECA showing specific calibration poses that have to be reproduced by the user for calibration during our first avatar calibration study (middle, image from thesis of Patrick Nossol) and the user at the end of the gamified calibration having attached boots, belt and should armor to their virtual body (right, image from thesis of Marius Meier-Krüger).

of the user's body (measured, for example, using the HMD and telling the participant to stand up straight). However, in humans also the limbs are differently scaled, therefore taking more measurements (see Fig. 3.7) can be appropriate, especially in context like virtual cloth try-ons or ergonomics (see [Pujades et al., 2019]). In the Virtual Caliper approach [Pujades et al., 2019] users have to take poses with the VR controllers, shown in a non-immersive instructional video, to take 3D measurements of their body. As reproducing these from 2D pictures/videos or even textual descriptions can be hard and error-prone, we implemented a similar system but with an ECA showing these poses directly in VR (see Fig. 3.7(middle)). In a small usability study (which was conducted with one person only due to Corona regulations at that time), we found that the animation were preferable over a text-based explanation of these poses only, by means of usability, using the System Usability Scale [Brooke, 1995].

Going beyond that, we also developed a gamified version of this visual pose-matching calibration process (see Fig. 3.7(right)). The idea here was that participants have to attach different accessories to their virtual body, like a belt or shoulder armor, and then have to hit some targets with a sword or kick them. By this we can measure, for example, shoulder and hip location and from the legs' and arms' swinging movements their length could be approximated. While many participants (80%) liked this system best it was also by far the most time-consuming one (M = 152 s, SD = 41), compared to the ECA guide (M = 121 s, SD = 42).

The gamified calibration, however, also introduces a new scenario which might not be aligned with the study content. In our recent studies, we therefore simply used the height-only calibration, which can be accomplished in a few seconds and yielded satisfactory results in our observations. However, here most participants weight-wise matched the available gender-matched avatar models. For a broader population, this calibration might still yield avatars not matching their body size, which potentially has unwanted psychological effects (see, e.g., [Mölbert et al., 2018]). This is especially true for context where the virtual body plays an important role (for examples see [Pujades et al., 2019]), unlike the ECA-human interactions we evaluate in this work.

Chapter 4

Verbal Behavior

As stated by Gratch et al. [2002], speech is an important aspect to create believable, embodied conversational agents (ECAs), as it serves not only as a medium for conveying information but also as a key factor in establishing connections and social presence between users and these agents. In this chapter, we will investigate various aspects that are crucial for designing and implementing ECAs, aiming to identify the most effective settings for for a natural verbal behavior. This chapter will encompass the voice used by ECAs and the prosody of the speech (Sec. 4.1) and the auralization (Sec. 4.2), with special attention to how speech sound is radiated. Aspects which are not directly related to the speech itself, like articulation movement etc. will be discussed in the following chapter concerned with co-verbal behavior.

4.1. Voice and Prosody

The contents of this section are based on and taken in part from work previously published in [Ehret et al., 2021].

ECAs are designed to exhibit interactive capabilities that enhance user engagement, which include plausible animations, non-verbal and verbal cues. Among these features, speech stands out as one of the key modalities, and developers strive to make it as natural and fluid as possible to further facilitate meaningful interactions with these agents. While the most natural option for the speech content is to record a voice actor, this is very labor- and cost-intensive [Georgila et al., 2012]. Therefore, text-to-speech (TTS) synthesis is often used (e.g., [Schröder et al., 2011; Wang et al., 2017; Shen et al., 2018]), which creates speech audio from text input only and can also be used in flexible real-time scenarios. While there exist approaches going even

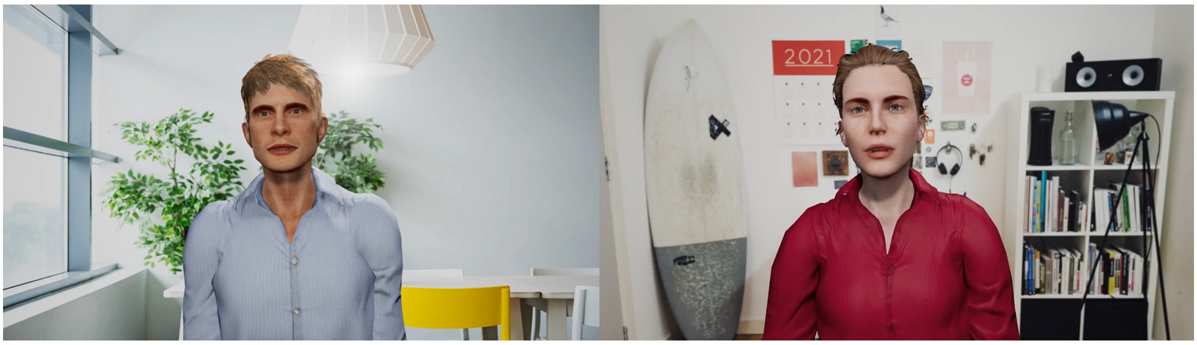


Figure 4.1.: Side-by-side visualization of two different frames of the used stimuli in the audio-visual condition E_{ECA}. The agents are animated using face recordings of real speakers and engage in a four-sentence conversation of about 30 s length, in this case organizing the next football training ($S4_{\text{training}}$).

further by, e.g., incorporating more information into the synthesis process, such as concept-to-speech (CTS) [Hiyakumoto et al., 1997], the present work will focus on the easier-to-use and more common TTS method.

There exists a large body of research comparing synthetic speech with prerecorded speech by trained speakers. For example, Chérif and Lemoine [2019] found that anthropomorphic agents with a human voice elicit stronger social presence than those with a synthetic voice. Chateau et al. [2005] evaluated the emotional response in participants comparing voice quality levels and Do et al. [2022] found that higher speech fidelity led to higher persuasiveness. Higgins et al. [2022] found detrimental effects of synthetic voices used for photo-realistic ECAs, while Lam et al. [2023] found that participants had a clear idea of which (differently-pitched) voices matched which visual ECA representation best. Krenn et al. [2017] looked into the social effects of synthetic voices incorporating dialects. Davis et al. [2019] compared non-native judgements of prosodically expressive as well as neutral human utterances with TTS. Malisz et al. [2019] used deep learning techniques to adapt the prominence of individual syllables of synthetic speech to the prominence in natural speech. By this they tried to improve the naturalness of the synthetic speech but did not find significant improvements in naturalness ratings. Miniota et al. [2023] trained a TTS model based on spontaneous speech to generate more natural synthetic speech and found that their solution was preferred over commercial solutions, albeit not rated as natural as recorded speech. An in-depth analysis of synthetic voices in human-agent interaction by Seaborn et al. [2021] provides a summary of many studies evaluating different dimensions when comparing synthetic and human speech. Other studies investigated the effect of synthetic speech on either computer-sided (e.g., [Gálvez et al., 2020]) or user-sided alignment/entrainment (e.g., lexical and syntactic alignment in [von der Pütten et al., 2016]). Von der Pütten et al. [2016] did not only examine the effect of synthetic as compared to prerecorded speech, but also how the ECA (in their case as robotic representation) was embodied, differing between an actual robot, a virtual robot, and no embodiment at all. They did not find an effect of synthetic speech on human-likeness, which can, however, at least partly be accounted to the robot-like visual representations used. Similar studies (e.g., [Cohn et al., 2020]) also enhance this comparison by a more articulated talking head present in the physical space of the participants, namely a Furhat [Moubayed et al., 2012]. While there is an open discussion as to what kind of voice to use for non-human devices like smart speakers [Cambre and Kulkarni, 2019], the focus on this dissertation lies on anthropomorphic agents.

Despite the rapid improvement of TTS technology in recent years, human listeners tend to rate synthetic speech as less natural [Kühne et al., 2020] while modern synthetic voices are reaching the level of human voices [Seaborn et al., 2021]. The preference of human voices may partly be attributed to an inadequate *prosody* of the synthesized speech, surfacing for example as the wrong placement of lexical stresses, pitch accents, and pauses, sometimes leading to a "broken" rhythm, or by inappropriate intonation contours (e.g., [Cutler, 1980]). When comparing synthetic and human voices for ECAs, Cabral et al. [2017] copied the natural human prosody in their synthesis and Davis et al. [2019] used different levels of expressive human prosody, but they both did explicitly not evaluate inadequate prosody as commonly present in off-the-shelf TTS solutions. To close the research gap on the effect of inadequate linguistic prosody for German native listeners, we investigate how prerecorded human speech featuring the same inadequate prosody as synthetic speech from off-the-shelf TTS solutions is rated regarding its perceived naturalness compared to TTS on the one hand and natural speech with "correct" or adequate prosody on the other. Thereby, we evaluate how strong the influence of such inadequate prosody is with the aim to draw attention to the role of (in-)adequate prosody when using off-the-shelf TTS in ECA research. Our test bed comprises four social contexts representing everyday situations, e.g., making a doctor's appointment, in which two ECAs engage in a four-sentence conversation of about 30 s length. Furthermore, we examine whether seeing virtual representations of the ECAs acting out this speech acts as a moderator and influences the expectations, and thus the ratings, of naturalness. We expect to find that synthetic speech will be more readily accepted within a virtual environment, since the combination may be felt as matching (cf. Gong and Nass [2007], who found synthetic speech to be best presented with a synthetic face). We call this the *masking effect of synthetic speech* by speaker embodiment. Moreover, we anticipate that the use of embodied ECAs also influences how severely inadequate prosody is assessed. We call this the *masking effect of prosody* by speaker embodiment. Furthermore, we expect the female voice to be judged as more natural in synthetic speech, since most smart speakers nowadays use female synthetic voices [West et al., 2019] and therefore participants are more accustomed to those producing incorrect prosody. To the best of our knowledge no study has been conducted before evaluating this isolated effect of inadequate prosody in TTS in combination with speaker embodiment. Although - as stated by Peeters [2019] - doing such research directly in virtual reality will increase ecological validity, we had to restrict the presented study to a video-based online survey due to the limitations resulting from the ongoing Corona pandemic.

We designed a study varying the *(S)peech* in three levels: synthetic speech as generated by a TTS system (S_{TTS}), speech recorded by a voice actor imitating the less adequate prosody as present in the synthetic stimuli ($S_{\text{human+TTS}}$), and human speech with adequate prosody (S_{human}). We also varied the *(E)mbodiment* of the speakers on two levels between audio-only (E_{audio}) and simultaneously watching ECAs acting out the speech (E_{ECA}) in an audio-visual condition. For our conversations, we used both female and male virtual interlocutors (*(G)ender* with the levels G_{female} and G_{male}).

We test the following hypotheses with respect to perceived naturalness (N):

H1 We expect participants to rate (i) a human voice as more natural than a synthetic voice (even if the prosody is inadequate) and to rate (ii) adequate prosody as more natural than inadequate prosody:
$N(S_{\text{human}}) > N(S_{\text{human+TTS}}) > N(S_{\text{TTS}})$.

H2 We expect that watching the ECAs speaking will increase the perceived naturalness of the synthetic speech:
$N(E_{\text{ECA}}) > N(E_{\text{audio}})$ for S_{TTS}.

H3 We expect participants to perceive the female voice as more natural in synthetic speech:
$N(G_{\text{female}}) > N(G_{\text{male}})$ for S_{TTS}

4.1.1. Evaluation of the Influence of Synthetic Voices and Embodiment

We designed a 3 × 2 within-subject study, comparing the three different levels of *Speech* and the two levels of *Embodiment* of the speakers.

Materials

We designed four dialogues between a woman and a man consisting of four sentences per dialogue portraying a short telephone call in German of about 30 s each. This allowed us to place the participants as passive observers between the interlocutors. The *Scenarios* were designed to represent everyday situations like making a doctor's appointment ($S1_{\text{doctor}}$), organizing a board game night with friends ($S2_{\text{gaming}}$), booking a flight ($S3_{\text{travel}}$), or organizing the next football training ($S4_{\text{training}}$). The dialogue for scenario $S1_{\text{doctor}}$ is given in Table 4.1, with the accented syllables in bold face and the *nuclear* accent in bold capitals (see Table A.2 and Table A.3 in the Appendix for the other scenarios). A nuclear accent is the final pitch accent in an utterance which determines the interpretation or pragmatic meaning of the utterance. Table 4.1 shows a distribution of accents representing a possible adequate prosody. The adequacy of prosody (especially in terms of accent placement, which is of major interest in our study) was checked in a brief informal survey prior to the experiment. Additionally, the prosody as produced by the TTS system is given, with inadequate accents in red. Since the experiment is conducted in German, we also provide an English translation, however, not specifying the accents since they are language-dependent.

We tested different commercial TTS engines and decided in favour of *Google Cloud TTS* using the voices *de-DE-Wavenet-F* as female and *de-DE-Wavenet-B* as male voice since they yielded the audibly most pleasing results while generating on average 2.5 misplaced nuclear accents per dialogue. In the last sentence of the example in Table 4.1, e.g., the TTS engine placed

Table 4.1.: Conversation in the first scenario (**S1**$_{\textbf{doctor}}$) given by a male ECA (A) and a female ECA (B). Accented syllables are written in bold face and the nuclear accent in bold capitals. The *adequate* prosody was used for S_{human} whereas *TTS prosody* was used for $S_{\text{human+TTS}}$ as well as S_{TTS}. For the latter, inadequate nuclear accents are highlighted in red. An English translation of the text is given in the right-hand column. The other scenarios can be found in the Appendix A.2.1.

S1	**German (adequate prosody)**	**German (TTS prosody)**	**English translation**
A	Guten **Tag**, ich **möch**te gerne einen Ter**min** für eine Kon**TROLL**untersuchung vereinbaren.	**Gu**ten **Tag**, ich möchte gerne einen Ter**min** für eine Kon**troll**untersuchung ver**EIN**baren.	Good morning, I'd like to make an appointment for a check-up.
B	Sehr **ger**ne, aber in **die**sem Monat kann ich Ihnen leider keinen Termin mehr **AN**bieten. Wir sind bereits **VOLL**.	**Sehr** **ger**ne, aber in **die**sem **Mo**nat kann ich Ihnen **lei**der keinen Ter**MIN** mehr anbieten. Wir **sind** bereits **VOLL**.	Very well, but unfortunately I can't offer you any more appointments this month. We are fully booked already.
A	**Scha**de. Wie **sieht** es denn im **FEB**ruar terminlich aus?	**Scha**de. **Wie** sieht es denn im **Feb**ruar ter**MIN**lich aus?	Too bad. What about the schedule for February?
B	**Gut**, hier **sind** noch einige Termine **FREI**. Sie **könn**ten zum Beispiel am **neun**ten Februar um **neun** **UHR** vorbeikommen.	**Gut**, **hier** sind noch **ei**nige Ter**MI**ne frei. Sie **könn**ten zum **Bei**spiel am **neun**ten **Feb**ruar um **neun** Uhr **VOR**beikommen.	It looks good, here we still have some free dates. For example, you could come by on the ninth of February at nine o'clock.

the nuclear accent on the first syllable of the final verb (*VORbeikommen*, 'come by'), representing both a wrong position of lexical stress (which should be on the second syllable, i.e., *vorBEIkommen*) as well as an inappropriate position of the nuclear pitch accent (which should be on the noun *Uhr* 'clock', as in the left-hand column of Table 4.1). These stimuli were used for the S_{TTS} level.

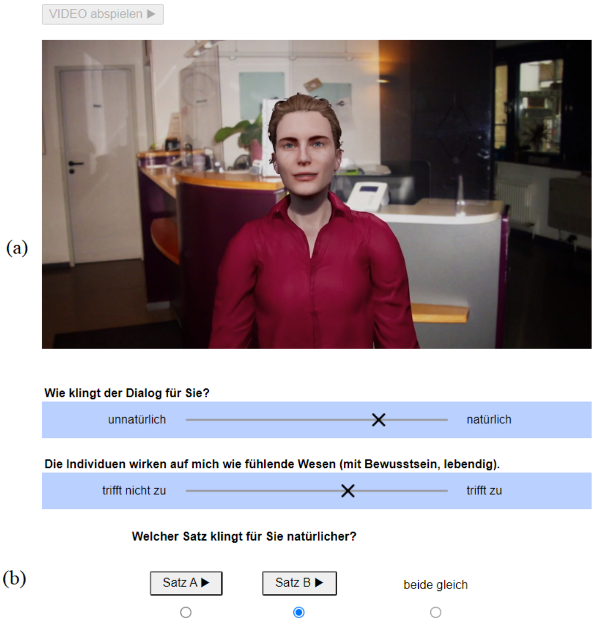
Additionally, we recorded a trained 36-year-old female speaker and a trained 51-year-old male speaker with an *AKG C451E* microphone (with CK4 Capsule) at around 50 cm distance to the speaker (see Fig. 5.3) in an acoustically optimized recording room (reverberation time $T_{30} <$ 200 ms) reading out the dialogue once with adequate prosody (S_{human}) and once imitating the prosody as produced by the TTS engine ($S_{\text{human+TTS}}$). For the imitation, the actors listened to the sentences produced by the TTS engine a few times and then spoke along with it.

While recording audio, we also captured the facial movements of the speakers to animate the respective ECAs during rendering. We used an iPhone to record face animations in 100 Hz as described in Sec. 5.1 (see Fig. 5.3). Since the sentences for $S_{\text{human+TTS}}$ were spoken in sync with the audio of S_{TTS}, we were able to use the face tracking for both conditions. By this process we minimized any qualitative visual differences between the speech conditions.

The audio for both *Embodiment* levels was processed with the *Virtual Acoustics*[1] framework, to generate a binaural signal of the virtual sound source approximately 70 cm away from the listener. A static artificial reverberation was added approximating the reverberation in a medium-sized room ($V = 56\,\mathrm{m}^3$, $T_{30} \approx 430\,\mathrm{ms}$).

For E_{ECA} we used two human models generated with Reallusion's *Character Creator 3* (see Fig. 4.1). The models were rendered in *Unreal Engine 4.22* in front of a static background and lit according to lights estimated from the background. For the conversations we tried to convey the impression of a hands-free phone call, using cuts between the frontal perspectives as depicted side-by-side in Fig. 4.1 and App. A.2.2. We decided to use this presentation since we assumed that this kind of cut sequences should be known from movies and allowed participants to listen to the agents from a frontal direction.

Procedure



Wie klingt der Dialog für Sie?

Die Individuen wirken auf mich wie fühlende Wesen (mit Bewusstsein, lebendig).

Welcher Satz klingt für Sie natürlicher?

Figure 4.2.: Screenshots of the study forms: (a) first part with the video stimulus currently playing and both visual analogue scales filled in; (b) form of the second part, with two buttons playing one stimulus each and radio buttons to select either stimulus or *both equally* (German: *beide gleich*)

The study was conducted as an online questionnaire realised using the *SoSci Survey* platform [Leiner, 2021] and made available to participants at www.soscisurvey.de. The study consisted of two parts with two different tasks. In the first part, participants had to rate the naturalness of 24 stimuli (3 *Speech* conditions × 2 *Embodiment* conditions × 4 *Scenarios*). The evaluation was carried out for each stimulus on a separate page. According to the *Embodiment* condition, 12 stimuli were presented as audio-only and 12 stimuli as video. Participants were able to control when to start a stimulus but it could only be played once. Each stimulus was rated on two visual analogue scales (VASs) to evaluate two aspects of naturalness (see Fig. 4.2(a)). The first scale was used to directly collect (N)aturalness ratings, answering the question "How does the dialogue sound to you?" (German: *Wie klingt der Dialog für Sie?*). Participants provided the judgements by placing a roll bar on the continuous horizontal scale (VAS) with the left pole labelled *"unnatural"* and the right pole labelled *"natural"*. The second scale was used to collect (A)liveness ratings. Participants had to judge to what extent the statement

[1] http://www.virtualacoustics.org/

"The individuals appear to be sentient (conscious and alive) to me" (German: *Die Individuen wirken auf mich wie fühlende Wesen (mit Bewusstsein, lebendig). "does not apply"* (left pole) or *"does apply"* (right pole) to them. This question is one of five items of the Social Presence Survey [Bailenson et al., 2001], connecting the naturalness here to this well-established measure. The responses on both scales were encoded as interval data ranging from 0 (left pole) to 100 (right pole). Hence, the higher the ratings or values the higher the degree of perceived naturalness/aliveness. The stimuli were presented in randomized order for each participant.

After finishing this part of the questionnaire, an intermediate questionnaire asked the following questions in random order:

1) "What aspects did you in particular focus on in the videos?":
multiple choice for *speech*, *individuals*, *lipsync*, *gaze*, *environment*, *other*

2) "Would you want to interact directly with one or both individuals?":
VAS from *"No, not at all"*(0) to *"Yes, absolutely"*(100)

3) "Which of the two Individuals would you prefer to interact with?":
single choice for *male*, *female*, *both equally*

4) "Would you prefer to see the individuals talking instead of just hearing them?":
VAS from *"No, not at all"*(0) to *"Yes, absolutely"*(100)

5) "Which version of the dialogue was easier to follow?":
single choice for *video*, *audio-only*, *both equally*

In the second part of the study, participants had to make forced choices, i.e., they had to choose which of two audio stimuli sounded more natural to them (German: *Welcher Satz klingt für Sie natürlicher?*). Therefore, participants were able to listen to each stimulus as often as necessary by clicking on it. After having listened to both stimuli at least once, participants had to pick either one stimulus or choose *"both equally"* (see Fig. 4.2(b)). The stimuli used were individual sentences from the first part of the study. Participants had to rate six pairs, in which the same sentence was spoken with a different *Speech* level (i.e., S_{human} vs. $S_{\text{human+TTS}}$, S_{human} vs. S_{TTS}, and $S_{\text{human+TTS}}$ vs. S_{TTS}) by both speakers (or *Genders*). Additionally, three pairs with identical *Speech* level were used to compare the naturalness of the speakers' *Gender* (G_{female} vs. G_{male}). Since we had not recorded the same sentence spoken by both voice actors, we used sentences with similar length. Finally, we added four filler pairs comparing identical stimuli (two filler stimuli for each speaker), to identify insufficiently attentive participants. Participants failing to rate fillers more than once with *"both equally"* were excluded from the analysis. Hence, in total 13 sentence pairs had to be rated: 6 comparisons with different *Speech* level (but same *Gender*) + 3 comparisons with identical *Speech* level (but mixed *Gender*) + 4 filler sentences. Ratings for the 9 (non-filler) sentence pairs were part of the analysis.

The procedure of the study was as follows: After reading a description about the content and purpose of the study, participants were asked to use regular stereo headphones and conduct an audio calibration using a sequence of TTS samples of numbers and letters. In this sequence, the participants had to adjust the audio volume so that only the numbers were comprehensible, without understanding the less loud letters in between. This created comparable hearing conditions for all participants, independent of hardware and potential background noise. In very quiet environments, the calibration led to a minimum playback volume of around 50 dBA.

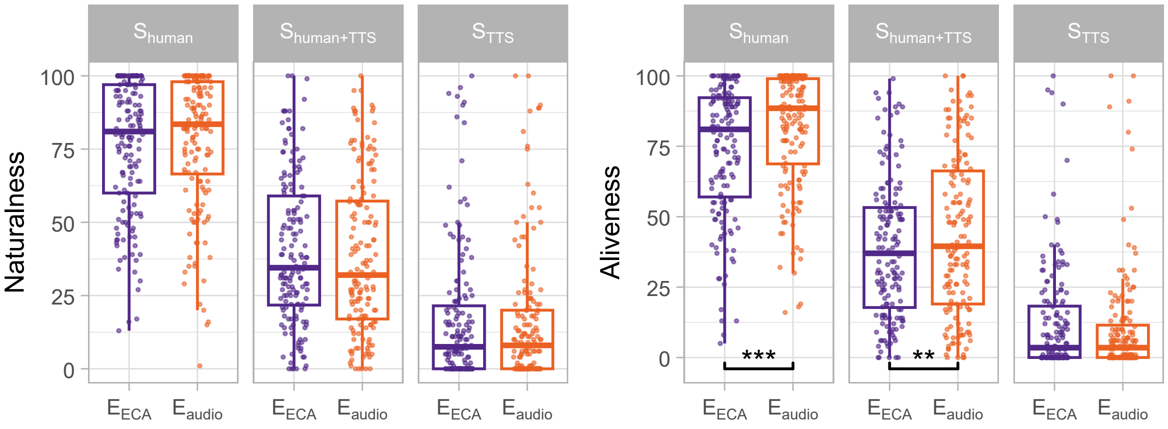
Next, example exercises for both parts were shown, so participants were aware of the procedure before being asked to give informed consent and filling in a demographics questionnaire. The remainder of the study was split in two parts. Both parts began with three warm-up conditions (taken from the study conditions), so participants got familiar with the controls of the exercise and were also introduced to the entire range of the stimuli. The study ended with two free-answer fields asking for suggestions to improve the naturalness of the dialogues and asking for general feedback.

4.1.1.1. Results

Fourty native speakers of German took part in the experiment, which were primarily recruited via university mailing lists. One participant rated more than one of the filler sentences not with *"both equally"* and was therefore excluded from the analyses. Eight of the remaining participants answered one of the fillers incorrectly, however, those participants were kept for the evaluation. The remaining 39 participants (25 female) had a mean age of 30.3 years (standard deviation (SD) = 13.4 years), and all of them reported normal hearing and normal or corrected vision. Twelve of the participants reported to have at least a basic knowledge of linguistics (one of them reported being advanced). Furthermore, six of the participants grew up in a bilingual environment. Participants took between 24 to 38 minutes to complete the entire study.

For the statistical analysis, we performed linear mixed-effects models and generalized linear mixed-effects models by using the *lmer()* and *glmer()* functions from the "lme4" package [Bates et al., 2015] for R [R-Core-Team, 2015]. Linear mixed-effects models were calculated to test for statistical significance of the naturalness (N) and aliveness (A) ratings on the visual analogue scales in the first part of the experiment. The models included *Speech* (S_{human}, $S_{\text{human+TTS}}$, S_{TTS}), *Embodiment* (E_{audio}, E_{ECA}), and *Scenario* ($S1_{\text{doctor}}$, $S2_{\text{gaming}}$, $S3_{\text{travel}}$, $S4_{\text{training}}$) as fixed factors and assume random intercepts and slopes for *Speech* by participants. Generalized linear mixed-effects models were performed for the statistical analysis of the distributions of naturalness choices between (i) different *Speech* levels and (ii) different speakers (*Gender*: G_{female}, G_{male}) in the second part of the experiment. The models included the type of *Comparison* (i.e., either between (i) different or (ii) identical *Speech* levels) as fixed factor and also assume random intercepts and slopes for *Comparison* by participants. We additionally tested all models against a model with the same random effect structure including *Participant Gender* (P_{female}, P_{male}) as another fixed effect. Correlations were computed using Pearson correlation coefficients.

Overall results of the first part of the experiment are depicted in Fig. 4.3 in terms of boxplots and individual data points of the naturalness and aliveness ratings split by *Speech* and *Em-*



(a)

(b)

Figure 4.3.: Boxplots of the ratings of (N)aturalness (a) and (A)liveness (b) (on a scale from 0 to 100), split by *Speech* and *Embodiment*. Boxes indicate quartiles with whiskers at full range, excluding outliers. Additionally, all individual data points are shown. Differences between all *Speech* levels were significant ($p < .001$), other significances are shown, $*** \ p < .001$, $** \ p < .01$.

bodiment. The figure shows that dialogues with adequate prosody spoken by a human voice (S_{human}) are clearly perceived as natural and alive, while the perceived naturalness and aliveness strongly decreases for dialogues with inadequate prosody ($S_{\text{human+TTS}}$, S_{TTS}). However, in the latter conditions human voices are still perceived as more natural and alive (medium scores) than synthetic voices which received the lowest scores.

In the following, we will first report the effects registered by the statistical analyses of the naturalness (N) and aliveness (A) ratings that are significant by the $|t| > 2$ criterion (corresponding to the established significance level of $p < .05$, cf. [Baayen et al., 2008]). Subsequently, we will report for both rating scales significant contrasts based on pairwise comparisons that exhibit a significance level of at least $p < .001$ unless otherwise specified. Statistical analyses of the naturalness (N) ratings (936 observations) register significant effects of *Speech* $[\chi^2 = 121.15, p < .001]$ and *Scenario* $[\chi^2 = 10.58, p > .001]$ as well as of the interactions *Speech:Scenario* $[\chi^2 = 4.47, p < .001]$ and *Speech:Embodiment:Scenario* $[\chi^2 = 2.61, p < .05]$. Likewise, statistical analyses of the aliveness (A) ratings (936 observations) register significant effects of *Speech* $[\chi^2 = 169.49, p < .001]$ and *Scenario* $[\chi^2 = 6.85, p < .001]$ as well as of the interaction *Speech:Scenario* $[\chi^2 = 3.2628, p < .01]$. Furthermore, aliveness ratings additionally reveal significant effects of *Embodiment* $[\chi^2 = 13.30, p < .001]$ and of the interaction *Speech:Embodiment* $[\chi^2 = 8.54, p < .001]$. Likelihood ratio tests comparing the presented models with a model including *Participant Gender* as another fixed factor revealed no significant effects (N: $\chi^2 = 15.55, p = .9$; A: $\chi^2 = 9.54, p = .99$).

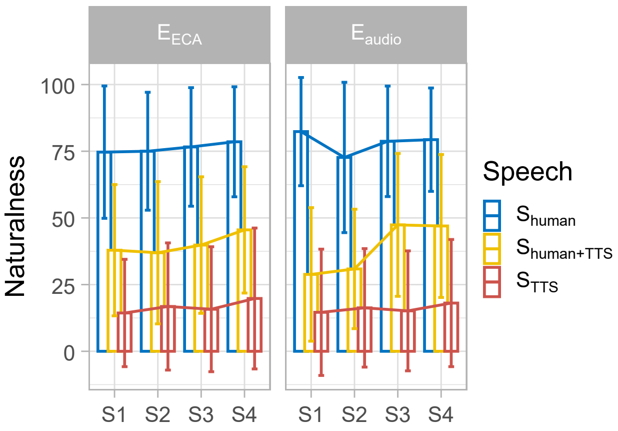


Figure 4.4.: Mean and standard deviation of the ratings for Naturalness (on a scale from 0 to 100) for each *Scenario* ($S1_{\text{doctor}}$, $S2_{\text{gaming}}$, $S3_{\text{travel}}$, $S4_{\text{training}}$), split by *Embodiment* and shown per *Speech* level.

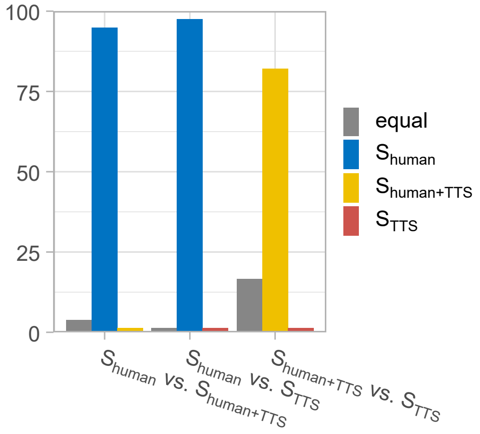


Figure 4.5.: Percentage of audio sample rated as more or equally natural in the second part of the study when comparing different *Speech* levels.

Pairwise comparisons of the effect of the *Speech* levels confirm a significant decrease in the perception of naturalness and aliveness from S_{human} to $S_{\text{human+TTS}}$ to S_{TTS}. Accordingly, we found N and A to be strongly correlated, $r(934) = .85, p < .001$. Further pairwise comparisons reveal that dialogues of scenario $S4_{\text{training}}$ are in general rated significantly more natural and alive than dialogues of scenario $S1_{\text{doctor}}$ and $S2_{\text{gaming}}$ (cf. Fig. 4.4). For dialogues with the $S_{\text{human+TTS}}$ *Speech* level the N and A ratings of scenario $S3_{\text{travel}}$ are also significantly higher than the ratings for scenarios $S1_{\text{doctor}}$ and $S2_{\text{gaming}}$. Moreover, for the naturalness ratings only, these differences between scenarios $N(S1_{\text{doctor}})$ and $N(S2_{\text{gaming}})$ vs. $N(S3_{\text{travel}})$, $N(S4_{\text{training}})$ are enhanced in the $N(E_{\text{audio}})$ condition. Further effects of *Embodiment* are registered for the aliveness ratings: Pairwise comparisons reveal that dialogues presented as audio-only ($A(E_{\text{audio}})$) are in general rated as more alive than dialogues presented as video ($A(E_{\text{ECA}})$). This effect is enhanced in the conditions with human voices ($A(S_{\text{human}})$: $p < .001$; $A(S_{\text{human+TTS}})$, $p < .01$).

Results of the second part of the experiment are depicted in Figs. 4.5 and 4.6. Fig. 4.5 shows the percentages of audio samples rated as more natural when comparing different *Speech* levels with each other. Listeners reliably chose utterances spoken by a human voice (S_{human}, $S_{\text{human+TTS}}$) as the more natural variant in all conditions. More precisely, a human utterance with adequate prosody (S_{human}) was preferentially selected (over 96%) whenever available. A human utterance with inadequate prosody ($S_{\text{human+TTS}}$) was only rated as more natural in comparison with a synthetic utterance (S_{TTS}). However, in the latter comparisons 17.1% of the cases were also rated as being equally natural. A likelihood ratio test comparing the generalized linear mixed-effects model including the type of *Comparison* (between different *Speech* levels: 234 observations) as fixed factor with a null model having the same random effect structure (see above) revealed that *Comparison* had a significant effect on whether the more or less natural variant was perceived as more or equally natural ($\chi^2 = 11.98, p < .01$). A further model comparison including *Participant Gender* as another fixed factor revealed no significant effect.

With respect to the comparison of the naturalness of different speakers (*Gender*: G_{female}, G_{male}), listeners' choices were overall less clear and more ambiguous when comparing the identical *Speech* levels with each other. In the S_{human} and S_{TTS} conditions listeners perceived both voices as equally natural in 59% of the cases. If listeners decided between the female and male voice, the female voice is more often rated as more natural in the S_{human} condition ($G_{\text{female}} = 33.3\%$ vs. $G_{\text{male}} = 7.7\%$), while the male voice is more often rated as more natural in the S_{TTS} condition ($G_{\text{female}} = 12.8\%$ vs. $G_{\text{male}} = 28.2\%$). In the $S_{\text{human+TTS}}$ condition, the ratings are quite balanced, although the male voice is most often rated as more natural ($\mathit{equal} = 35.9\%, G_{\text{female}} = 25,6\%, G_{\text{male}} = 38.5\%$). A likelihood ratio test comparing the generalized linear mixed-effects model including the type of *Comparison* (between the identical *Speech* levels: 117 observations) as fixed factor with a null model having the same random effect structure (see above) revealed that *Comparison* or rather the *Speech* level had no significant effect on whether the male or female speaker was perceived as more or equally natural ($\chi^2 = 5.29, p = .71$). However, a further model comparison including *Participant Gender* as another fixed factor revealed a significant effect ($\chi^2 = 5.33, p < .05$). Fig. 4.6 shows the percentages of audio samples (different speakers (*Gender*: G_{female}, G_{male})) rated as more or equally natural when comparing the same *Speech* levels with each other split by the participants' gender (*Participant Gender*: P_{female}, P_{male}). The graph resembles the overall results for the different choices by the two participant groups: Female listeners more often rated both voices as equally natural ($P_{\text{female}} = 58.7\%$ vs. $P_{\text{male}} = 38.1\%$). With respect to the speakers' gender, female listeners judge the male voice more often as more natural ($G_{\text{female}} = 16\%$ vs. $G_{\text{male}} = 25.3\%$), while male listeners judge the female voice more often as more natural ($G_{\text{female}} = 38.1\%$ vs. $G_{\text{male}} = 23.8\%$).

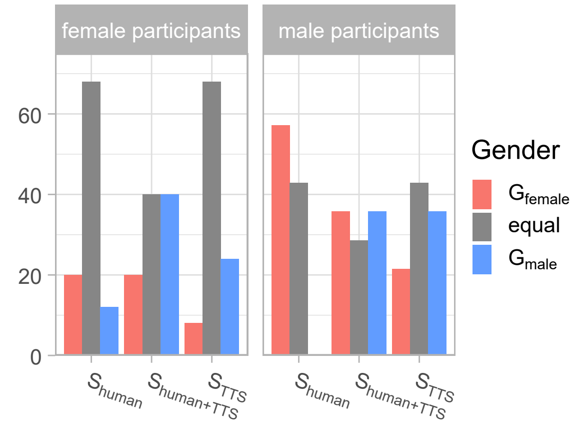


Figure 4.6.: Percentage of audio sample rated as more or equally natural in the second part of the study. The data is shown per *Speech* level and split between female and male participants.

The intermediate questions asked after the first part of the study revealed how many participants paid attention to the following aspects in the videos: *speech* (89.7%), *lipsync* (43.6%), *gaze* (28.2%), the *individuals* (25.6%), the *environment* (15.4%). 7.7% of the participants specifically mentioned that they focused on intonation and pronunciation, albeit not given as default answer option. Furthermore, the data shows interesting differences in attention levels between participant gender for *lipsync* ($P_{\text{female}} = 56.0\%$, $P_{\text{male}} = 21.4\%$) and *gaze* ($P_{\text{female}} = 36.0\%$, $P_{\text{male}} = 14.3\%$), while 100% of the male participants stated that they paid attention to *speech*. In general, the participants had diverse opinions on whether they would want to interact directly with one or both individuals (VAS mean = 46.9, SD = 28.4) and showed no clear preference for an interaction with one of the individuals (*both equally* = 48.8%, G_{female} = 25.6%, G_{male} = 25.6%). Furthermore, the VAS answers correlated with their ratings of naturalness, $r(934) = .14, p < .001$, and aliveness, $r(934) = .17, p < .001$. Similarly, participants did not show a clear tendency on whether they would prefer to see the individuals talking instead of just hearing them (VAS mean = 45.6, SD = 32.8). Accordingly, most participants answered that they were able to follow both versions of a dialogue equally well (71.8%), showing a slight preference for the audio-only (20.5%) version (vs. 7.7% for video).

4.1.1.2. Discussion

Our first hypothesis **H1** can be confirmed: Participants in general rated (i) a human voice as more natural and alive than a synthetic voice and (ii) adequate prosody as more natural and alive than inadequate prosody. This also means, that inadequate prosody triggered lower naturalness and aliveness ratings for the human voice: $N(S_{\text{human}}) > N(S_{\text{human+TTS}}) > N(S_{\text{TTS}})$. It is interesting to see that $S_{\text{human+TTS}}$ (average $N(S_{\text{human+TTS}}) \approx 39.3$) has been judged to be closer to S_{TTS} (average $N(S_{\text{TTS}}) \approx 16.4$) than to S_{human} (average $N(S_{\text{human}}) \approx 77.2$). While this effect might be considered an artefact of participants not being accustomed to using ECAs, it is also reflected in the direct comparisons between the *Speech* levels (cf. Fig. 4.5). Here, the two levels with inadequate prosody ($S_{\text{human+TTS}}$ vs. S_{TTS}) were rated significantly more often as equally natural than in the comparisons with the prosodically adequate condition (S_{human}). This finding indicates a strong (unfavourable) impact of suboptimal prosody on perceived naturalness in general.

Our second hypothesis **H2** cannot be confirmed since there was no significant difference in naturalness and aliveness ratings for synthetic speech due to the level of *Embodiment*. This means that neither a masking effect for synthetic voice nor for inadequate prosody based on lowered expectations towards ECAs was found in our study. However, and surprisingly, only *listening* to dialogues spoken by human voices (instead of also watching the ECAs speaking) significantly *increased* the perceived level of aliveness. This might be due to the fact that the perception of aliveness of the human voice is potentially masked by seeing an ECA reenacting this speech. This aspect was also emphasized by a participant's free comment ("If you *see* an avatar you obviously cannot rate it very human-like anymore"). Since the measure of rating the individuals as more "conscious and alive" was taken from the Social Presence Survey [Bailenson et al., 2001] this result is somewhat surprising since normally it is expected that social presence increases with better visualization (i.e., photographic, anthropomorphic and behavioral realism, cf. [Oh et al., 2018]). Additionally, it needs to be conceded that the movements (gesture, lipsync, etc.) of the ECAs, although being recorded from real humans, may not be perfectly convincing. This was also the reason why we decided to not use photo-realistic rendering for the ECAs to avoid mismatches in rendering and behavioral realism [Kätsyri et al., 2015].

Furthermore, there was a significant difference of the naturalness ratings between $S1_{\text{doctor}}$ and $S2_{\text{gaming}}$ compared to $S3_{\text{travel}}$ and $S4_{\text{training}}$, especially due to the clearly diverging judgements in the $S_{\text{human+TTS}}$ condition. In a post hoc analysis we found that those scenarios which received lower ratings show more severe misplacement of accents: In $S1_{\text{doctor}}$, as mentioned above, the combination of a wrong position of a lexical stress *and* of the nuclear accent (on *VORbeikommen*, 'come by') was probably felt as the most serious mistake, closely followed by the wrong nuclear accent placement on the adverb *SIcherheitshalber* ('in order to be on the safe side') instead of *ERwin* in $S2_{\text{gaming}}$ (see Table 4.1, and Table A.2 and Table A.3 in the appendix). A nuclear accent on an adverb is generally rare in German, at least, if it is not used contrastively. The difference in naturalness ratings between the four scenarios is again more pronounced in the audio-only condition. This finding is in line with our expectation that seeing an artificial agent speaking will affect what participants expect from them with regard to speech quality.

Our third hypothesis **H3** cannot be confirmed either: Participants perceived the female voice as least natural in synthetic speech. This could potentially originate from the same rationale mentioned earlier, namely that female synthetic voices are more common. One participant stated that "machine voices that you already know from navigation systems or platform announcements sound very unnatural due to previous experiences with those." Two participants even stated that they already knew this specific synthetic voice from *Siri*. It is, however, interesting to note that male participants were in general less inclined to rate both voices as equally natural, preferring the natural female and the male synthetic voice.

When asked for suggestions for improvement on the presented dialogues, six participants proposed to formulate them in a more natural manner and ten participants noted that they should be performed with a faster pace and should be pronounced less precisely (five participants). Still, a quarter of the participants (10) rated the intermediate question whether they "want to interact with the individuals" with a mark of 66% or higher. While the variance of the answers to this question was very high (11 participants rated it lower than 33%), it is interesting to see that the answers to this question correlate with the answers given to the naturalness and aliveness questions by the same participants. This outcome suggests that a high degree of naturalness is important to facilitate interactions with conversational agents.

With this study we aim to raise awareness of the negative effect of inadequate prosody when utilizing synthetic speech for ECA research. Extending the observations by Seaborn et al. [2021] that synthetic and human voices are not yet on a par, we added one previously neglected, yet important dimension for this discrepancy, namely inadequate prosody which is quite common in synthesized utterances. While our work did not focus on manually fine-tuning the off-the-shelf synthesis to produce adequate prosody for the same sentences, the difference in naturalness ratings of the human speech (S_{human} vs. $S_{\text{human+TTS}}$) and the differences between the scenarios with different severeness of accent misplacement ($S1_{\text{doctor}}$ and $S2_{\text{gaming}}$ vs. $S3_{\text{travel}}$ and $S4_{\text{training}}$) indicate that naturalness of speech is strongly decreased by inadequate prosody. To this end, we strongly recommend practitioners to pay close attention to inadequate prosody when it comes to ECAs' speech. In case of using natural speech, despite the labor- and cost-intensity, trained native speakers may produce the best results. In case of using synthetic speech, practitioners should try to reduced or even omit inadequate prosody by either reformulating sentences to achieve a more suited prosody or by manually fine-tuning the synthesis results. This recommendation is especially true since a decrease in naturalness due to inadequate prosody potentially decreases the desired effect when embedding ECAs, for example, in training as mentioned in the introduction. Therefore, we see at the current state of the art recorded speech as the best option to elicit the highest possible social presence. With the current advancement in speech synthesis, especially for speech sounding more like spontaneous speech and less like a radio moderator reading a text, synthetic speech could become a viable option. However, in this case a special focus in pre-testing should be to evaluate whether the produced prosody sounds natural.

Nevertheless, these findings would need to be reproduced using ECAs interacting with the participants in virtual reality since this would also improve the ecological validity of the linguistic results as stated by Peeters [2019]. A step towards this goal was taken by [Higgins et al., 2022], who found a difference of social presence rating only in their video-based experiment but not in the VR experiment they conducted before. They attributed this to the artificially-synthesized but equally expressive speech material used in the VR study compared to TTS in the video-based one. But maybe also prosody was mediating these effects. Furthermore, while we conducted the study in German due to the higher availability of native speakers, it would be interesting to reproduce it in English since the German synthetic voices are potentially inferior to the English ones, the latter being further developed.

4.2. Auralization

While the previous section focuses on the characteristics of the speech signal, particularly on prosody and the nuances of both synthetic and human voices, it is equally important to consider appropriate auralization techniques [Kleiner et al., 1993] for the voices of ECAs. These techniques can significantly contribute to creating a more immersive and realistic interaction experience. Auralization describes how the sound of different virtual sound sources is presented to a human listener who is immersed into a virtual audio scene, incorporating sound generation, transport and reception. The human auditory system and the generation of appropriate auralizations have been a research subject for decades and are investigated extensively (e.g., [Blauert, 1997; Pausch et al., 2018]). Research indicates, that presenting speech sound and potentially other sound sources to users in an IVE increases the presence, i.e., the feeling of being physically present within this IVE [Poeschl et al., 2013; Nordahl and Nilsson, 2014]. However, this sound has to be appropriate for the situation and must be presented in a plausible way [Dicke et al., 2010]. Therefore, often panning or ambisonics approaches are implemented [Gerzon, 1985] which, however, lack correct distance perception cues, especially for close-by sound sources [Pelzer et al., 2011]. To that end, binaural auralization approaches [Lentz, 2008] can be used that produce individual signals for each ear incorporating the characteristics of the listener's head, the so called head-related transfer function (HRTF)(see [Begault, 2000; Vorländer, 2008; Steadman et al., 2019; Geronazzo et al., 2018]). Thereby, the natural hearing experience is simulated and thus humans are able to localize spatially distributed virtual sound sources. However, Immohr et al. [2024] found binaural to be preferred in direct comparison but surprisingly did not find significant differences in perceived social presence or plausibility. Nevertheless, the subjective impression of an acoustically responsive environment is modeled by simulation that applies acoustic phenomena during propagation, and integrates reflections at surface boundaries [Lentz et al., 2007]. Especially in the context of IVEs, these simulation methods are based on Geometrical Acoustics, where both sound source and receiver are considered infinitesimally small [Vorländer, 2008]. Point sources radially emit a signal from their location into space and are considered omnidirectional, since the orientation of the source is not taken into account (cf. [Mehra et al., 2014]). In order to maintain the direction-related spectral attenuation that every real-world sound source inherits, a directivity filter can be applied [Rindel et al., 2004]. Directivity filtering generates perceivable differences to the user, for example, a human speaker sounds clearest when talking directly towards the listener while sounding muffled and less loud when talking away from the listener. This directivity for a human speaker/singer can be measured as described by [Kob, 2002; Behler et al., 2012] and can then be simulated for virtual sound sources during binaural auralization.

Recent work, investigating moving sound sources with static directivities, report improvements, when accounting for the varying orientation (rotational movement) of the source [Vigeant et al., 2011; Ackermann et al., 2019]. Postma and Katz report significant differences in the room acoustics clarity and distance perception when presenting auralizations based on recordings that capture a singer's voice simultaneously at many locations and thereby naturally include directivity [Postma and Katz, 2016; Postma et al., 2017]. These findings therefore encourage the use of directivities in general. In the domain of music, Ackermann et al. [2019] added directivities to isolated dry recordings of instruments based on the musicians' motion, and provide evidence that listeners can reliably distinguish between static and moving auralizations. However, this directivity is often ignored when auralizing virtual humans. Shin et al. [2019] found that a more realistic 3D sound has a positive influence on social presence. While they used recorded 3D sound, which naturally incorporates the sound source directivity of the recorded sources in the examined static setup, this has to be actively simulated for dynamically moving ECAs. As Mehra et al. [2014] stated, sound source directivities improve the realism of the auralization and might thereby, if also applied to ECAs' speech, improve their acoustic realism and thus their perceived social presence.

Our research focus lies on the investigation of how changing the directivity of the speech sound sources affects the social presence and naturalness of ECAs (subconsciously) perceived by an immersed human user. To this end we conducted three studies. First we evaluate the influence of adding this measured directivity of a speaker in two studies presented in Sec. 4.2.1 and Sec. 4.2.2. This is further called *static directivity*, since the directivity filter is static over time, while, however, the orientation and position is dynamically adapted to the ECA's head. In a second step, we then also adapted the directivity filter based on the currently uttered speech content, coined *dynamic directivity*[2]. We present a third study in Sec. 4.2.3 evaluating the effect of this dynamic directivity for ECAs.

[2] This choice of words stands in contrast to related publications (e.g., [Postma and Katz, 2016; Postma et al., 2017; Ackermann et al., 2019]), where dynamic directivity is attributed to a dynamic/moving sound source rotation and not a time-variant dataset. We simulated sound source movement in all conditions, but wanted to highlight the difference in the directivity filters by usage of these terms.

4.2.1. Pilot Study on the Influence of Static Speech Directivity

The contents of this section are based on and taken in part from work previously published in [Wendt et al., 2018].

A within-subject pilot study was conducted in the five-sided AixCAVE (see App. A.5.1). It had two independent variables: *Auralization* and *Gender*. We used two different ECAs (*male* and *female*) that were auralized using either no directivity (*omnidirectional*) or using the directivity of a human speaker (*static directivity*) as provided by [Kob, 2002] (see Fig. 4.9). This led to four different conditions. The participants were immersed in a virtual stockroom with one ECA (see Fig. 4.7). Participants had to fulfill a search task. Therefore, the ECA uttered demands for specific items in various directions facing towards and away from the participant to exhibit the differences in the auralization techniques. We measured the perceived social presence of the ECAs using the social presence survey (SPS) [Bailenson et al., 2001] after each condition which the participants could directly answer virtually within the AixCAVE.

Experimental Design and Task

After reading a brief introduction and answering a demographic questionnaire, participants enter the AixCAVE and have the opportunity to familiarize with both ECAs side-by-side in an otherwise empty scene. The ECAs say a brief welcoming sentence for the participants to get accustomed to their speech. Subsequently, a virtual stockroom is displayed (see Fig. 4.7), which has exactly the size of the AixCAVE (see App. A.5.1), so the participants can navigate by means of physical walking. The ECA to be used in the respective condition is placed close to the middle of the room, so he/she does not stand directly in between two shelves and therefore gives the participant enough space to pass, avoiding collisions. The shelves are filled with several boxes and 18 items. This number was chosen so that the algorithm, explained below, has a sufficient number of items to pick from. The items are placed at well reachable places on the shelves, evenly spread around the ECA. The items are randomly swapped between conditions to avoid learning effects.

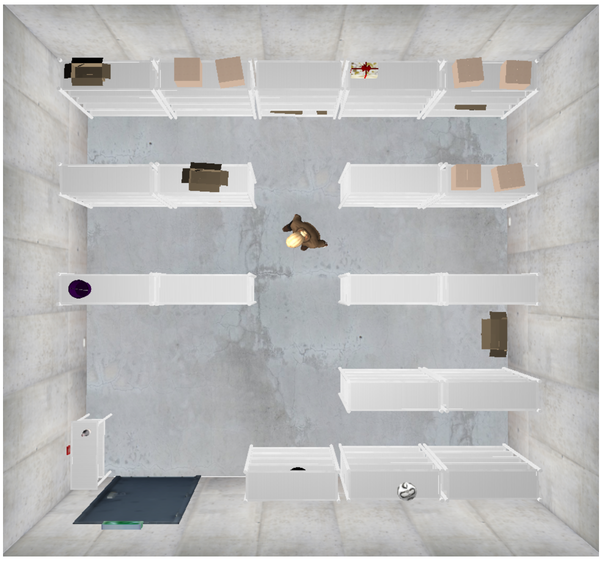


Figure 4.7.: Top view of the stockroom, with an ECA in the middle and 18 items placed on shelves. The items that are not on the top shelves are not visible.

The ECA utters a request for one particular item at a time. Therefore, the ECA first turns and looks towards the item they will ask for, before speaking a sentence like *"Please bring me the green basket"*. These sentences are predefined and differ slightly for each item, as recommended by Bönsch et al. [2017b]. By turning towards the item, differences in auralization become noticeable, since the sound does not change if the ECA and thereby the directed sound source is facing the participant for every utterance. Following this idea, the item that has to be picked up next is determined based on the angle

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

which is based of the positions of the item ($\mathbf{p}_{item}$), the participant ($\mathbf{p}_{user}$) and the ECA ($\mathbf{p}_{agent}$) projected onto the floor plane. Participants are asked to find 12 different items, that are chosen such that we have an equal number of cases with $\theta < 45°$ (front), $45° \leq \theta < 135°$ (side) and $135° \leq \theta$ (back), namely four items each. This way facing directions in all four quadrants around the ECA based on the respective participant position are used. Left and right are not distinguished, since they do not exhibit different sounds related to directivity, which is approximately symmetrical. To find the item to ask for next, we first randomly chose a facing direction and then pick one of the remaining items in that quadrant. In case no more items were present in this quadrant, a random one is chosen. Therefore, more items need to be present in the scene than should be picked.

The participants have to pick up the demanded item. Therefore, they have to walk towards it and use a grabbing metaphor with a pointing device (see App. A.5.1). With the item attached to the pointing device (see Fig. 4.8), participants have to walk back to the ECA and bring the item close enough to the ECA. Once the item's distance to the ECA in the floor plane is below 40 cm it disappears, the ECA faces the participant and randomly says one of the three predefined thank-you sentences. If not all 12 items have been picked up yet, the ECA waits 1.5 s and then turns towards another item and asks for it. This time is chosen such that the ECA has enough time to finish the thank-you sentence before they starts turning.

After finishing all 12 pick-ups the scene fades out and an empty scene (only a blue floor plane) with a questionnaire is displayed which the participants are asked to answer using the same pointing device. After each of the four conditions, participants are asked to answer the 5-item SPS questionnaire (see App. B.2). As soon as the fourth questionnaire is answered, the familiarization scene is displayed again and the participants are asked to answer a post-study questionnaire outside of the AixCAVE asking for preferences of specific conditions.

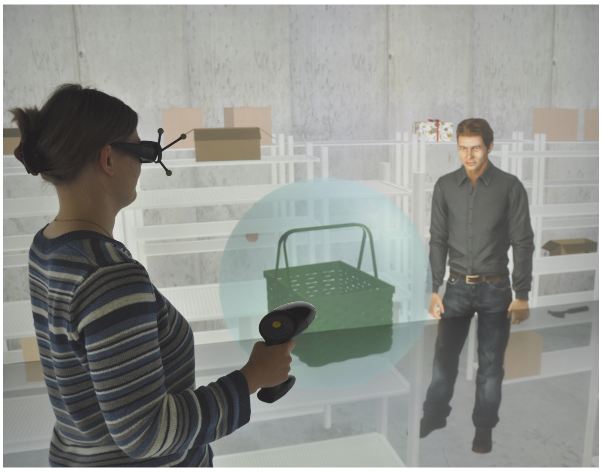
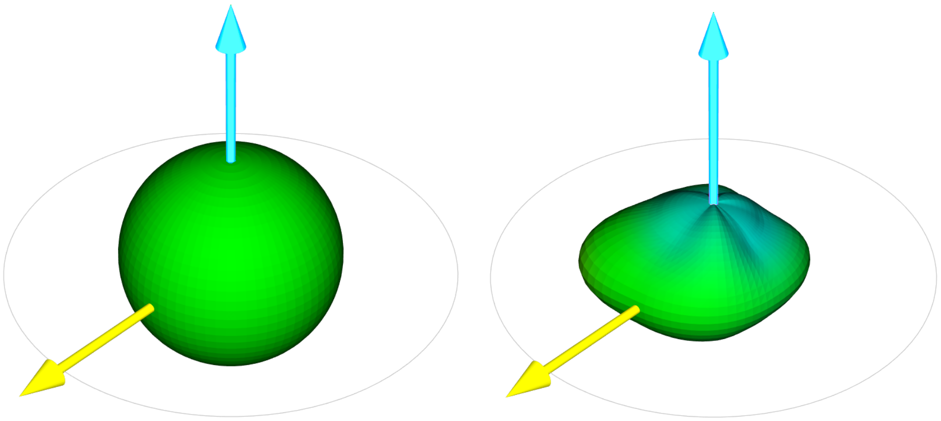


Figure 4.8.: A participant returning an item, which is attached to the pointing device, to the virtual agent.

The participants are equally distributed on the randomized sequences of conditions, to counter any order effects. However, both con-



(a) Omnidirectional (b) Human Singer Directivity

Figure 4.9.: The used directivities, shown here at 125 Hz with the yellow arrow pointing forward and the cyan one upward

ditions of one *Gender* are always done right after each other. This way, it is potentially easier for the participants to specify their preference for *Auralization* in the post-study questionnaire, as they experience the same voice with and without directivity right after each other. This leads in total to 8 different possible sequences. Additionally to the SPS questionnaires, the distance that the participants keep to the ECA is measured and the minimal distance per condition is stored as an objective measure (see Sec. 2.3). Therefore, the virtual stockroom is deliberately designed in a way, that the participants, who are asked to avoid collision with virtual objects, have to pass close by the ECAs (see Fig. 4.7). Thereby we hope to measure a correlation between the SPS and the minimal distance.

Study Material

The study is conducted in the AixCAVE (see App. A.5.1). To render and animate the ECAs, *SmartBody* [Shapiro, 2011] is used, from which the human models *Brad* and *Rachel* are utilized. *SmartBody* can also perform lip-syncing. Throughout the study the ECAs blinked periodically and performed an idle motion. The speech audio is produced using the text-to-speech engine *CereVoice*[3] [Aylett and Pidcock, 2007], which also produces synchronized viseme data (see Sec. 5.1.1) to animate the face. Although we stated in Sec. 4.1 that this is a potentially detrimental choice, TTS was used for cost and time efficiency.

Since the ECAs were, by design, often talking towards the wall, we also had to include the room response to the acoustical speech signals into the auralization. To that end, we created three image sound sources [Allen and Berkley, 1979] for the closest walls of the room with respect to the speaker sound source, which had the largest influence on the volume of the

[3] `https://www.cereproc.com/`

reflected speech. Adding also image sources for the floor, ceiling and the farthest wall, led to performance problems and was therefore omitted, since during listening tests their absence was not noticeable. The image sound source position $\mathbf{p}_{\text{image}}$ and direction $\vec{d}_{\text{image}}$ are computed by mirroring them at each of these walls at $\mathbf{p}_{\text{wall}}$ with normal $\vec{n}_{\text{wall}}$, as:

$$\mathbf{p}_{\text{image}} = \mathbf{p}_{\text{source}} - 2\frac{(\mathbf{p}_{\text{source}} - \mathbf{p}_{\text{wall}}) \cdot \vec{n}_{\text{wall}}}{|\vec{n}_{\text{wall}}|}\vec{n}_{\text{wall}}$$

$$\vec{d}_{\text{image}} = \vec{d}_{\text{source}} - 2\frac{\vec{d}_{\text{source}} \cdot \vec{n}_{\text{wall}}}{|\vec{n}_{\text{wall}}| \cdot |\vec{d}_{\text{source}}|}\vec{n}_{\text{wall}}$$

where $\mathbf{p}_{\text{source}}$ and $\vec{d}_{\text{source}}$ are the position and direction of the speech sound source, which is placed in the ECAs' mouth and moved with the animation of the head. Furthermore, these image sound sources use the same directivity pattern as the primary source since the patterns are symmetrical and therefore don't have to be mirrored. This approximates the early reflections within the room, the late reverberation does not need to be added since it naturally exists due to the echoic properties of the AixCAVE and the fact that the virtual stockroom has exactly the size of the AixCAVE. As directivity filter a measured directivity of a human singer [Lentz, 2008] was used (see Fig. 4.9(b)). Furthermore for the omnidirectional condition, a directivity was used that is uniform in all directions and frequencies (see Fig. 4.9(a)) and is normalized to have the same amplitude in the frontal direction as the singer's directivity. Using directional filters for both conditions guaranteed that the acoustic signals were processed equally for both conditions and no difference was introduced by additional filtering. We confirmed with expert listening tests, that the difference between the auralization conditions was well noticeable.

4.2.1.1. Results and Discussion

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

Furthermore, we noticed in the data logs, that the prerequisite of equally distributed utterance directions was violated in 9 of 32 conditions. In these cases, one of the directions was only used three times and thereby another direction five times. This should be adapted for the exhaustive study by adding more items and thereby more potential item directions to pick from. During the pilot study no one had problems finding the requested items. We also added the possibility to repeat the lastly uttered sentence by a button press, which however nobody utilized. Additionally, we noticed that many participants moved while the ECAs were speaking, although asked not to do so in the task description. This might have influenced the hearing experience and should therefore be further prevented during the full study. However, having the ECA turn while speaking could enhance the audibility of the directivity. For the agents' speech we used a synthetic voice. When asked afterwards about this synthetic voice,

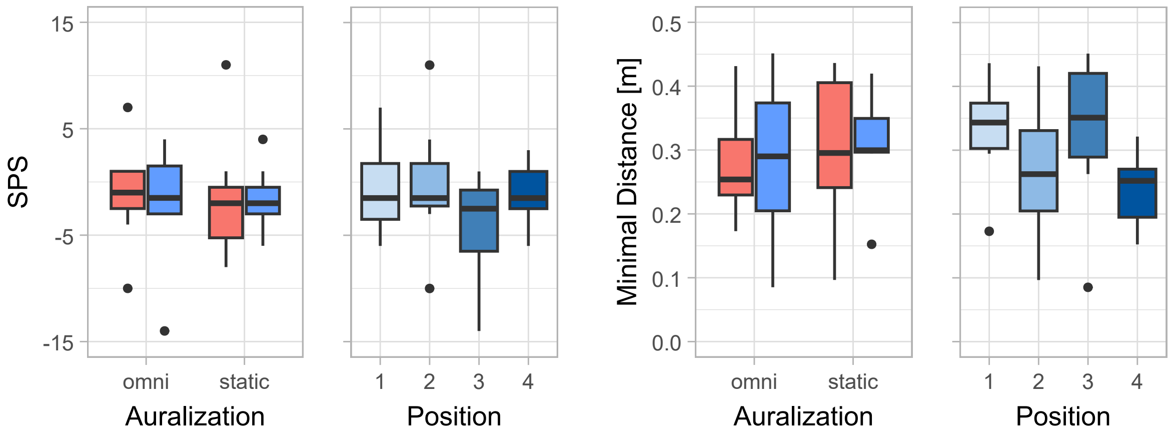
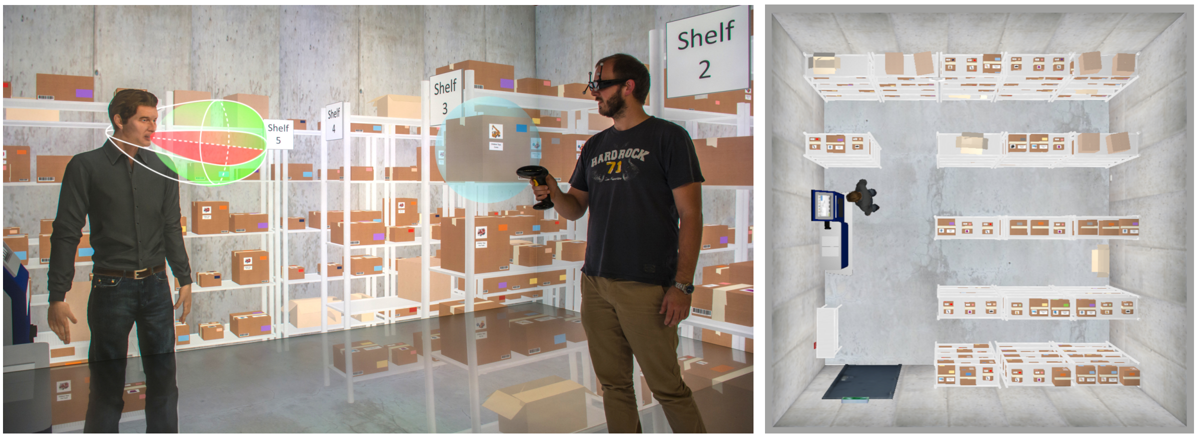


Figure 4.10.: Social Presence Survey (SPS) scores (left) and minimal distance kept to the ECA (right) with regard to auralization (omnidirectional vs. with static directivity) and gender (female (red) vs. male (blue)) and with regard to presentation position (so being presented as 1st, 2nd, 3rd, or 4th condition).

seven of the participants stated, in line with Sec. 4.1, that they would have preferred recorded speech, since the synthetic voice had both a negative influence on their "feeling of being there" (presence) and "their feeling of interacting with a real person" (social presence).

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

Looking at the recorded answers of the SPS questionnaires (see Fig. 4.10(left)), no trends for improvements of social presence between *omnidirectional* and *directivity* conditions is noticeable. A statistical evaluation is omitted due to the small number of participants ($n = 8$). The difference in SPS ($SPS_{dir} - SPS_{omni}$), between all pairs of conditions has a mean of -0.2 (SD: 4.01), while SPS can take on values from -15 to 15. This does not seem very promising for a large scale study, so better choices for questionnaires or scores should be evaluated to measure an effect if it exists at all. Furthermore, the considered minimal distances to the ECAs (see Fig. 4.10(right)) seem to rather exhibit order effects (M: -0.076 SD: 0.069) than being influenced by the *Directivity* (M: 0.025 SD: 0.102). So probably an additional training condition should be added for each *Gender* before the two conditions varying the *Directivity* begin. Then again, this objective distance measure might not prove insightful after all. We used ECAs with different genders to counter gender effects, which at least for the distance kept to an ECA cannot be ruled out (cf. [Bönsch et al., 2016]).



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

Figure 4.11.: (a) A participant holding a picked up item (within the blue sphere), which he was asked for by the agent. The directivity of the agent's speech sound source is exemplarily visualized, for a more precise visualization of the used directivity see Fig. 4.9. (b) Top view of the stockroom, with an agent standing next to the scanner and shelves filled with 237 collectable packages.

4.2.2. Study on the Influence of Static Speech Directivity

The contents of this section are based on and taken in part from work previously published in [Wendt et al., 2019].

Based on the findings from the pilot study, we conducted a second within-subjects study. The main improvement was to make the study task more interactive to foster more social interactions and thereby social presence. Participants have to assist the ECA in collecting a number of packages for an imaginary order. Therefore, the ECA gives the participants incremental hints for what and where to look and can be asked by the participants for further information using natural language. This led to more and longer speech acts of the ECA. Furthermore, we slightly adapted the scene to more realistically fit the scenario and used a better speech synthesis, while the rest was kept as in the pilot study. Analogously to the pilot study, this study had one independent variable: *Auralization*, with the two levels *omnidirectional* and *static directivity*. However, we used a better TTS solution to mitigate the detrimental effect thereof. Furthermore, we deliberately had the ECA turn while speaking, to make the differences in auralization more pronounced and only used male ECAs to avoid to also measure potential gender effects (see [Bönsch et al., 2016]). Lastly, we still measured the perceived social presence of the ECAs using the social presence survey (SPS) [Bailenson et al., 2001] (see App. B.2), however, also added questions regarding speech realism for each condition. SPS was used since it was, to our knowledge, the best available questionnaire measuring the concept of social presence.

We specifically designed this study to test the following hypothesis:

H1 *The perceived social presence of ECAs auralized using static directivity is higher.*
Since this setup is closer to our everyday experiences of the sound of speech, we expect this to subconsciously improve the participants' perception of the ECA.

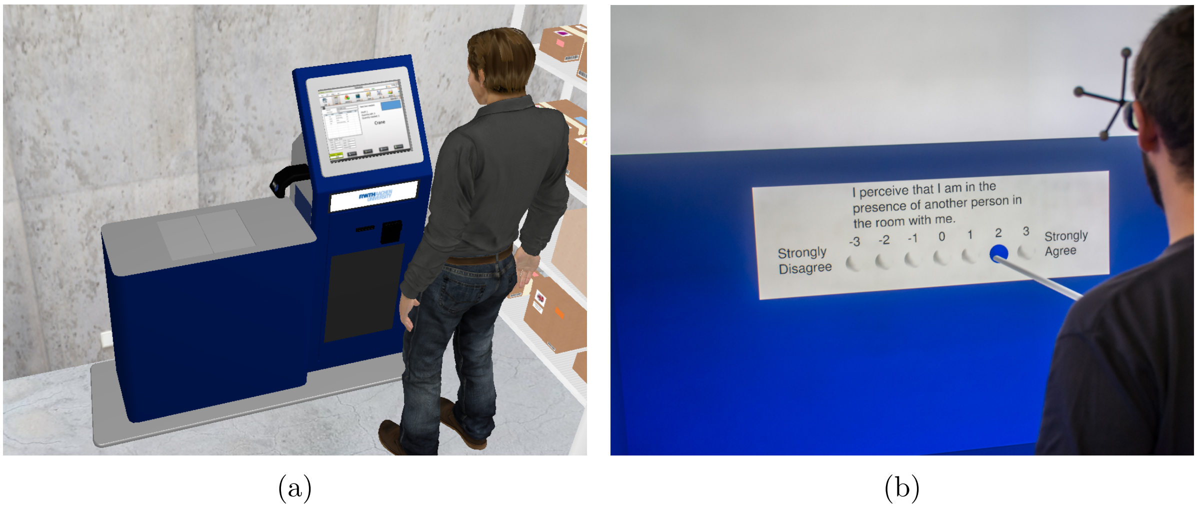
H2 *Participants prefer the ECA that is auralized using directivity.*
We expect that **H1** leads to participants preferring the ECA which is auralized with directivity, when asked to decide between the two differently auralized conditions in the end.

H3 *Participants rate the ECA that is auralized using static directivity as sounding more realistic.*
When specifically asked for the realism of the ECA's speech sound, we expect participants to rate the ECA using the directivity pattern as appearing more realistic.

Study Task

The collection task was again performed within a virtual stockroom exactly matching the dimensions of the AixCAVE, albeit with a different layout (see Fig. 4.11(b)). The stockroom was filled with five rows of shelves with in total 237 packages of 16 different types and a scanner terminal at which the ECA stood during the entire interaction (see Fig. 4.12(a)). The packages from different categories could be differentiated by distinctly colored labels, varying package sizes, a picture of the product inside and a distinct text label describing the content. Participants had to collect all required packages for one of two predefined orders requiring to collect 17 items in total each. Therefore, the ECA read out what is needed next while looking at the terminal, e.g., *"For this order we also have to find a teddy bear."*, with a small introduction statement for the first item, e.g., *"Please help me collect the items for order XS47."* In total, 10 requests per order were given, where some required multiple items of the same kind to be collected. In that case all of the packages were in the same shelf row. The ECA could be asked to give additional information, i.e., the shelf number, label color or size of the item. Unnoticeable for the participants, they were observed by the experimenter by means of real-time video and audio streams. This enabled the experimenter to control the ECA via a Wizard-of-Oz paradigm by triggering reasonable utterances depending on the participants' actions or questions using natural language. Thereby the ECA gave the participants incremental hints for what and where to look.

One of these utterances per request included a turning of the ECA towards the shelf row where the requested item was placed and thereby highlighted the directivity pattern, if present, since the direction-dependent changes are most noticeable during turning. If the participant did not ask for additional information, the experimenter at some point deliberately triggered the additional information to also present those sentences that involve turning of the ECA. Since the experimenter did not change between participants, consistency was increased. Additionally to these sentences, which are differently formulated for each request to avoid repetition



I perceive that I am in the presence of another person in the room with me.

(a) (b)

Figure 4.12.: (a) The ECA stands in front of the scanner terminal. Participants scan the collected packages by holding them close to the scanner mounted on the terminal's left side. Then the hatch opens and the packages falls through it. (b) After each condition a set of questions is presented in the AixCAVE. Participants answer these questions with a 6DoF input device using a ray-casting-based selection and a button on the device for the confirmation of their answers.

(cf. [Bönsch et al., 2017b]), the experimenter could also trigger sentences to tell the participant how many items are left, or confirm or disapprove of a question of the participant. When triggered by the experimenter, these phrases were randomly chosen from a pool of answers. For example, for the confirmation case one of the utterances *"Yes."*, *"Correct."* or *"That's right."* was used.

Navigation to the requested items was realized by means of natural walking as the dimensions of the virtual stockroom exactly matched those of the AixCAVE (see App. A.5.1). Participants had to find one of the required packages and pick it up using the 6DOF input device and carry it to the side of the scanner terminal (see Fig. 4.12(a)). Unrequested packages could not be picked up. If the package was close enough, the scanner beeped, the hatch temporarily opened and the item fell into the shaft. The ECA acknowledged this by uttering one of three thanks phrases and asked for the next item, unless all 17 items had been collected. When the ECA had turned to a shelf, it fixated the package once being picked up by the participant, but turned back to the scanner once the item was closer than 1.5 m. This way participants did not play with the ECA's gaze, i.e., wave items in front of its head, and thereby break immersion, which was experienced in the pilot study.

Study Procedure

Participants had to first give their informed consent and were then asked to fill out a demographic questionnaire and read a task description, leaving them naïve to the examined effect. Subsequently, participants entered the AixCAVE and were given some time to get familiar with the ECA, which said a welcoming sentence and looked at the participants. The scene was, apart from the ECA and a blue floor with a red circle, empty. Once the participants felt accustomed, which they indicated by stepping on the red circle, the scene faded out and the virtual stockroom (see Fig. 4.11(b)) was shown in which the participants had to perform the collection task described in Sec. 4.2.2. Once all items were collected, the scene faded out and participants had to answer a questionnaire with subjective measures (see Fig. 4.12(b)) while staying in the AixCAVE. The questionnaire items are displayed individually and the participant uses a 6DoF input device and ray casting to pick answers on a 7-point Likert Scale from "Strongly Disagree"(-3) to "Strongly Agree"(3). Additionally to the 5 items of the SPS questionnaire [Bailenson et al., 2001], we added three items asking for realism:

Real1 The motions of my co-worker looked realistic.

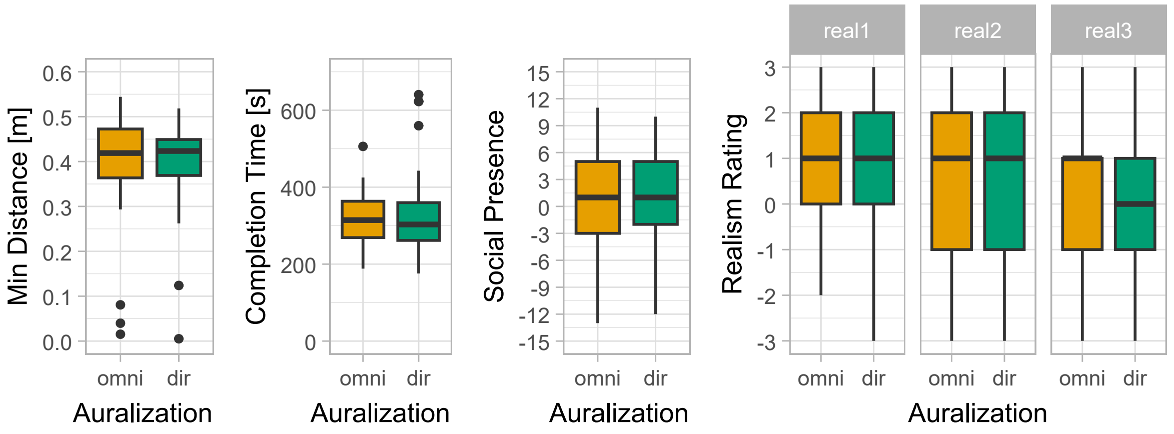
Real2 My co-worker sounded realistic.

Real3 The communication with my co-worker felt realistic.

These items were adapted from Poeschl and Doering [2013] and should enable the participants to subjectively rate the realism, so that potential changes can be detected. Thereby *R1* and *R3* were added to not give away the purpose of the study and not prime participants for the second condition. After answering these questions either the second condition was started or the participants were asked to leave the AixCAVE, if both conditions were completed. We counterbalanced the order in which we presented the two levels of *Auralization* and also the mapping of the prescripted orders to the levels of *Auralization*. Additionally to the subjective measures, we measured the minimal distance participants kept to the ECA, as these proxemics could potentially also be used to gain objective insights in the perceived social presence (see Sec. 2). The time needed for each condition was also logged, as task-related measure. After leaving the AixCAVE, participants were asked to fill out a post-study questionnaire asking them what they think was investigated and several questions to rate the experience and which of the two conditions they liked better concerning different aspects of the ECA.

Study Material

The study was conducted analogously to the pilot study, so only differences from the material explained in Sec. 4.2.1 are stated here. As ECA this time only the human model *Brad* from *SmartBody* [Shapiro, 2011] was used. The speech audio was generated using



(a) (b) (c) (d)

Figure 4.13.: Comparative results between the two auralizations: (omni)directional and static (dir)ectivity. The minimal distance kept (a), time needed for task completion (b), social presence rating (c) and answers to the three realism questions (d) are shown as box plots.

Google Cloud Text-To-Speech[4] [Shen et al., 2018] and the *Sphinx-4* library[5] was used to generate the matching lip sync data, which was used by *SmartBody* for the required lip-syncing.

4.2.2.1. Results and Discussion

We conducted the study with 36 participants (9 female and 27 male, mean age = 24.61 years, SD = 4.18). All of them had normal or corrected to normal vision and except for one, which was excluded from the analysis. All had normal hearing and at least basic English language skills. The participants were recruited on the university campus and compensated with free candy and drinks. Each participant spent approximately 15 minutes immersed in the AixCAVE. The participants were naïve to the subject of the study. Even after completing the immersive part of the study, only 6 participants (17%) suspected that the study investigated the effect of ECAs' speech or movement when explicitly asked to speculate on the examined effect at the beginning of the post-study questionnaire. Most subjects assumed that general human-VA-interaction (37%) or even Artificial Intelligence (11%) was investigated.

Comparing the objective measures between both conditions with *omnidirectional* (omni) or *directional* (dir) auralization (see Fig. 4.13), the minimal distance d kept from the ECA ($d_{\text{dir}} = 0.39\,\text{m}$, standard deviation (SD) $= 0.11\,\text{m}$ and $d_{\text{omni}} = 0.39\,\text{m}$, SD $= 0.12\,\text{m}$) and the time t needed for completion ($t_{\text{dir}} = 334\,\text{s}$, SD $= 118\,\text{s}$ and $t_{\text{omni}} = 319\,\text{s}$, SD $= 67\,\text{s}$) using a paired-samples t-test did not yield any significant differences (both $p > .51$). The same holds for

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

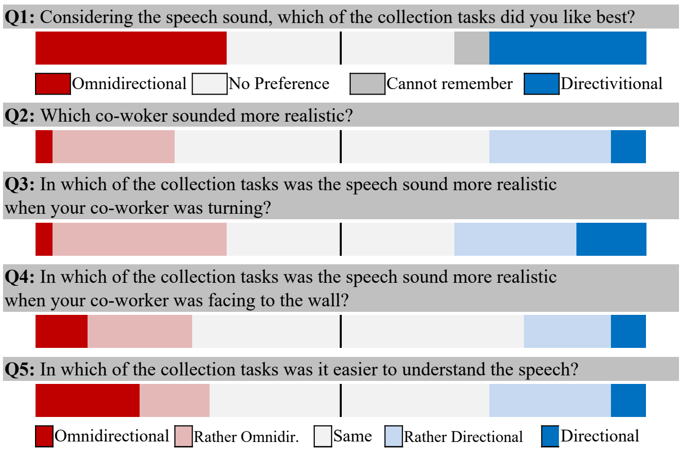


Figure 4.14.: Answers to questions *Q1* to *Q5* comparing the two conditions with regard to the speech sound.

the answers to realism questions *real1*-*real3* (see Fig. 4.13(d)). These questions asked for motion realism (*real1*), realistic sound (*real2*) and a realistic communication (*real3*) with the co-worker. *Real2* is the most crucial one with regard to our research focus, with means $real2_{\text{dir}} = 0.63$, SD = 1.70 and $real2_{\text{omni}} = 0.69$, SD = 1.76. However, neither questions showed significant differences (all $p > .71$). As second subjective measure, the Social Presence Survey (SPS) [Bailenson et al., 2001] score for both conditions was observed as $\text{SPS}_{\text{omni}} = 0.7$ (SD = 5.6) and $\text{SPS}_{\text{dir}} = 1.0$ (SD = 5.8). The SPS is measured using five questionnaire items on a 7-point Likert Scale (-3: Strongly Disagree, 3: Strongly Agree), which are summed (while the scores of item 3 and 5 are inverted, cf. [Bönsch et al., 2019]) and thereby yields values from $[-15, 15]$. However, analyzing these also did not show any significant differences between the two conditions ($t(34) = 0.55, p = .59$).

Furthermore, the results to questions *Q1* to *Q12*, which were posed in the post-study questionnaire right after leaving the AixCAVE are represented in Fig. 4.14 and Fig. 4.15. For *Q1* to *Q5*, these answers are corrected by the order in which the participants experienced the two conditions. In the questionnaire the items were labeled with "1^{st}" and "2^{nd}" instead of "Omnidirectional" or "Directional". The answers to these five questions show no significant preference for one of the conditions. We used a one-sample t-test, with Ordinal values "Omnidirectional"(-2), "Rather Omnidirectional"(-1), "Same"/"No Preference"/"Cannot remember"(0), "Rather Directional"(1) and "Directional"(2) and tested against the null hypothesis that the mean of the population is 0, where due to the sample size a test for normality can be omitted. However, for *Q1* the discriminating answers ("Omnidirectional": 31% and "Directional": 26%) don't show large differences. The same holds true if we compare them by the two different task orders/set of sentences ("order 1": 29%, "order 2": 29%). If we compare them by order of presentation, a slight preference for the condition shown in second place is apparent ("1^{st}": 20%, "2^{nd}": 37%). So neither of those invalidated the null hypothesis.

Using a one-sample t-test on the answers to *Q6* to *Q12*, with a null hypothesis test value of 3.0 (the middle of the used 5-point Likert Scale), results in significant divergences from the

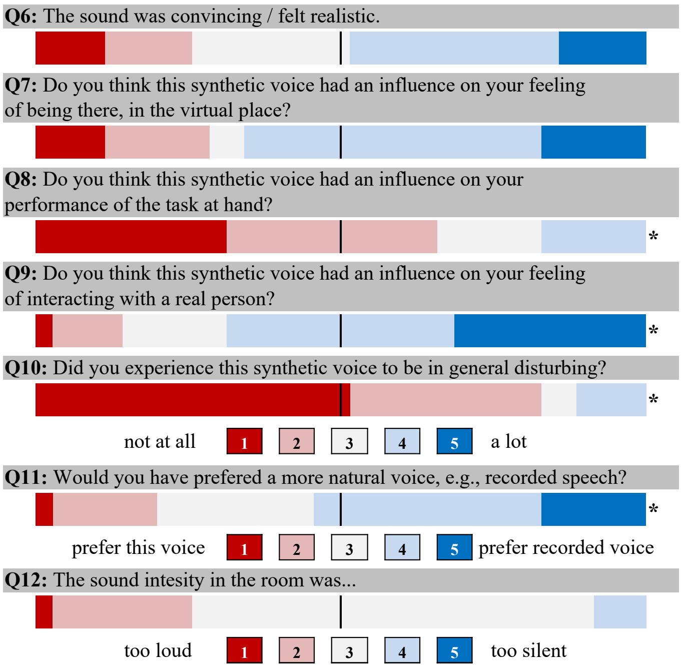


Figure 4.15.: Answers to the post-study questionnaire items *Q6* to *Q12* concerning the sound and the voice. They were all answered on a 5-point Likert Scale with different extremal labels as indicated below each block of questions. Significant divergences from a mean centered normal distribution are marked with *.

mean for *Q8* ($t(34) = -4.385$, $p < .001$), *Q9* ($t(34) = 4.465$, $p < .001$), *Q10* ($t(34) = -7.250$, $p < .001$) and *Q11* ($t(34) = 2.692$, $p = .011$). These are marked with * in Fig. 4.15.

When asked in the post-study questionnaire whether participants noticed a change in the auralization of the voice of the ECA, only 31% stated that they noticed a difference at all.

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

Discussion

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

A further challenge we faced is that evaluating the influence of the subtle auralization change using objective measures is complicated. We did not expect to find an influence on the completion time, since a higher realism of the voice should not change the difficulty of the task. However, also the other objective measure, the kept distance to the ECA, did not yield any insightful information either. This is probably due to the fact that proxemics is a better measure to differentiate effects of eeriness [Zibrek et al., 2017] than subtle changes of the voice. We noticed though that all except of two participants kept an "appropriate" (0.25 m to 0.50 m) minimal distance to the ECA (cf. [Bönsch et al., 2018a]), which could be an indication that they in general accepted the ECA as human-like counterpart.

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

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

Generally, the studied auralization techniques are advanced with regard to sound realism. Therefore, some participants remarked that the overall realism of the interaction and the scene was too low, thus the added realism due to the directivity was unnoticeable with the applied measures. This becomes also apparent in the answers to the realism questionnaire items and the free field comments, which span from *"The voice of the ECA was also reducing the quality of the experience, it was very 'robotic' sounding."* to *"I did not focus on the speech/sound at all, as the speech/sound itself was the most natural thing in my opinion"*. Furthermore, this also hints to another deficiency, the usage of a synthetic TTS voice. While we stated already in Sec. 4.1 that synthetic voices can have detrimental effects, we still used synthetic speech here for cost-efficiency reasons. We took special care that the best available synthetic voice was used and that prosody was natural. Additionally the focus in this study was on the naturalness of the auralization and not on that of the voice, which was kept constant. Nevertheless, we hypothesize that potentially more significant results could have been gathered if a recorded voice would have been used. Therefore, we opt for recorded speech in the following study.

Another possible explanation for not finding any significant differences between the probed auralization conditions, is that there is no or only a small effect when adding directivity to the auralization of an ECA during the task at hand, at least on the perceived social presence and realism. In the following study, we therefore examine further whether in a direct comparison between these conditions' effects on perceived realism can be measured when participants are primed and know on what to focus. Therefore, we designed a study in which no artificially social task is involved, but only a monologue of an ECA is experienced, using recorded speech and motion. Beyond that, we want to examine the effect of dynamic directivities. That means that the directivity pattern is influenced by the currently uttered phoneme. It remains to examine whether that is at all distinguishable from the static directivity that was used in this study.

4.2.3. Evaluation of Dynamic Speech Directivity

The contents of this section are based on and taken in part from work previously published in [Ehret et al., 2020].

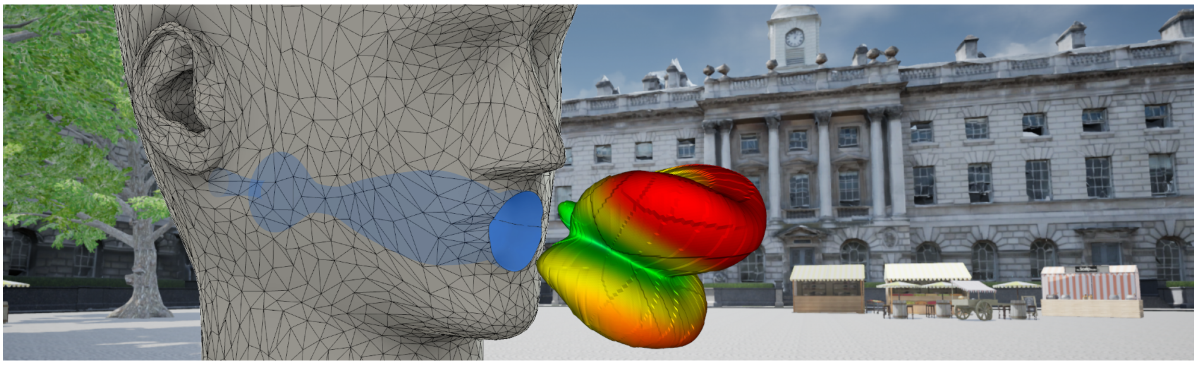


Figure 4.16.: An exemplary speaker directivity for the vowel 'a' at 1600Hz (modulation per direction shown as distance and color (from attenuation (green) to amplification (red)) displayed in front of the used head model including vocal tract within and mouth opening (blue). In the background the used study outdoor scene can be seen.

Aiming for higher naturalness and going beyond static directivities, where directionality is added but does not change based on the content of the speech, we simulated dynamic, phoneme-based directivities. During speech production, the human vocal tract greatly changes. Different vowels are associated with various mouth openings, which result in variance of the directional pattern [Katz et al., 2006; Arai, 2001]. As this is one step further to full realism, we wanted to evaluate whether this additional effort is perceivable and potentially beneficial. To this end, we conducted a study gathering preference and naturalness ratings of *dynamic (D)*[6] auralization,

[6] As stated before, this choice of words stands in contrast to related acoustics publications (e.g., [Postma and Katz, 2016; Postma et al., 2017; Ackermann et al., 2019]), where dynamic directivity is attributed to a dynamically moving sound source rotation and not a time-variant dataset.

in contrast to *omnidirectional (O)* audio, which emits the sound equally in all directions and the aforementioned *static (S)* directivity. Additionally, we tested whether the participants were able to reliably distinguish these auralizations.

4.2.3.1. Acoustics Simulation

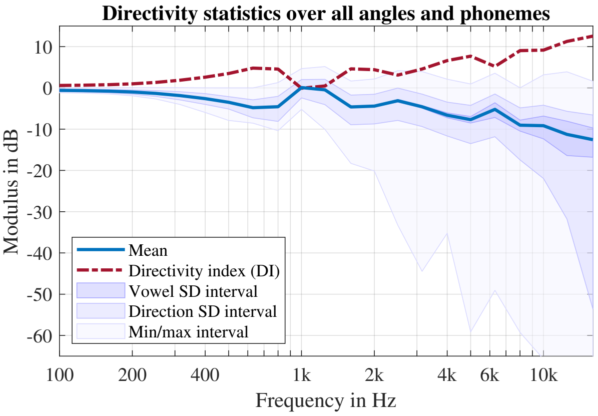


Figure 4.17.: Simulated directivity statistics over all angles and phonemes described by Arai in [Arai, 2001]. SD intervals are given as mean +/- standard deviation (SD), either between the simulated vowels or between all directions.

Directivity datasets for different vocal tract constellations can be either acoustically measured or simulated with physics-based approaches. The first method is commonly used for technical devices, like loudspeakers and musical instruments [Weinzierl et al., 2017], and employs an array of microphones surrounding the sound source recording simultaneously. A difference analysis reveals the directional pattern for the given tone or frequency, which is usually formulated as a relative change with respect to the frontal direction in an anechoic environment [Shabtai et al., 2017]. The measurement procedure suffers from a limited frequency range, where valid data can be acquired. Low frequencies are limited by the measurement chamber's ability to mitigate external noise, and high frequencies are limited by the spatial resolution of the measurement array. Thereby the high limit does not usually cover the full audible range. If the sound source can be described by a 3D model, the acoustic radiation can be determined by simulation, for example, with the Boundary Element Method (BEM). To simulate a phoneme-dependent directivity, a 3D head model of a human was selected and combined with different settings of a simplified tube model representing the vocal tract following Arai [2001], using COMSOL Multiphysics Version 5.4 for the audible frequency range (30Hz to 16kHz in third-octave resolution). A virtual sensor array arranged in an Euler grid of 1° angular resolution acquired the transfer functions from the tube's end where the vocal chords are located. As the model is symmetrical, only one side of the head and vocal tract model was simulated. The results were post-processed by Matlab Version 2018b to normalize the filter values to the frontal direction and mirror the half-sided dataset to cover the full sphere around the sound source. Finally, the data was exported in the OpenDAFF format[7], which provides an interface to access discrete directional data stored as a lookup table. Similar work was also done in [Johannsen, 2021].

In the end, three vowels were selected that cover the mouth opening range on the IPA vowel chart[8]. The strong similarity between directivities with the same mouth opening but differ-

[7] OpenDAFF: www.opendaff.org

[8] International Phonetic Alphabet (IPA): www.internationalphoneticassociation.org/

ent tongue position prompted us to neglect this dimension. Therefore, three representative directivity datasets for open mouth (a), half-open mouth (e) and closed mouth (i) have been simulated and all other vowels were mapped to these.

Fig. 4.17 shows the statistical evaluation of the simulated phoneme-dependent directivities with frequency modulations for all directions and phonemes. The curves indicate moderate damping in the lower frequency range, an articulated region around 1 kHz and slow roll-off towards higher frequencies. The directivity index (DI) is a measure to reveal frontal focusing, where levels above 0 dB are stronger towards the front. Hence, both the lower and the higher frequency range of the DI curve indicate a focusing to the front and attenuation toward other directions [Blauert, 1997]. Our mean deviations between different directions are well above the theoretically audible threshold of approximately 1 dB. In contrast, variations between vowels as indicated by the deep blue interval only show a slightly noticeable difference that exceeds 1 dB just for frequencies above 4 kHz.

Auralization

To render the acoustic virtual environment and to reproduce the binaural signal at the user's ears, we again employed the Virtual Acoustics real-time auralization framework[9]. The agent represented a dynamic, moving sound source that emitted the recorded speech signal. Directivity lookup tables were preloaded and could therefore be assigned to the sound source without delay. This way, switching between directivities had an instantaneously perceivable effect. The realization of the directivity filter bank used a transform of 256 interpolated frequency values from the third-octave resolution into the time domain, which provided an impulse response with the fingerprint of the directional attenuation. Hence, the directivity was made audible by convolving the speech signal with dynamic filter coefficients tied to the emission angle as calculated from the source's position and orientation, as well as the receiver's location. Free-field propagation simulation was applied covering properties like spherical spreading loss (amplification factor depending on relative distance), Doppler shift (resampling depending on relative movement) and the binaural filtering of the incident wave front at the receiver's ears (depending on listener's location and orientation). Again cross-talk cancellation [Masiero and Vorländer, 2014] using the 12-loudspeaker audio system of the aixCAVE (see Sec. A.5.1) was used for reproduction.

If the communication between an ECA and a user is primarily face-to-face, the directivity variation in the direct sound can be expected to have little effect on the signal. Consequently in our study, different emission angles were enforced by rotation animations of the ECA, and the subjects were encouraged to move in a defined area (see red circle in Fig. 4.18(b)). In an indoor environment, reflections off walls need to be taken into account. They can be determined according to the image source method by Allen and Berkley [1979]. If the ECA, for example, talks away from the user towards a wall, the damped direct sound is overlaid with the frontal speech sound reflected off the wall. To that end, we created five image sound sources for the walls and the floor. The ceiling was considered non-reflecting, due to

[9] Virtual Acoustics: `www.virtualacoustics.org`

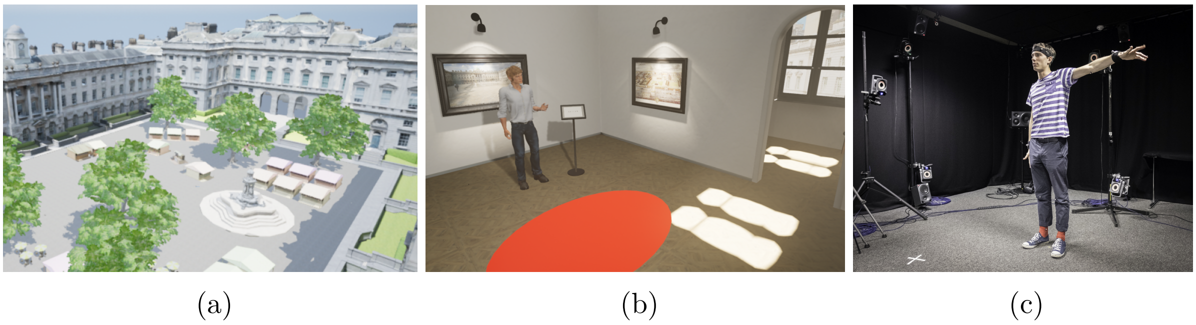


Figure 4.18.: The different Settings: (a) The Somerset House courtyard and (b) the museum room with an ECA, a museum stand and a red area the participants should move on. (c) The actor performing the full-body movements of the speech.

computational limitations and it not being visible in the 5-sided AixCAVE. These image sound sources used the same directivity pattern as the primary source since the filters are, by design, symmetrical. This method approximates early reflections within the room, but further aspects of reverberation were not simulated due to the complexity of providing real-time update rates during dynamic directivity switching. Furthermore, the experimental environment (AixCAVE) has a reverberant characteristic that cannot be acoustically treated without interfering with other components. This reverberant characteristic is, however, comparable to the virtual room used in the study. Nevertheless, a problem arises if directivity datasets are exchanged in indoor situations because the overall energy emitted into the room changes in an non-physical way. This is most apparent when an omnidirectional directivity dataset is replaced with a human directivity dataset during auralization because the higher frequencies are dampened drastically to the back of the ECA and result in a harsh drop in that region. This effect is expected and has been intentionally included in the study to investigate the effect of energy change on naturalness ratings.

4.2.3.2. Study Design

We conducted a within-subject user study in the aixCAVE (see Sec. A.5.1) to investigate the influence of the different ECA speech Auralizations, as well as free-field *outdoor* versus acoustically reflective *indoor* Setting, on naturalness and detectability. Thereby Auralization is varied from unrealistic *omnidirectional* over adding *static* directivity to realistic phoneme-dependent *dynamic* directivity. First, an ECA took the role of a tour guide giving a 90-second speech, while allowing the participants to move freely and change Auralizations. Participants provided naturalness and preference ratings. Second, we tested whether participants were able to detect differences between the Auralizations, by pairwise testing two Auralizations in an A/B/X task.

Hypotheses

We tested the following hypotheses:

H1 Because *static* and *dynamic* auralizations better simulate sound propagation, participants will rate the naturalness of these conditions higher than that of an *omnidirectional* auralization.

H2 Because the impact of *dynamic* auralization is subtle in face-to-face settings, participants will rate the naturalness of *static* and *dynamic* directivities equally.

H3 Related to **H1**, participants will prefer auralizations with higher naturalness.

H4 Because reflections may obscure directionality cues, the differences in naturalness ratings will be stronger in the free-field *outdoor* condition compared to the *indoor* case.

H5 In contrast to **H2**, participants will be able to reliably distinguish *static* from *dynamic* directivity in a direct comparison.

Materials

We situated the study in a virtual version of the Somerset House in London, using a freely available scanned model from Sketchfab[10]. This model was made more lifelike, using booths and trees shown in Fig. 4.18(a). Additionally, we modeled a virtual museum room with pictures of the Somerset House as well as the model visible through the windows (Fig. 4.18(b)). This way the exact same speech could be given by the virtual guide in both settings. The museum room was modeled with identical dimensions to the aixCAVE to match the local reverberation environment. For more images of the scene see App. A.4.1. The speech content was a 90 seconds long talk about the history and some architectural highlights of the Somerset House. An additional 30-second sequence of short words was used to feature English vowels moving from open to closed[8]. This sequence in particular required the phonetic intricacies of a real human speaker rather then synthesized speech, to be able to generate the nuances between the different vowels. Therefore, the verbal content was recorded by a native English speaker using a calibrated microphone with no frequency weighting (NTi Audio Norsonic M2230) positioned at 0.72m in front of the speaker's mouth under acoustically dry conditions. The speaker's face movements were simultaneously recorded using optical markers and a 14-camera Vicon Tracking system. Due to equipment limitations, the full body movement was recorded in a second pass wherein the speaker re-enacted the speech with co-speech gestures and head and body rotations, to showcase all parts of the directivity filters during replay (Fig. 4.18(c)). The movements were transferred to a virtual human model, created with Reallusion's Character Creator 3 (see Fig. 4.18(b)), using Autodesk Motionbuilder. Unfortunately, the recorded face motion did not fit the 3D model so the lip syncing was manually re-created by an artist using

[10] Somerset House site survey scan 2019 by Kimchi and Chips art collective: `https://skfb.ly/6svNI`

Reallusion's iClone 7. Unreal Engine 4.22 was used for presentation. Furthermore, the timings of the vowels in both speeches were manually annotated and a mapping was created from all occurring vowels to the three vowels used for directivity simulation, following their position on the IPA vowel chart[8]. Consonants in between were auralized using the directivity of the vowel before and diphthongs were split in the middle.

Tasks

Participants engaged in two tasks: the *Comparison* and the *Detectability* task.

In the **Comparison** task, the ECA gave the above mentioned 90-second speech with associated movements in front of the participants. The speech was always identical throughout the experiment to avoid distractions. During the speech participants had the option to switch between three levels of Auralization: *omnidirectional* or featuring *static* or *dynamic* directivity. The switching was done by three dedicated buttons on the interaction device and had no perceivable delay. A sign next to the ECA displayed a letter (A, B, or C) corresponding to the auralization (see Fig. 4.18(b)). Auralization-letter mapping was randomized across trials. Participants were encouraged to switch as often as they liked. Once the speech was over, a 4-item questionnaire was displayed next to the ECA, asking first "Which variant do you prefer?" Next, three 7-point Likert scales asked "How natural was Variant A/B/C" on a scale from "very unnatural"(1) to "very natural"(7). All of these questions had to be answered using of a virtual pointing ray from the interaction device to continue. Alternatively, participants could answer a subset of the questions and repeat the 90-second speech. Those who repeated the trial only did so once, although they had the chance to do so mor often. During the speech, participants were asked to move on a red ellipse (2.3m × 1.5m) displayed underneath them, to encourage listening from different directions. The task was repeatedly performed (three times) in both Settings, *outdoor* and *indoor*, counterbalanced for order of presentation across participants.

During the **Detectability** task three equidistant ECAs were placed on platforms in front of the participants, with signs indicating A, X, and B (see Fig. 4.19). If the participant clicked on a sign using a pointing ray and dedicated button, the respective agent started to speak and slowly rotate. Any other currently speaking agent was stopped. The speech content used here was a series of short words with all English vowels. Participants were told that A and B are always different and that X always matches either A or B. The participants were allowed to listen to each agent as much as they liked until they felt able to decide whether X sounded the same as A or as B. This answer was given via button press, only after each model had been played at least once.

Procedure

Participants first gave informed consent, provided demographic information and read task descriptions. Next, they entered the AixCAVE and performed a practice trial of the *Com-*

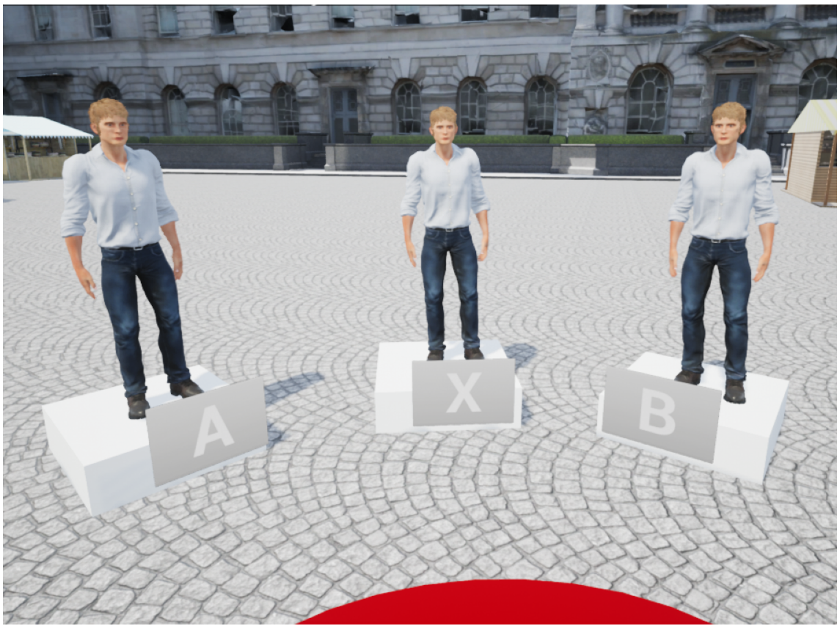


Figure 4.19.: The setup of the *Detectability* task.

parison task, which included virtual interface instructions as well as the standard post-trial questionnaire. The practice trial compared the omnidirectional auralization to two very artificial sounding high- and low-pass filtered auralizations, with the expectation that participant would prefer the omnidirectional one. Participants who preferred the artificial auralizations were excluded from analyses. Experimental *Comparison* trials were blocked for *outdoor* versus *indoor* Setting, randomized for order of presentation across participants. For each of the three experimental trials within each Setting, each Auralization was randomly paired with A, B or C. After six *Comparison* trials, participants were asked to answer the 5-item Social Presence Survey (App. B.2) and leave the AixCAVE for a 5-minute break. Next, participants returned to the AixCAVE for nine trials of the A/B/X *Detectability* task, which presented each possible combination of Auralization pairs three times in randomized order. All participants performed the *Comparison* task prior to the *Detectability* task to avoid direct comparisons influencing our measures of naturalness and preference. Finally, participants filled out a post-study questionnaire, addressing the ease of the *Detectability* task, as well as the intensity and realism of the audio, on 7-point Likert scales.

Measures

In addition to the in-VR decisions and questionnaire responses, we recorded the position and orientation of the participant and the currently speaking ECA. From these, we summed up the distance participants' heads moved between two frames during the *Comparison* task, as well as the change in emission angle. The emission angle is the angle under which the speech sound source was heard, relative to the ECA's head, and can be computed as:

$$\measuredangle_{\text{emission}} = \measuredangle(\ \mathbf{p}_{\text{participant}} - \mathbf{p}_{\text{agent}}, \ \vec{d}_{\text{agent}} \ )$$

with the position $\mathbf{p}$ and forward direction $\vec{d}$ of the participant's and agent's head. We compute the maximum emission angle encountered during each of the Auralizations in every trial, as well as the sum of emission angle changes between all adjacent frames.

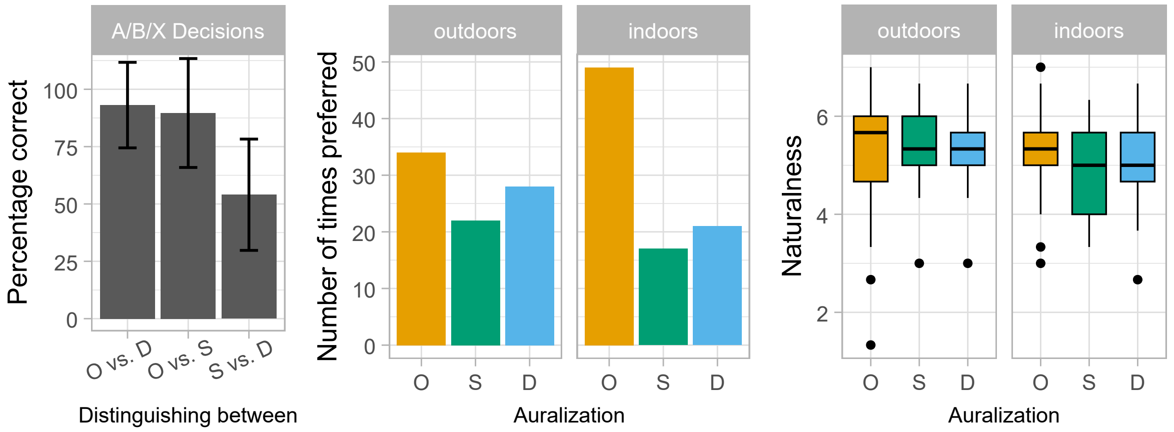
4.2.3.3. Results and Discussion

32 participants (9 female) were primarily recruited via university mailing lists. Three participants were excluded from all analyses for failure to select the correct response during the practice trial. Furthermore, answers to three *Comparison* task trials were removed where the participant failed to listen to all three auralizations. The remaining 29 participants (8 female) had a mean age of 28.24 (SD = 9.85), and all reported normal hearing, normal or corrected vision and some English skills (2 "basic", 27 "fluent").

The participants rated the sound's realism on a 7-point Likert Scale from "not at all"(1) to "a lot"(7) (M = 5.00, SD = 1.10) and the sound intensity on a scale from "too silent"(1) to "too loud"(7) (M = 4.07, SD = 0.60) overall appropriate. Furthermore, the ease of the *Detectability* phase was rated on a 7-point Likert Scale from "not at all"(1) to "a lot"(7) (M = 3.65, SD = 1.78) as reasonable. Social Presence was rated with a neutral 0.1 (SD = 4.08) of the possible scale from -15 to 15.

The results of the *Detectability* task can be seen in Fig. 4.20(a). A planned chi-square test of independence was performed to examine the relation between the different A/B/X-comparison-pairs and the ability to detect whether X was using the same Auralization as A or B. There was a significant difference between response accuracy across comparison pairs, $\chi^2(2, N = 261) = 48.97$, $p < .0001$. Furthermore two-sided binomial tests showed that *omnidirectional (O)* can be distinguished from both *static (S)* and *dynamic (D)* speaker directivity significantly better than by chance, ps < .0001. The A/B/X decisions between S and D speaker directivity, on the other hand, were not significantly different from chance (.5), $p = .52$.

Fig. 4.20(b) highlights the frequency of preference selection in the *Comparison* task. A planned chi-square test did not reveal a main effect of the Setting: *outdoor* vs. *indoor*, $\chi^2(2, N = 171) = 4.30$, $p = .116$. The relation between preference and naturalness ratings of the different Auralizations was further analysed using a planned two-way repeated measures ANOVA. There was a statistically significant interaction between preference and Auralization on naturalness ratings per trial, $F(4, 504) = 39.6$, $p < .0001$. Therefore, the naturalness ratings per Auralizations were analysed for each preference rating. P-values were adjusted using the Bonferroni multiple testing correction method. The effect of preference was significant for all three preference outcomes (ps < .0001). Pairwise comparisons, using paired t-test, show that the naturalness rating of the preferred Auralization is always significantly higher than the ones of the other two Auralizations (ps < .0001). Furthermore the differences between the



(a) (b) (c)

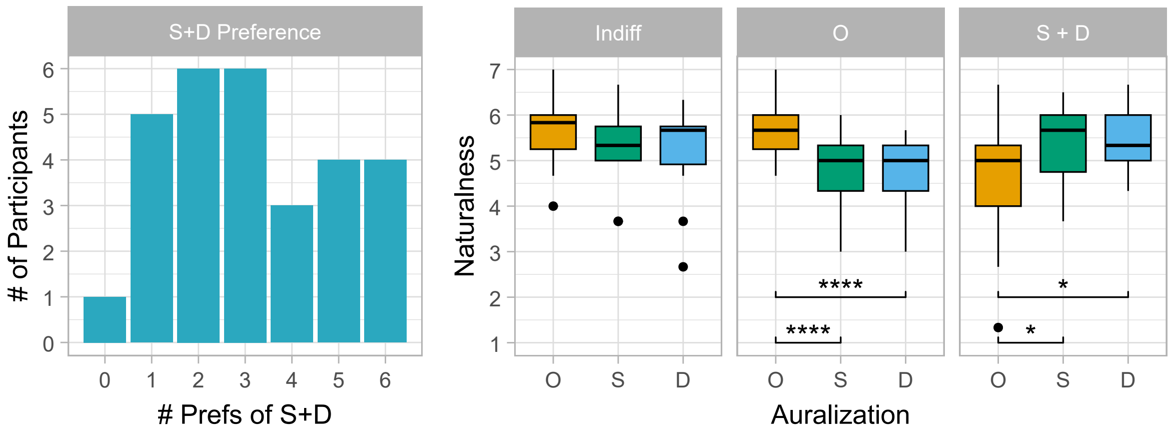
Figure 4.20.: (a) The percentage of correct answers when trying to differentiate between the different Auralizations *omnidirectional* (O), *static* (S) and *dynamic* (D) in an A/B/X comparison. Error bars show the standard deviation. (b) The number of preferences per Auralization split by the two Settings: the courtyard scene (*outdoor*) and the museum scene (*indoor*). (c) Box plot of naturalness ratings per Setting and Auralization averaged over 3 trials based on a 7-point Likert scale from “very unnatural”(1) to “very natural”(7). Boxes indicate quartiles, with whiskers at full range.

non-preferred Auralizations were non-significant (S vs. D when choosing O ($p = 1.0$), O vs. D when choosing S ($p = .13$) and O vs. S when choosing D ($p = .09$)) .

Fig. 4.20(c) shows the mean values of the naturalness ratings, averaged over the 3 trials per Setting. A planned 2x3 (Settings by Auralizations) repeated measures ANOVA revealed no significant effects (Fs $< .932$, ps $> .40$) and only a marginal trend for Setting ($F(1, 168) = 3.231, p = .074$), with slightly higher naturalness ratings for *outdoor*.

Looking at the preferences of single participants (see Fig. 4.21(a)), we saw that the number of times single participants preferred either Auralization seems to be bi-modal. We combined S and D here following the results of the *Detectability* task. This way we found that there is a group of participants strongly preferring $S + D$ while some other participants were consistently in favor for O with a noticeable gap in between. Due to this observation we split the population into three groups: those preferring O (N=12) or $S + D$ (N=11) more often and individuals preffering these two auralization (groups) equally often (N=6), labeled as *Indiff.*

A post-hoc 3x3 repeated measures ANOVA showed a significant interaction between this preference grouping (further called Preference Type) and Auralization, $F(4, 165) = 8.58, p < .0001$. Therefore, the effect of the Auralization was analyzed for each Preference Type. P-values were adjusted using the Bonferroni multiple testing correction method. The effect of Auralization was significant for O-preferred ($p < .001$) and S+D-preferred ($p = .009$) but not for the in-



(a)

(b)

Figure 4.21.: (a) How many participant preferred *S+D* how often. Following the results of the *Detectability* task, *S* and *D* are combined. (b) Mean naturalness ratings per preference type (where *O* and *S+D* preferred this Auralization more than 3 times and *Indiff* both exactly 3 times) and Auralization averaged over 3 trials based on a 7-point Likert scale from “very unnatural”(1) to “very natural”(7). Significance levels are displayed as , $*\ p < .05$, $**\ p < .01$, $***\ p < .001$, $****\ p < .0001$.

different group ($p = 1$). Pairwise comparisons, using paired t-test, show that the naturalness ratings were significantly higher for the preferred Auralization in these two groups (O-preferred: $ps < .001$; S+D-preferred: *O* vs. *S* $p = .02$ and *O* vs. *D* $p = .01$), while no other differences were significant ($ps > .39$).

In our next post-hoc analysis, we examined participant movement during the *Comparison* task while the agent was speaking. Changes in auralization become most noticeable when the participant experiences a variety of emission angles. Participants moved an average of 123m (SD = 86m) with an accumulated change of emission angle of $13,814°$ (SD= $3,657°$). The mean maximum emission angle per Auralization and trial was 104° (SD = 37°). However, there was no correlation evident in the data between this maximum emission angle and individual naturalness ratings per Auralization and trial ($p = .13$). There was also no significant correlation between the accumulated movements and naturalness ratings grouped by Auralization ($ps > .22$), with only a marginal negative correlation between accumulated emission angle change and the naturalness rating of *O*, $r(56) = -.24, p = .067$. Furthermore, there was also no significant correlation regarding movement and the quantity of preferences for either Auralization ($ps > .19$).

Discussion

The results of the *Detectability* task show that participants were able to differentiate ECAs having an *omnidirectional* auralization from those using a directional one. However, participants were unable to reliably distinguish between *static* and *dynamic* directivity, even when given control over the rotation of the agents. This pattern contradicts hypothesis **H5**. However given the subtle differences in the simulated directivity data of the different vowels (see Fig. 4.17), participants inability to distinguish both directional auralizations is unsurprising.

In general, we did not observe higher naturalness ratings for the two directional auralization methods, leading us to reject **H1**. This can potentially be attributed to the fact that the low-pass filters applied in the directional auralizations in cases of lateral or even backward emission angles may detract from their perceived naturalness. While directional auralizations may be closer to reality, users seem to prefer that speech is unchanged. This finding allows for speculation about the role of low-pass filtering on speech intelligibility [Jax and Vary, 2006], with participants preferring the condition with higher speech intelligibility over the more realistic one. We also found, however, that several participants consistently select either strongly in favor of the directional auralizations or the omnidirectional auralization, while others remained indifferent (see Fig. 4.21(a)). We therefore performed a post-hoc split based on this preference, and found significant differences in the naturalness ratings in favor of participants' preferred auralization (see Fig. 4.21(b)).

Even under this split, there was no significant difference between *static* and *dynamic* auralizations. One possible explanation is that participants were not listening enough to non-frontal directions, as the virtual environment implied face-to-face communication. The measured maximum emission angles, however, exhibit sufficiently large movements as to elicit directional effects (mean above 90°). Furthermore, a post-hoc analysis did not show any significant relationship between participant movement and naturalness or preference ratings. A marginal trend revealed that participants who moved more (i.e., those who had a larger summed emission angle, indicating effort to detect differences), rated the naturalness of *omnidirectional* slightly lower. We take this result to suggest that our study circumvented the possible impact of insufficient movement, which would have otherwise hidden auralization differences at non-frontal emission angles. Therefore, we cautiously support **H2**.

Furthermore, we were able to support **H3**. Participants rated the naturalness of their preferred auralization significantly higher than that of the other two. This aligns with proposals that listeners prefer naturally sounding ECAs [Zibrek and McDonnell, 2019], and affirms the need to optimize for higher naturalness.

Given the results above, however, it may not be necessary for ECAs to implement dynamic, phoneme-dependent directivity for a comparable face-to-face interaction. In our case, the added effort over a static directivity seems unrecognized by the listeners. This would make auralizing ECAs' voices easier, eliminating the extra link between the animation system and the acoustical simulation to switch directivity filters based on the currently uttered phoneme.

We were unable to confirm **H4**, pertaining to the strength of naturalness differences across free-field *outdoor* versus reflective *indoor* Setting. We expected that the different directivity spectra per propagation path of the additional indoor reflections would entail perceivable deterioration of the directional effect. Furthermore, the energy mismatch caused by the missing damping of an *omnidirectional* source results in an unnatural room acoustic characteristic. Contrastingly, participants rated naturalness *outdoor* slightly higher where these effects did not occur. As no significant differences were found, however, we cannot make any conclusions about changes related to *indoor* reflections. The real reverberation of the inside cavity of the AixCAVE, where the walls and floor have acoustically problematic properties, may have interfered with the successful investigation of this effect. Future studies might be able to better control the experimental environment acoustically.

4.3. General Discussion

In this chapter we presented four experiments concerned with verbal aspects of ECA speech. These are one key factor to enhancing the naturalness and perceived social presence of ECAs.

The first user study examining synthetic voices and especially faulty prosody generated by TTS showed that if naturalness is a key requirement, for example, to elicit higher social presence, then recorded human voice is preferable over synthetic voices. The results of the three experiments presented after that also yielded results along these lines. These experiments were conducted prior to the one presented at first, therefore they partly used synthesized speech material. For the remainder of this thesis and also as general guideline, we propose to use recorded speech, where possible. However, sometimes the resources will not allow the usage of recorded speech and also some very dynamic scenarios potentially including large language models (LLMs) for content creation will have to rely on TTS. Nevertheless, these results are based on state-of-the-art synthesizers at that time. As technology progresses TTS solutions have to be reevaluated, however, paying special attention to prosody as identified in this thesis as one key factor of naturalness degradation.

We also presented three studies concerned with the directional auralization of speech sound sources. We found hints that the integration of dynamic, phoneme-dependent directivities was not distinguishable from a static (averaged) speaker directivity. We therefore suggest that the additional effort to switch directivities during speech is not required, and we expect this to hold for comparable scenarios. Furthermore, we found no evidence that participants prefer directional speech sound in general. While nearly half of our participants preferred the auralization including directivities, there were also many participants with a strong preference for omnidirectional speech. The fact that those groups consistently reported their respective preference gives rise to the notion that subjective preference is more related to other factors not considered here (e.g., speech intelligibility [Sewell et al., 2023]) than the realism of directional rendering. Nevertheless, we found that participants generally preferred the auralizations they rated as more natural, which affirms the need for high naturalness in speech auralization. Another possibility is that some participants simply preferred the increased speech intelligibility due to missing (potentially muffling) filters when using the omnidirectional auralization. As we are ultimately engaged in a virtual interaction, we do not have to adhere to physical laws and could implement ECAs that can be well understandable even if they are talking away from the user if that is beneficial for the application purpose at hand. So, “super human” capabilities should be considered for contexts where speech intelligibility is very important. Nevertheless, we propose the usage of static directivity filters, as this objectively increases the realism of the audio simulation and is easily implementable. It is therefore used for the remainder of this thesis.

Another learning from the lastly presented study is that future studies would be well served by using a more acoustically controlled environment, with a more robust reproduction (i.e., headphones), to better distinguish reflective and free-field settings. Therefore, we opted for HMD presentations using headphones for the following investigations.

In conclusion, these findings highlight critical aspects of verbal behavior of ECAs to enhance social presence, paving the way for future improvements in human-ECA interactions across various domains. By prioritizing user preferences for naturalness and intelligibility in speech interactions, developers can create ECAs that resonate more deeply with users, ultimately leading to richer and more meaningful engagements.

Chapter 5

Co-Verbal Behavior

When humans speak, their verbal behavior is nearly always accompanied by co-verbal behavior, for example, by gestures but also gazing etc. For embodied conversational agents (ECAs), this also becomes a necessity since they are virtually embodied and therefore these embodiments should behave in accordance with the given speech to not appear live-less and unnatural. Wang and Ruiz [2021] give an overview about non-verbal behavior, comprising, gestures, gaze, facial expression, proxemics, posture, and mimicry. We will further call these co-verbal to stress the co-occurrence with speech, but will focus on the same behaviors as Wang and Ruiz [2021] in this chapter. We will first discuss related work and our implementation of facial animations, both to convey emotions and articulation movements (Sec. 5.1), which is then followed by an explanation how to realize appropriate gazing behavior in Sec. 5.2. Subsequently, we will delve into co-verbal gestures and the different ways to produce those in Sec. 5.3, including the presentation of a user study we conducted to evaluate the performance of different manipulations to improve recorded gestures. Next, we will focus on two important communicative functions of co-verbal behavior essential for effective face-to-face communication which go beyond the uni-modal behavior in the previous sections: In Sec. 5.4, we will cover related work on back-channeling, where ECAs provide feedback, such as nods or vocalizations, to indicate understanding and engagement, which signals active listening and fosters a supportive communication environment. In Sec. 5.5, we will examine turn-taking cues in ECAs, referring to how speakers and listeners continuously manage who speaks next to ensure a smooth conversation and prevent unwanted interruptions. In this section we also present a user study we conducted, to examine the influence and interdependence of different non-verbal modalities involved in communicating turn-taking. While back-channeling also includes the aforementioned use of mimicry and posture, we do not include an explicit section about proxemics [Bönsch et al., 2018a] here, since the main focus of this work is on verbal interactions with the interlocutors staying at given positions throughout the interaction.

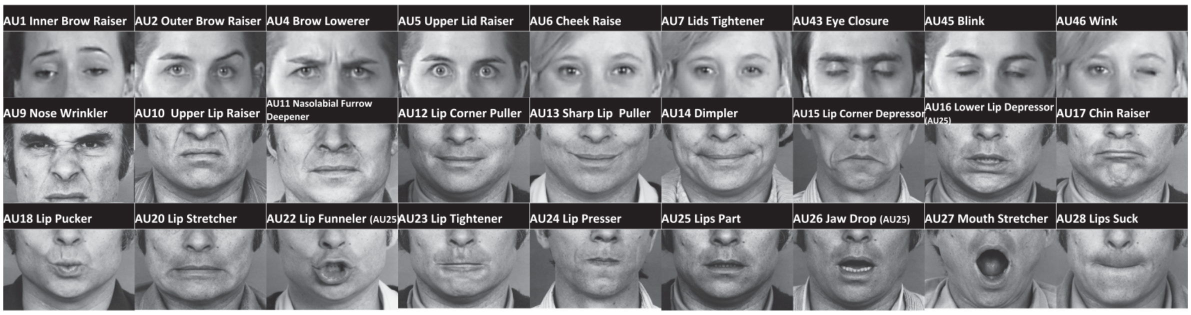


Figure 5.1.: A list of upper and lower face action units (AUs) as defined by the Facial Action Coding System (FACS) and their interpretation. Image taken from [Martinez et al., 2019]

5.1. Facial Animation

Human faces naturally move a lot during interactions. So special care has to be taken to adequately reproduce these movements in ECAs for them to not appear life- and emotion-less. Thereby these facial animations can constitute articulation movement (Sec. 5.1.1), potentially subtly, convey emotions (Sec. 5.1.2), or stem from other behavior of the ECA, like gazing (Sec. 5.2), which also requires the eye region to move according to the gazing.

For realizing these facial animations on ECAs, two main approaches exist. The first involves virtual bones in the face (see, e.g., [Aneja et al., 2019]), similar to those used for the remaining skeleton, that can be used to move small part of the face, like raising the eyebrow or opening the mouth. While this bone-based approach is often still used for the eye ball, since they are often required to be rotated to specific angles, for facial animations blend shapes[1] are more commonly used. These are a collection of different manipulations of the face vertices. For example, a "puff" blend shape can hold 3D displacements for the vertices of the cheeks to move them outwards to make a puffing face. This transformation of the face would be particularly hard to achieve using only bone-based transformation. To produce a facial animation multiple of these blend shapes can be blended together simultaneously, by giving each an activation value $\neq 0$ and then adding up the individual vertex displacements scaled by these activation values (see [Lewis et al., 2014]). One commonly used system for defining the different blend shapes is the Facial Activation Coding System (FACS) [Ekman and Friesen, 1978] (see Fig. 5.1), which was originally designed to analyse facial expressions in humans (see, e.g, [Hamm et al., 2011; Martinez et al., 2019]) but can also be used to drive facial animations of ECAs [Zhou et al., 2018]. Here the individual blend shapes, which are considered the most distinct individual movements a human face can perform, are called action units (AU).

[1] Blend shapes are also often called morph targets.

5.1.1. Articulation Movements

One crucial facial animation for ECAs is adequate articulation movement, i.e., an accurate lip synchronization [Bear and Harvey, 2017]. First of all, talking embodied agents that do not move their mouth are at least intuitively confusing and potentially detrimental to perceived realism (see, e.g., [Bönsch et al., 2017b]). Further Thézé et al. [2020] showed that the McGurk effect, i.e. incongruent visual speech overwriting what is heard, e.g., from “vase” to ”base” with another visual animation, also persists with virtual humans and artificial voices. Additionally, research indicates that viewing an articulated face while listening to speech enhances intelligibility (see [Gonzalez-Franco et al., 2017; Sewell et al., 2023]), which is crucial for effective communication and engagement in VR-based interactions with ECAs. Therefore, special care has to be taken to animate the lips appropriately.

Creating lip sync for ECAs involves various approaches, with phoneme mapping being one of the most straightforward methods. This technique dissects the speech signal into its smallest units, known as phonemes, and maps each to corresponding facial expressions of the mouth region called visemes. Implementations can vary significantly, with some using as few as three visemes [Llorach et al., 2016] while others employ up to 20 [Zhou et al., 2018]. Here more visemes clearly perform better (cf. [Bear and Harvey, 2017]). These mappings can be based on hand-crafted rules (e.g., [Llorach et al., 2016]) or data-driven methods using machine learning as, for example, in [Taylor et al., 2017] or [Zhou et al., 2018]. A review on used phoneme-viseme mappings can be found in [Bear and Harvey, 2017], also showing that for speaker identification more than one viseme should be mapped to the same phoneme. However, simply blending between visemes does not yield natural animations due to the need to consider co-articulation of phonemes. To address this, Xu et al. [2013] proposed hand-crafted animations for all pairwise combinations (bigrams) of visemes to improve realism. Additionally, alternative data-driven models have emerged that generate lip movements directly from raw audio input and potentially transcripts (e.g., [Karras et al., 2017; Taylor et al., 2017; Song et al., 2019]), allowing for the capture of more nuanced lip movements by inherently considering neighboring sounds rather than relying on a fixed set of visemes. There further exists a large body of related research just concerned with generating correct lip sync in videos (e.g., [Kumar et al., 2017; Suwajanakorn et al., 2017]).

While accurate lip movement is essential for effective communication, it is equally important to incorporate additional elements that contribute to realistic speech production. In particular, incorporating tongue animation is crucial for accurate articulation since it plays a significant role in human speech production [King and Parent, 2005]. Beyond that, research by Chen et al. [2019] highlights that head movement enhances the perceived naturalness of speech performance if it is in accordance with the speech [Hadar et al., 1983]. Fares et al. [2022] trained a data-driven model to generate continuous eye brow and head movements semantically fitting the speech. Moubayed et al. [2009] stated that adding natural head nods and eyebrow movement to the presentation of speech can significantly enhance speech intelligibility, even when prosodic cues were deliberately removed from the speech. Liu and Ostermann [2011] supported this notion by demonstrating how head movements and mouth shapes sampled from a database of recorded speech improved naturalness in their studies involving video rather than 3D embodiments.

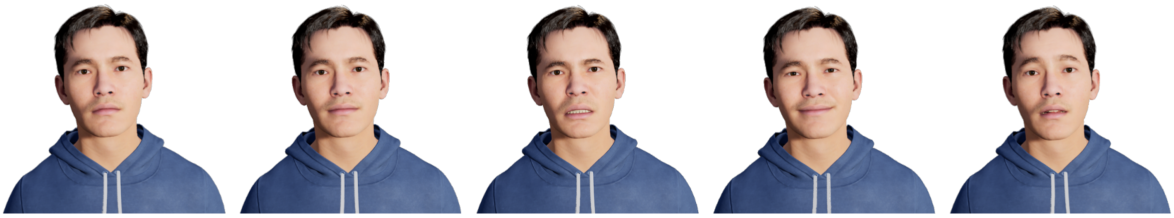


Figure 5.2.: Facial emotions realized via blend shapes in our system exemplarily shown on a MetaHuman. From left to right: neutral, slightly smiling, disgusted, happy, surprised.

In addition to these individual features, holistic communication dynamics encompass broader non-verbal elements that enhance animation realism. Breathing-related movements of the head and chest are also important aspects often overlooked in animated performances, as identified by Bernardet et al. [2019]. Lastly, visual speech should adapt based on speaking styles and volumes. For example, an agent's facial movements should differ when whispering compared to shouting (see [Miyawaki et al., 2022]). The *JaLi* framework proposed by Edwards et al. [2016] articulates this concept by stating that speech can be conveyed through either jaw or lips movements——or a combination thereof——depending on intensity and style.

5.1.2. Conveying Emotions

One important aspect of facial expressions and animation is the transport of emotions (see [Pelachaud, 2009]). Ito et al. [2024] found that emotional expressions were well able to shape the interaction of a human user with an ECA. This was in their study specifically helpful, when the ECA displayed emotions that discouraged cooperative attitudes, whereby participants tended to exhibit more selfish behavior, which was desirable when the ECA was intended to provide assistance. However, negative emotions presented as facial expressions, can also significantly reduce trust into the ECA [Luo et al., 2023]. Sato et al. [2024] furthermore found that emotions expressed by ECAs are beneficial in negotiation training. Additionally, the dynamics of emotional expression should be considered [Pelachaud, 2009], so that not only static expressions are held. One example of these dynamic emotional expressions for the upper face was developed by Liu et al. [2024a], where a data-driven model is trained to generate continuous emotional expressions fitting the speech that can be rendered on top of the lip sync. A database for different realizations of emotions can be found in [Pan et al., 2024]. Beyond facial animation also other aspects like texture can be used to convey emotions, for example, [Jung et al., 2011] used blushing as additional cue. Nevertheless, Lang et al. [2012] also found that using facial expressions has to be carefully considered, as they can be very context- and also culture-dependent. Furthermore, some emotional expressions like smiling can also be audible in the voice, we should be considered when designing emotional expressions [Torre et al., 2019, 2021].

In the *Character Plugin* (see App. C.3) already introduced in the previous chapter, we provide the possibility to dynamically add emotional expressions to ECA performances, based on FACS. Additionally to individually triggering these AUs, we provide ready-to-use expressions for basal emotions: happiness, sadness, surprise, fear, anger, disgust (see examples in Fig. 5.2). These expressions were based on the involved AUs provided in [Martinez et al., 2019], however, we adapted them to fit our needs. Additionally we also provide a slight smile to apply by default to the ECAs, since the neutral facial expression often looks too stern for social applications (see two left-most emotions in Fig. 5.2). During application development, developers will, however, need to carefully check and evaluate these expressions on the virtual human models used, to verify that the emotional expression matches what is needed. If this needs to be formally evaluated tools like the Geneva Emotion Wheel[2] can be used, as we did—however, for a predecessor implementation—in [Bönsch et al., 2020c].

5.1.3. Recording Facial Movement Along With Speech

Due to these intricacies of articulation movement above and since we used prerecorded speech in several studies, we opted for recording face movement along with the speech audio. In a first attempt (see [Ehret et al., 2020]) we tried to use an optical tracking system and markers in the speaker's face. Due to a lack of motion capture post-processing expertise and the recorded data not fitting the virtual model of the face, we in the end manually recreated the face movement using Autodesk Motionbuilder, which was very time-consuming. An alternative semi-automatic approach can be found in [Deng et al., 2006].

In a pursue for more automatic methods, we evaluated two video-based solutions, when being challenged again to record facial animations along with speech for [Ehret et al., 2021]. In our tests, using a recording based on Apple's *TrueDepth Sensor* [Esselink et al., 2023] and the *ARKit* turned out to work better than capturing the face using purely RGB-video based solutions like *OpenFace 2.0* [Baltrusaitis et al., 2018] as proposed in [Struijk et al., 2018]. Furthermore, we evaluated a data-driven approach based on speech only using *Oculus Lipsync*[3]. The approach using the ARKit, however, performed superior in our pre-tests and required only little post-processing. Therefore, we used it for most studies, employing Epic's *Live Link Face* app [4] for recording (see Fig. 5.3). This iPhone app records face animations in 100 Hz and writes it into a file, so the activation of the different facial blend shapes can be used later on for rendering. We implemented functionality in our *Character Plugin* (see App. C.3) to easily load and replay these files onto ECAs. Additionally, we added the possibility to select a specific tracking frame as reference pose (neutral) so only the differences to this frame were used as activation values for the blend shapes. This functionality was later on also natively added to the Live Link Face app.

[2] `https://www.unige.ch/cisa/gew/`

[3] `https://developer.oculus.com/downloads/package/oculus-lipsync-unreal/`

[4] `https://apps.apple.com/us/app/live-link-face/id1495370836`

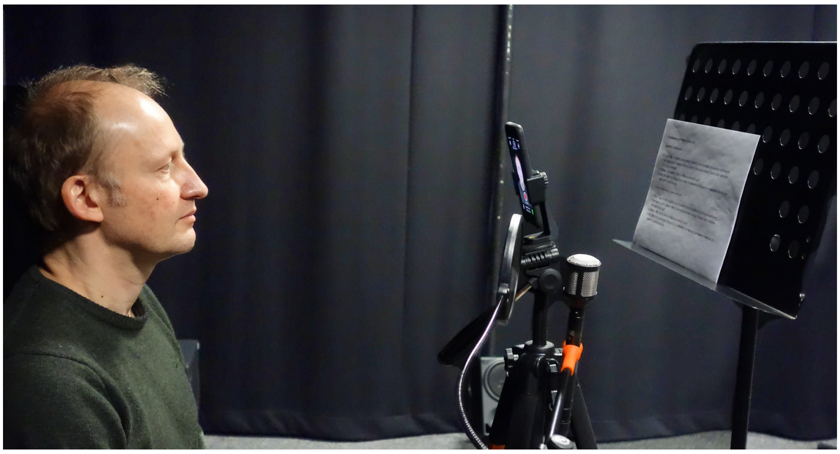


Figure 5.3.: A voice actor speaking his part in the dialogue for the voice and prosody study shown in Sec. 4.1 being recorded with an *AKG C451E* microphone with pop filter and an *iPhone SE* for facial tracking in an acoustically dry environment.

5.2. Gazing

For ECAs gazing is a particularly important behavior. In contrast to 2D representations, e.g., on a computer screen, where the Mona Lisa Effect (see, e.g., [Horstmann and Loth, 2019; Kum et al., 2024]) can elicit the feeling in multiple observers to be directly looked at, we have to carefully steer where ECAs are looking in VR since rendering is adapted if the user moves their head. Previous research has shown that human observers are well capable of quickly estimating where in the IVE an ECA is looking (see, e.g., [Loth et al., 2018; Langton et al., 2000]) and extensive research has been conducted on implementing natural gazing for ECAs [Ruhland et al., 2015]. Following Heylen [2006], gaze thereby serves two functions at the same time: observing the world and constituting a behavior that is observed. This is especially evident in mediated communication, where two humans communicate in VR being embodied in avatars. Here it was found that reproducing the actual gaze patterns of the users significantly improved realism [Andrei et al., 2023] and also the fluency of the conversation [Garau et al., 2003]. For ECAs, Heylen et al. [2005] found that natural gaze behaviors (looking at the interactant and away periodically) is preferred over starring agents and Wang and Gratch [2010] found that just staring at the user is not a feasible strategy to establish mutual attention and was rated as bad as an ECA constantly ignoring the user. Analogously, Andrist et al. [2013] found that gaze aversions can signal intimacy, thinking, or manage turn-taking [Jokinen et al., 2013]. Furthermore in scenarios with moving ECAs, gaze pattern have to be adapted, for example, alternating between the walking direction and the interaction partner [Bönsch et al., 2024].

Beyond that, gaze can be an important cue to infer and direct attention. Kompatsiari et al. [2022] showed that ECAs are capable of guiding the attention of human interactants using only gaze with similar results being presented by, e.g., [Bailly et al., 2010] and [Hartz et al., 2021]. Considering human gaze patterns, Huang et al. [2015] showed that it is possible to predict the desire for specific items from gaze even before a speech request for a specific item is uttered.

Furthermore, Koleva et al. [2015] showed that gazing at objects while listening is an important listener feedback mechanism, so that a human user feels understood. Narang et al. [2019] used gazing adaptations with moving ECAs to communicated the willingness to interact derived from the gazeing behavior of the user. In particular, carefully designed gaze patterns can be used to implement ECAs, that elicit the feeling of attentive listening (see, e.g., [Heylen, 2008; Lala et al., 2017; Oertel et al., 2021]) or ECAs can use gaze cues of the human interactant to adapt their behavior accordingly (see, e.g., [Grillon and Thalmann, 2008]). Along these lines Mejía et al. [2023] developed rules for ECA gazing when the user approaches to touch the ECA, finding that first looking at the approaching hand and then into the user's eyes is the best strategy.

However, as stated before gaze can also be influenced by far more factors, for example, humans tend to avert their gaze on high cognitive load [Doherty-Sneddon and Phelps, 2005] and coordinate their gaze behavior with laughter [Maraev et al., 2023]. Additionally, gaze behavior can also express emotions, like fear or anger [Sander et al., 2007]. When observing an ECA, human gaze behavior was further found to changed based on the ECA's ethnicity and gender [Huang et al., 2024]. When implementing ECA gaze behavior, research by Roth et al. [2018a] shows that gaze patterns that are based on a social model are preferred over random behavior and Kyrlitsias et al. [2020] found it to have a significant effect on social presence.

The main part of gaze behavior is the execution of gaze shifts, i.e., saccades, (e.g., [Lee et al., 2002; Andrist et al., 2013]) and the coordination of eyes, head, and torso movements (e.g., [Heylen, 2008; Pejsa et al., 2015; Hendrikse et al., 2018; Sidenmark and Gellersen, 2019]). These can be produced using statistical models (see [Ruhland et al., 2015]) or using data-driven approaches. For the latter, a recent approach by Goude et al. [2023] uses the saliency of the virtual scene rendered from the perspective of the ECA to automatically generate plausible gaze patterns. A similar approach was already proposed by Itti et al. [2006], where they recreated gaze behavior of a human watching a 2D video and Deng et al. [2005], who used texture synthesis to generate gaze paths. But sometimes gaze patterns are also not directly linked to focusing on specific things in the environment, but rather based on other behavior, for example, when gazing around the room while telling something. For example, Le et al. [2012] generated head, eye and eye lid movement based on a speech signal using a data-driven approach and Ferstl [2023] trained a neural network to generate head animations to look at something based on the current emotional state that should be conveyed. Additionally to these saccades, believable eye blinks should be produced so that the gazing looks plausible [Trutoiu et al., 2011].

To allow the ECAs developed for the study presented in [Ehret et al., 2023] (see Sec. 5.5.2) to show appropriate gazing behavior, first we implemented a general rule-based gaze controller that allows the ECAs (in our case visualized by *MetaHumans* or *Character Creator 3* models) to fixate any point in 3D. Therefore, we rotate the eyes, head, and upper body towards the gaze target following the movement dynamics described in the work by Pejsa et al. [2015]. The eyes and head start to move first and the eyes always move all the way to the target The torso starts to move delayed, with a delay time that is linearly dependent on the angle of the movement (see [Pejsa et al., 2015]). We also allow for alignment factors to be defined, so that the head can be fully rotated to face the gaze target or be rotated, for example, only 50%

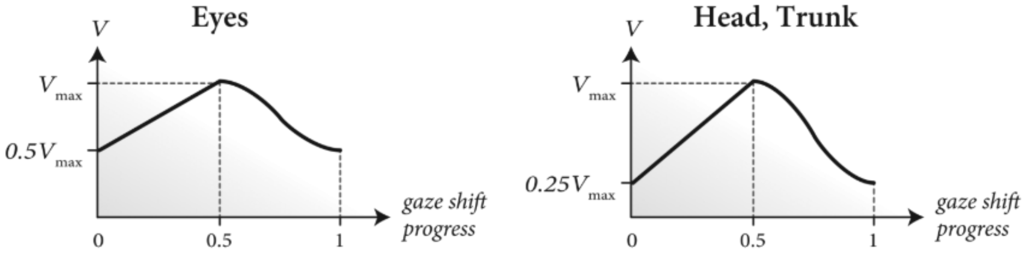


Figure 5.4.: Angular velocity profiles of different body parts involved in gazing, as shown in [Pejsa et al., 2015].

from its forward direction towards the gaze target, so that the eyes will still have a rotation relative to the head. During extensive testing 90% proved to be a reasonable default value for the head alignment and 10% for the torso (see Fig. 5.5) for the stationary scenario we used this (see Sec. 5.5.2). Note that these numbers are not added but are always relative to the default forward direction, so the head has to move less relative to the torso, if the torso has a higher alignment factor. Although [Sidenmark and Gellersen, 2019] report that gaze shifts with angles below 25° tend to be performed by eye movement only, this model looked plausible in our scenarios. We based the movement velocities on those described in [Pejsa et al., 2015], however, increased them partially as the produced movement otherwise looked too slow and thereby unnatural. In their model (see Fig. 5.4) the angular velocity is linearly increased from a start velocity of $f_{\text{start}} \cdot V_{\text{max}}$ to a maximum velocity (V_{max}) during the first half of the movement. And then decreased back to the start velocity using a polynomial function. Since the polynomials in the second half resemble quadratic easing in and out as implemented in Unreal by `FMath::InterpEaseInOut()`, we used this instead for simplicity. The values used are:

Body Part	$f_{\textbf{start}}$	$V_{\textbf{max}}$
Eyes	0.5	$300°/s$
Head	0.25	$150°/s$
Torso	0.25	$60°/s$

Table 5.1.: Parameters used to compute the angular velocity profile, see Fig. 5.4.

To produce physiologically plausible movements we restricted the movement to the physiological motor ranges, so for example the eye can only move up to 35° horizontally and up to 10°, while the head can be rotated up to 50° horizontally and 30° vertically. These values were estimated from the movements the used Character Creator 3 models could perform, while still looking natural. Given these constraints not all gaze targets are possible to fixate, especially in the back if the torso only rotates up to 10% towards the target. Therefore, we compute in each step how far the eyes are aligned with gaze target once they finished their movement. In case that they cannot reach the target due to rotational constraints, we alleviated the alignment

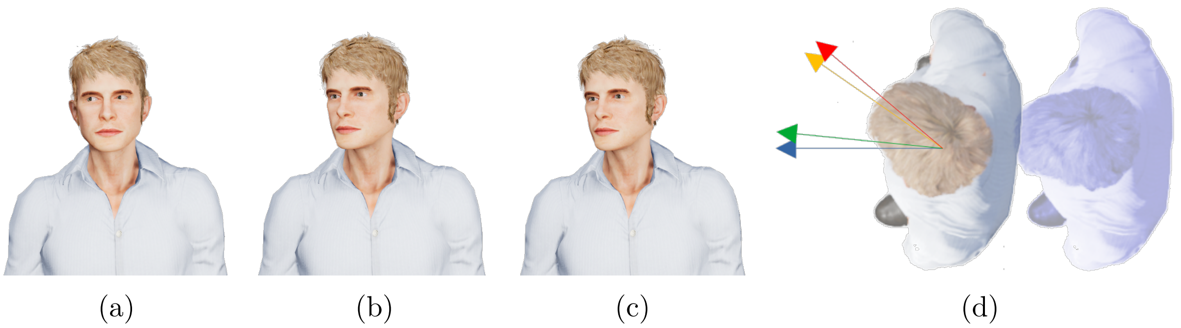


Figure 5.5.: Gazing Alignment: Gazing with eyes only (a); additionally aligning the head 90% with the gaze target (b); also aligning the torso 10% with the gaze target (c); top-down view of (c) with the forward direction in blue (+ blue silhouette shown behind with no alignment applied), the torso orientation in green, the head orientation in orange and the eye orientation in red.

factors of head and torso. First trying whether with a further rotation of the head (within its motor range) the target can be reached and after that whether the torso should be rotated further as well if the added head rotation did not suffice. Furthermore, the ECAs are normally not perfectly still but use idle movement so there is a continuous sway that has to be counteracted. Additionally, we wanted to support moving targets, like the users themselves. This proved difficult during the implementation, since the skeletal animation is not updated in the main thread of the Unreal Engine, but runs in a separate thread. To avoid oscillation due to using outdated data in the computations, we had to move this computation into the animation system by implementing custom “animation nodes”[5]. But this also results in unknown durations for single saccades. Therefore, we updated the current angular velocity (see Fig. 5.4) not by the gaze shift progression time but by the amount of angular distance covered compared from the original angular distance towards a new gaze target. Convergence of both eyes towards a gaze target as described in [Pejsa et al., 2015] is automatically achieved in this system, as the eyes are individually rotated towards the gaze target. For MetaHumans, which use blend shapes for animation of the eyes, the eye ball rotations have to be converted into activation values, using linear interpolation[6]. These blend shapes then also move the eye lids accordingly. For facial setups where eye balls are controlled via bones (e.g., Character Creator 3 models), this has to be done manually, so especially closing the eye lid according to downward gazing (see Fig. 5.6).

To also allow natural gaze patterns when the ECAs do not have specific locations in the scene to fixate, we implemented the saccade patterns described by Lee et al. [2002]. This means that the the ECA looks forward (or potentially at someone or something) and periodically averts

[5] The actual skeletal animation in Unreal is entirely done using animation Blueprints. They use the visual programming paradigm of the Unreal Engine, which uses nodes representing computations with data connections in between.

[6] For MetaHumans a blend shape activation of 1.0 corresponds with 25° rotation in either direction, so up/right/down/left

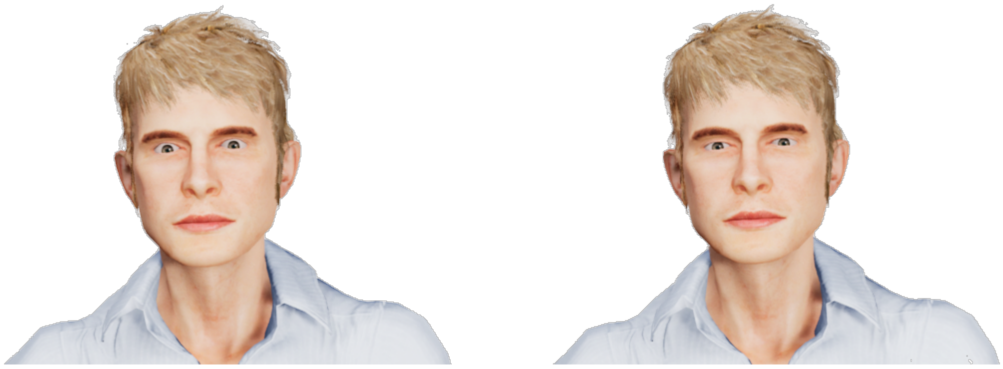


Figure 5.6.: Gazing down by the same angle without adapting the upper eye lid (left) and with appropriate adaptation (right).

the gaze by means of eye rotations only. Lee et al. [2002] provide statistics of saccades during two different settings: speaking, listening. These provide the mean and standard deviation for the length of direct/forward gazing and for gaze aversions and additionally statistics about the distribution of aversion directions. We added a third saccade mode for thinking inspired by the implementation and parameters used in *smartbody* [Thiebaux et al., 2008]. Furthermore, human eyes are never completely still but perform so-called microsaccades, small saccades of ~0.4° at ~1.4 Hz, which are barely noticeable (see, e.g., [Martinez-Conde et al., 2004; Troncoso et al., 2008]. These are also performed while fixating and therefore we use the statistical distributions of those described [Troncoso et al., 2008] by default to add this liveliness. As the ECAs are supposed to sometimes look directly at the user, we furthermore implemented periodic switching of which user eye the agents look into as is natural when humans look each other in the eyes. To this end we do not use the HMD position as gaze target when the ECAs look at the user, but estimate the positions of the eyes, left and right of the HMD position, and randomly switch every few seconds ($M = 2.0\,\mathrm{s}$, SD = 1.0) between those.

As last eye behavior also natural blinking has to be implemented for the ECA to not seem off. We added blinking following the statistics described by Bentivoglio et al. [1997] (20 ± 9 blinks/min, blink duration: 250 - 350 ms) using cubic ease-in/out, where the opening takes twice as long as the closing of the eyes. Trutoiu et al. [2011] describe that blinks also naturally occur when performing larger saccades. Therefore, we evoke a blink (resetting the timer for a next blink) every time a new gaze target is specified that causes a gaze shift of more than 10°.

We tested this rule-based gaze controller in several applications and used it during several studies with good results. We did not evaluate gaze behavior directly, but there were no concerns uttered regarding the naturalness of the performed gaze movement during post-study free-field comments. However, as shown by Ferstl [2023], who found different head animations to be depending on the emotion to be conveyed by the ECA, the exists no one-fits-all solution to gaze animations. Therefore, these parameters have to be changed according to the specific context so that the movement (speed) looks appropriate, which is easily possible within our presented plugin. We also found the alignment parameters to not be capable of generating all wanted behavior. For example in conversation, we often face our conversation partner by means of head rotation and then do gaze aversions only using the eyes while keeping the head fixed (potentially in a non-frontal direction). For this case we added the functionality to define separate gaze targets for the head and the eyes, to give maximum flexibility during application development. Furthermore, there is a large inter-personal difference in gazing behavior, as different people tend to use very different gaze patterns, for example, during conversation. So the presented parameters can only be understood as a reasonable starting point, which potentially requires adaptation as needed. But during the implementation we put special focus that these parameters can be easily adapted per ECA instantiated in an IVE.

We further direct the interested reader to [Ruhland et al., 2015] and [Admoni and Scassellati, 2017] for additional reading about implementing gaze behavior.

5.3. Gestures

The most obvious aspect of ECAs' co-verbal behavior is the use of co-verbal gestures (see [Wagner et al., 2014]). Following McNeill [1992], these can be grouped in four different categories:

- **deictic gestures** pointing towards references in the (virtual) world
- **iconic gestures** representing objects or concepts like forming a ball with one's hands
- **metaphoric gestures** visualizing a concept like "large" by moving the hands apart
- **beat gestures** following the rhythm of the speech while not transporting semantic meaning

However as stated by Neff [2016], producing fitting co-verbal gestures for a given speech act to be applied an ECA is non-trivial. Not only do they have to be contextually and temporally aligned with the speech, but they potentially convey information not present in the speech or even alter the meaning of the speech act. Furthermore, human observers are very trained at deciphering co-verbal gestures, and are sometimes already unsatisfied with a human actor not acting out a role perfectly, so producing convincing gestures with an ECA is a hard problem [Neff, 2016]. For example Ferstl et al. [2021] found that degraded/robotic gestures can potentially have detrimental effects on the perception of ECAs' human-likeness. However, properly evaluating co-verbal gestures is, in and of itself, a fundamentally complex task. Here, Debarba et al. [2022] used the configuration transition method proposed by Slater et al. [2010] and found that for many participants already a non-perfect simple tracking system was deemed good enough when used to record co-verbal gestures for ECAs. Similarly, Wolfert et al. [2024] found that direct comparison is superior to rating the gestures afterwards using a questionnaire to rate appropriateness and human-likeness. For example, He et al. [2022] compared generated co-verbal gestures with simple idle movements during ECA speech and found a difference on participants' gaze behavior but not on the subjectively-rated perceived human-likeness.

Nevertheless, it was found that having virtual teachers gesture improves learning performance in vocabulary training (when iconic gestures were used) [Bergmann and Macedonia, 2013] and math [Cook et al., 2017]. While in an obvious application deictic gestures can be used to give better directions [Cassell et al., 2007], they can also be used to elicit higher emotional responses in human observers [Wu et al., 2014]. Additionally, Neff et al. [2010] demonstrated that gesture rate and amplitude can influence perceptions of extraversion of ECAs, while Thomas et al. [2022] explored how robotic movements further affect a broad range of personality traits. Ravenet et al. [2013] showed that various nonverbal behaviors, including gesture properties, can be used to express different interpersonal attitudes in ECAs, such as dominance or friendliness. Another important aspect is synchronicity of speech and gesture [Chu and Hagoort, 2014]. Leonard and Cummins [2011] conducted a study on perceiving the temporal shift of beat gestures. It turned out that a temporal backwards shift of the gesture was already negatively recognized at 200 ms. In the case of a shift to the front, however, it was only recognized when exceeding 600 ms. Interestingly, a slight forward shift of 200 ms was even described as more natural than the original by some participants. A comprehensive overview over co-verbal gestures can be found in [Wagner et al., 2014].

As shown above, adding co-verbal gesture to an ECA's speech performance is important and can fulfil very different functions. In general, co-verbal gestures can either be recorded or generated, for instance, through machine learning techniques [Wolfert et al., 2022]. While generated gestures may suffer from repetitiveness and potentially lack semantic depth, recorded gestures require extensive manual labor, for example, to set up and to remove unwanted artifacts. In the remainder of this chapter we will first detail different approaches for generating co-verbal gestures, which was not the preferred technique for this thesis as we opted for the higher control of using recorded gestures. After that we will briefly describe our method to record co-verbal gestures and present a study evaluating different gesture modifications of these recorded gestures and the influence on perceived naturalness.

5.3.1. Automatic Gesture Generation

One way of generating co-verbal behavior is to blend together predefined gestures using a rule-based system (e.g., *BEAT* [Cassell et al., 2001b]). This was further developed in the *SAIBA* framework [Kopp et al., 2006], which allowed to specify situations and derive intentions for specific ECAs which communicated these by means of realizing predefined behavior. This was further refined in [Marsella et al., 2013] and [Lhommet and Marsella, 2013], who added, for example, more in-depth prosodic and rhetorical analysis to trigger appropriate behavior. It was further extended in [Ravenet et al., 2018] by adding more communicative and metaphorical gestures into the execution. All of the aforementioned utilize the *behavior markup language* (BML) to decouple behavior planning and behavior realization. By this, they rely on existing gestures that are combined to create a speech performance. Similar approach also use existing gestures but instead of using rule-based decisions on which gestures to use, they utilized prediction models, to see which class of gestures would be most appropriate for the current utterance (e.g., [Chiu et al., 2015; de Coninck et al., 2019]). Although predefined gestures are used, these can be slightly adjusted to, for example, move the hands to specific positions or within given physical constraints (see, e.g., [Xu et al., 2014]). Another approach was taken by Neff [2022], who manipulated existing gestures based on perceived muscle tension, to generate different expressive outcomes. A review of realising speech performances using existing gestures can be found in [Jung et al., 2011].

Another possibility to generate co-verbal gestures dynamically is to use mimicry or the so-called "chameleon effect" where the ECA copies the movements of a human user with a given time delay (see, e.g., [Roth et al., 2018b]). This will, however, be addressed again in Sec. 5.4, which is concerned with the production of back-channels.

More recently these systems using existing gestures, were substituted by data-driven approaches, where the movement is generated for each utterance individually. These systems firstly relied on the speech prosody to mainly produce fitting beat gestures (see, e.g., [Chiu and Marsella, 2011, 2014; Bozkurt et al., 2016]) They already performed significantly better than just performing random beat gestures [Bozkurt et al., 2016]. However, they were not capable to generate gestures for the other aforementioned gesture categories since the needed context information is simply not present in the pure prosody. Therefore, the inputs were extended by giving a text transcript (e.g., [Ishi et al., 2018; Kucherenko et al., 2020]) to create co-verbal gestures better fitting the content of the speech. Thereby different neural network architectures are used, like auto encoders (e.g., [Kucherenko et al., 2019]), auto-regressive models (e.g., [Alexanderson et al., 2020]) or more recently generative adversarial networks (GANs, e.g., [Habibie et al., 2021; Bhattacharya et al., 2021]). Also the inputs used for these models and thereby the focus of the produced gestures vary between the approaches. For example, Alexanderson et al. [2020] added style input like the handedness or radius of the gestures to be produced as additional input, while Yoon et al. [2020] focused on being able to specify different speaker ids and generating gestures with distinctive styles based on these. Furthermore, Bhattacharya et al. [2021] focused on reproducing different affective states of the speaker and Voß and Kopp [2023] designed a system which is particularly trained to generate appropriate metaphoric and deictic gestures. For the latter, the system, for example, needs to know reference positions in the space around the ECA to be able to correctly point to those. Another direction is evaluated by Krome and Kopp [2023], evaluating trade-offs between latency and quality to produce gestures in real-time, for example, if the speech content cannot be specified beforehand and is generated on-the-fly. A recent approach by Liu et al. [2024b] generates facial and full-body animation simultaneously and especially allows to generate fluent and appropriate motion in between specified passages or poses, yielding flexibility for the developers. However, there also exist solutions tackling very specific motions, like synthesizing head motion that fits the emotional state of a speech act [Sadoughi et al., 2017], or movements matching a musical presentation on the piano or a violin [Shlizerman et al., 2018]. For a further overview of gesture generation we refer the interested reader to [Wolfert et al., 2022].

5.3.2. Gesture Recording

However, to this date the best performances can still be generated by capturing the motion of a real human (see [Li et al., 2019b]), including potential manual post-processing. Thereby room-scale opto-electronic tracking systems are used most commonly, but also small systems relying only on tracking technology worn by the actor are possible (see [Zhou and Hu, 2008]), albeit with potentially not as good results. We already discussed the different tracking possibilities in Sec. 3.4. For this work, when gestures had to be recorded, we utilized the 6-tracker setup explained there, using the HMD, *Valve Index Controllers* (capable of rudimentary finger tracking) and 6 additional *Vive Trackers* being worn at the ankles, arms, pelvis, and chest. However, obtained results contained some artifacts originating from the low cost tracking technology. Those were small jumps and glitches due to tracking inaccuracies, which were still visible after generating full-body animations from the raw tracking data using inverse kinematics. Therefore, we developed some techniques to improve recorded gestures by meaningfully manipulating the collected tracking data which will be presented in the following section (Sec. 5.3.3) alongside with an evaluation conducting a user study in Sec.5.3.4.

5.3.3. Gesture Manipulations

The contents of this section are based on and taken in part from work previously published in [Ehret et al., 2025b].

During the gesture manipulation process, we modify recorded gestures to enhance their quality and eliminate issues such as jumps and glitches. To this end, the manipulation process works on the raw tacking data of the used nine tracked entities and inverse kinematics only applied afterwards. When manipulating gestures to improve their naturalness and thereby the social presence of ECAs expressing those gestures, one should be aware that these manipulations can also change other dimensions. One key concept here is perceived personality. Smith and Neff [2017] showed the influence of different motion adjustments of the perceived personality, more specifically on the *Big Five* personality traits, which are: extraversion, conscientiousness, agreeableness, emotional stability, and culture. So improving the gestures to not show unwanted artifacts is also important for the cause of not portraying an unwanted personality, which potentially influences the responsive behaviors of users in VR [Patotskaya et al., 2023]. While we tested multiple approaches on improving the recorded data, we settled on two straight-forward main methods: *peak removal* and *smoothing*, which are presented in the following.

The idea of **peak removal** is to remove erroneous data (position or orientation) resulting from temporal inaccuracies during tracking. Therefore, outliers (see [Wang et al., 2019a]) are identified among nearly constant data points by calculating the (Euclidean or angular) distances between consecutive tracking points, termed steps. A candidate is flagged as an outlier if its distance exceeds 4.4 times the median distance [Wang et al., 2019a], which effectively identifies strong movements as proven through extensive testing. If the median distance is zero for calm gestures, the first non-zero value is used instead. For all of the identified candidates we have to decide whether they are indeed an unwanted peak or simply a large change during the tracked movement. Therefore at first, consecutive candidates are grouped, potentially describing a larger peak. Then the (angular) distance from the start to the end point of this

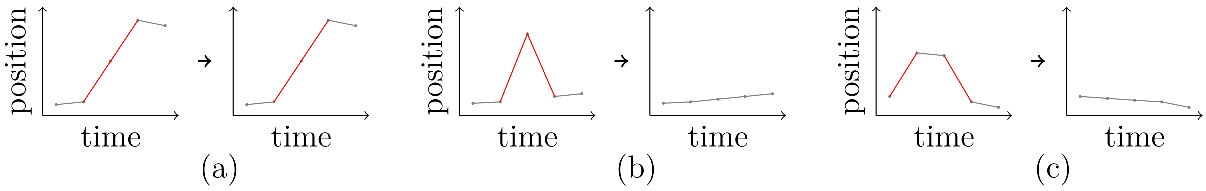


Figure 5.7.: Peak Removal Review: Recorded motion of a tracker with five individual tracking points. Two steps, defined as the positional change between two consecutive tracking points, above a distance threshold are identified as potential outliers and highlighted in red. No peak removal in (a), as the potential peak is recognized as fast but meaningful motion. Peak removal with linear interpolation in (b), as the potential peak is identified as measurement noise, and in (c) due to identifying an outlier plateau, which indicates an elongated deviation from the regular recorded tracker motion.

entire potential peak is considered. If this distance divided by the number of steps falls below the threshold defined above, this peak is considered an outlier and the values for all data points in between are discarded and recomputed by linear interpolation (see Fig. 5.7(b)). Otherwise, the data is kept (see Fig. 5.7(a)).

Since sometimes unwanted peaks can also form a plateau (see Fig. 5.7(c)), we enhanced this approach by adding the possibility to increase these potential peak region by one or more steps that were not classified as outlier candidates. For this, we introduced a parameter $t_{\text{extension}}$, which describes the time (and thereby number of steps) which can be added on each side of such a potential peak to possibly connect it to another potential peak. This, however, leads to multiple potential peak candidates adding different amounts of non-outlier steps to the region. In case multiple of these sets around the same data points detect a peak, we use a cost function to determine which set to use for peak removal. This cost function prioritizes sets with the least number of steps and, if multiple with the same number exist, with the smallest distance between start and end points. For peak removal a value of $t_{\text{extension}} = 0.001\,\text{s}$ proved beneficial during testing and was applied. Since the algorithm always rounds up, this means that only one additional step was considered to only remove severe glitches and not lose deliberate movement (given the used tracking rate of 30 Hz). While this parameter worked well for pronounced peaks (see Fig. 5.8), most recordings remained unchanged.

Therefore, temporal **smoothing** by means of averaging rotational and positional data with a sliding window was used. For this a window size $t_{\text{window}} = 0.3\,\text{s}$ proved beneficial, as proposed in [Motionbuilder, 2024].

While this temporal smoothing worked well for most body parts, the rudimentary finger tracking of the Valve Index Controllers occasionally resulted in unnatural hand postures (see Fig. 5.9). In these cases, only single fingers were curled or fully extended over a longer period, so that the above-mentioned methods did not improve it. Therefore, we complemented it with spatial smoothing, particularly for the fingers.. To this end, we computed the smoothed extension of individual fingers (excluding the thumb) as the means of their tracked extension and that of the neighboring fingers (see Fig. 5.9).

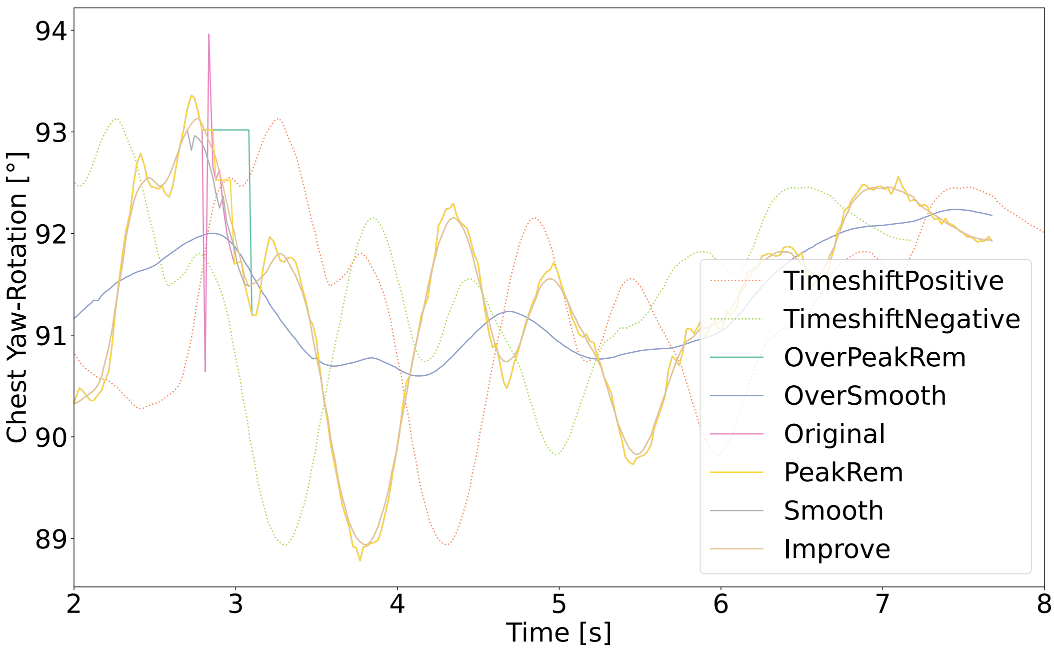


Figure 5.8.: Yaw rotation of the chest during one sentence of the HTR texts. The different manipulations can be seen in different colors. The manipulations shown are those used for the gesture manipulation study in Sec. 5.3.4. However, the originally recorded data is nearly entirely overlaid by the peak-removal data and the same is true for Improve and smooth.

To improve the recorded full-body gestures, we optimized the raw tracking data by first applying the peak removal with consecutive temporal and afterwards finger smoothing before feeding this data into an inverse kinematics solver.

In addition, we also wanted to degrade the gestures to emphasize effects in the gesture manipulation study (Sec. 5.3.4). Furthermore, these degraded co-verbal gestures will be relevant in a further study evaluating different degraders for ECA performances (Sec. 6.1). During pilot testing, we found that the above described methods could also be used for that purpose if parameters were exaggerated. We achieved an **over-smoothing** when setting $t_{\text{window}} = 2.0\,\text{s}$. This led to very slow and unnatural movement.

Furthermore an **over-peak-removal** could be realized by setting $t_{\text{extension}} = 0.1\,\text{s}$. Thereby, many short expressive movements can be misclassified as erroneous peaks that should be removed, and therefore the movement—that potentially is an important feature to characterize a specific gestures—is replaced by a short period with next to no movement, since the algorithm interpolates between start and end posture. This leads to a resulting movement that looks disjoint and “robotic”, partially lacking fluidity and continuity. As this is crucial for accurately representing natural movement, this results in severely degraded movement data. One examples of these manipulations can be seen in Fig. 5.8.

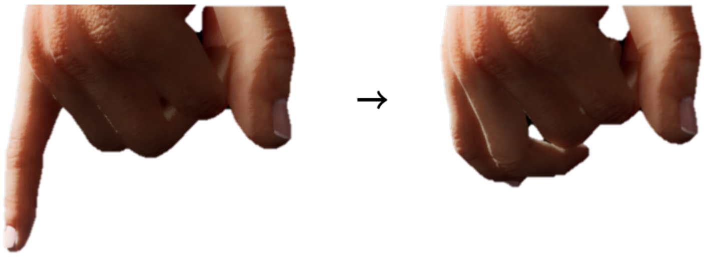


Figure 5.9.: Finger Smoothing: In the left image the raw finger tracking data is shown (with the pinkie unnaturally extended) and in the right image the smoothing result can be seen.

5.3.4. Gesture Manipulation Study

The contents of this section are based on and taken in part from work previously published in [Ehret et al., 2025b].

After introducing the different gesture manipulation/improvement techniques in the previous section, we present a user study with the research objective of evaluating whether the aforementioned improvements help in making the recorded co-verbal gestures more authentic and natural and whether they are preferred over the originally recorded data. Furthermore as a second objective, we also evaluate whether the degraded gestures are indeed perceived as worse. To put this into context with previous research, we also added time shifts as additional modifiers. Leonard and Cummins [2011], e.g., found a detrimental effect of gestures starting too late (further called positive time shift) while a shift of the gesture starting before the speech (further called negative time shift) in some cases even had a positive effect with regard to perceived naturalness. Following the results of Leonard and Cummins [2011] we used time shift by ±500 ms of the otherwise improved gestures. This duration was selected to provide a sufficient shift so that it is noticeable in the positive direction while being potentially even beneficial in the negative case. This resulted in seven different modifiers:

- M_{Smooth} and $M_{\text{PeakRemove}}$: the individual modifiers
- M_{Improve}: combining the two above
- $M_{\text{OverSmooth}}$ and $M_{\text{OverPeakRemove}}$: over-exaggerating M_{Smooth} and $M_{\text{PeakRemove}}$
- $M_{\text{TimeshiftPositive}}$ and $M_{\text{TimeshiftNegative}}$

These are complemented by M_{Original} using the raw tracking data for animation. Since all of the methods only clean up (or degrade) the 6DoF data of the seven tracking points, all of the modifiers use identical Inverse Kinematics to map the data onto the skeleton of the ECA.

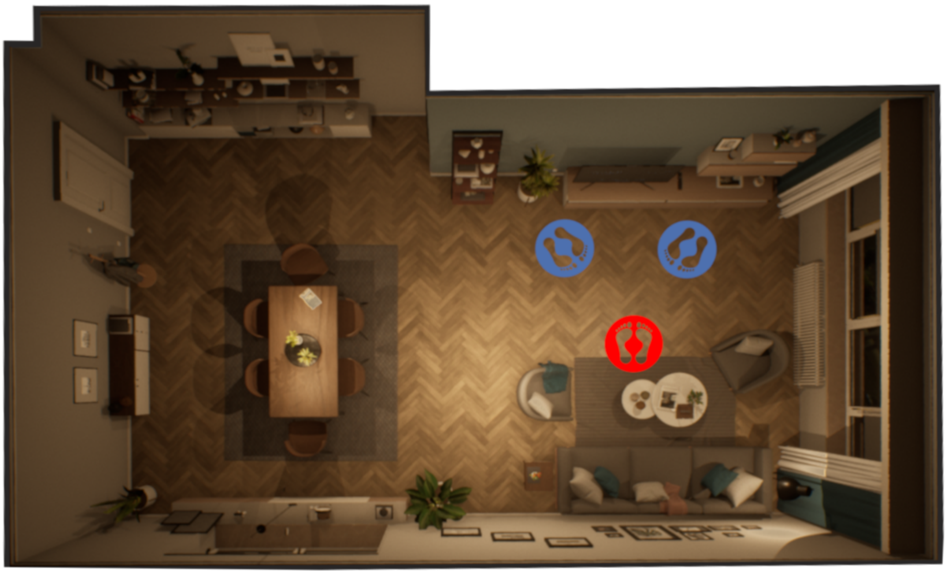


Figure 5.10.: Room setup used during the Gesture Manipulation Study. During the Listen Phase only one ECA was present on either of the blue positions, while during Comparison Phase the ECA was placed on both blue footmarks. Participants were asked to stand on the red footmarks, approximately 1.5 m away from the ECAs. Footmarks were not displayed during the studies.

Study Material

As speech material, we used the 34 family stories of the HTR paradigm [Schlittmeier et al., 2023], lasting approximately one minute per text (see App. A.1). While face tracking data was available and could directly be used to animate the ECA, full-body movements were missing. Thus, we recorded co-verbal movements for each sentence, using consumer components only as described in Sec. 3.4, namely a Vive Pro 2 HMD, two Valve Index Controllers, that support rudimentary finger tracking, and six Vive Trackers, which were attached to the feet, elbows, pelvis, and chest.

These recorded gestures were manipulated with the methods described in Sec. 5.3.3 before using inverse kinematics to apply them to the ECAs. The study was set within a virtual living room (see App. A.4.3) and as ECA models we used MetaHumans[7]. The study was conducted with an HTC Vive Pro Eye. Since the HMD obstructed the view of their own body, we embodied participants in the scene using a gender-matched avatar model from Character Creator 3. For this we used the avatar implementation described in Sec. 3.4 with controllers only (0-tracker setup), so without additional trackers. We utilized Unreal Engine 4.27 for rendering and the StudyFramework plugin (see Sec. 3.2) to facilitate the study procedure. Spatial audio rendering was done with Virtual Acoustics [Schäfer et al., 2023] (version 2022.a)

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

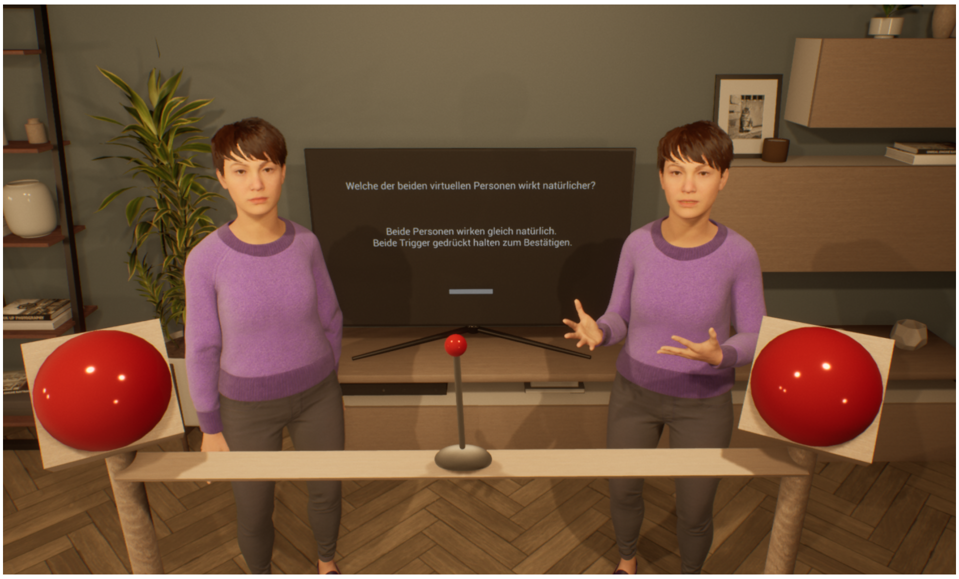
using the generic head-related transfer function (HRTF) of the IHTA artificial head [Schmitz, 1995] and a static directivity (see Sec. 4.2) for the virtual speakers.

Study Design

The study evaluated the influence of the different gesture manipulations on perceived naturalness, a prerequisite for social presence, and preference. It consisted of two phases: a *Listen Phase* and a *Comparison Phase*. The study was conducted in accordance with the Declaration of Helsinki and participants gave their informed consent and demographic information before putting on the HMD and calibrating their avatar.

During the **Listen Phase** participants had to listen to family stories being told by one ECA standing in front of them (see Fig. 5.10). Although the texts of the AuViST dataset can be presented by two speakers, we opted for just one speaker presenting the whole story to avoid idle and turn-taking movements to superimpose the effects of gesture manipulation. To that end, only one ECA (either female or male) was visible during each presentation. To cover up the main focus of this study and make participants carefully listen to the ECA, participants verbally answered the nine questions per text, which were displayed on the virtual television screen in the scene. Afterwards, they completed two questionnaires, which served as the measures for this phase, using Likert scales presented within the virtual scene Those were the *Anthropomorphism* dimension of the German version of Godspeed Questionnaire [Bartneck et al., 2009] (5 items, see App. B.7) and the *Coherence* construct from the ASA questionnaire [Fitrianie et al., 2022] (4 items, see App. B.11) in its German translation [Albers et al., 2024]. We conducted this phase in a 2×5 within-participant design. The first factor was ECA *Gender*, with the two levels G_{male} and G_{female}. The second factor was a subset of five *Modifiers*, namely M_{Original}, M_{Improve}, $M_{\text{TimeshiftPositive}}$, $M_{\text{OverSmooth}}$, and $M_{\text{OverPeakRemove}}$. We chose this subset of modifiers to reduce the number of conditions, expecting these to have the strongest influence. Conditions were counterbalanced in order between participants and each participant did two familiarization runs with the setup up-front, seeing both ECAs in the M_{Original} condition. Texts were randomly assigned to these conditions.

During the **Comparison Phase** two identical ECAs were placed in front of participants (see Fig. 5.11). Both ECAs spoke the same random sentence from an HTR text, while using one of the full set of eight modifier levels for the displayed co-verbal gestures. Participant could start either ECA by touching the corresponding button with the controller. This caused any other playing sentence to be stopped. Sentences could be played as often as desired until participants felt confident to answer the **question** “Which of the two virtual persons is more authentic/natural?” as only measurement in this phase. We varied the adjective (authentic or natural) as a between-participant factor to explore whether it made a difference. To answer this question participants had to pull the lever (see Fig. 5.11) to either side or keep it in the middle, if they evaluated them as equally authentic/natural. The lever always snapped to either of the three positions, closest to where it was released. This selection had to be confirmed with a dedicated button on the controller, in which case the data was logged, the scene faded out, and the next comparison was started. In total, each participant evaluated 36



Welche der beiden virtuellen Personen wirkt natürlicher?

Beide Personen wirken gleich natürlich.
Beide Trigger gedrückt halten zum Bestätigen.

Figure 5.11.: Participants' view in the Comparison Phase, asking "Which of the two virtual persons seems more natural?". By pressing the respective button, the left or right ECA spoke one sentence, including gesturing. Once decided for either ECA, the lever in the middle had to be pulled to the left or right, or kept in the middle if undecided. The decision was then confirmed by a controller button press.

comparisons, since each of the eight modifiers was paired with all other modifiers and itself exactly once. The order of these pairs was counterbalanced and whether female or male ECAs presented the sentence per modifier pair was randomized. After finishing all 36 comparisons, participants were debriefed and the study was over.

Results

In the study, $n = 33$ participants (22 male, 11 female) took part, who had a mean age of 31.5 years (SD = 14.4, 17-65 years). The one 17-year-old participant provided written parental consent. All participants self-reported (corrected-to) normal vision. One person participated in the study with hearing aids and all others reported normal hearing or at least no experienced limitations in everyday life hearing. Two participants (6.1%) were only fluent in German, while the others had German as their mother tongue (the whole study was conducted in German). All participants received an expense allowance of 15€ for approximately 90 minutes of participation in the study. All of the participants were able to complete the study without any significant problems (e.g., technical or cybersickness). A simulation-based a priori power analysis (using *simr*) [Green and Macleod, 2016] predicted a power of 82.50% (95% CI [80.00%, 84.81%]) for the questionnaire results assuming a medium effect (Cohen's $f = 0.25$) and a sample size of $n_1 = 33$. This suggests that the sample size of n_1 was adequate for the planned analysis.

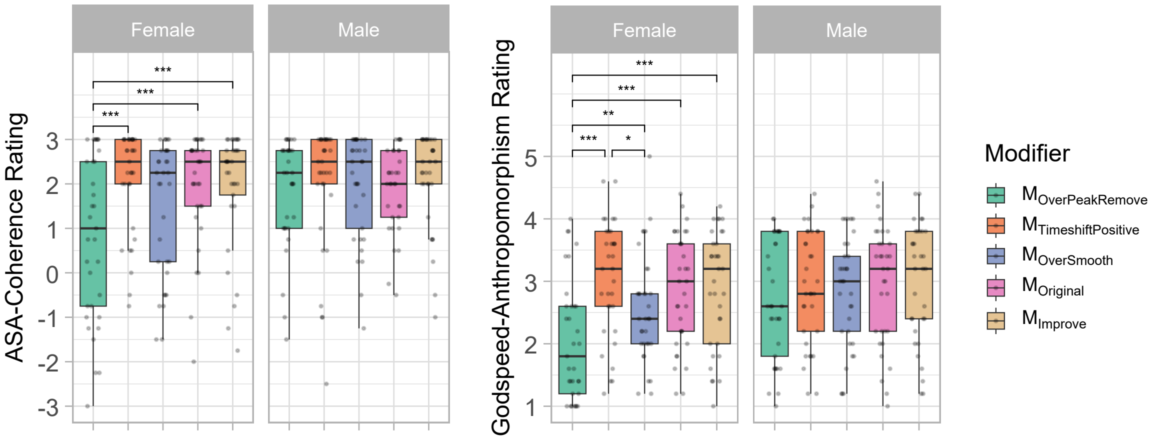


Figure 5.12.: Subjective ratings during the Listen Phase shown as boxplots, displaying median, interquartile range, and potential outliers, divided by speaker gender and modifier, with individual ratings shown as dots. Significant pairwise differences are shown as $***$ for $p < .001$, $**$ for $p < .01$, and $*$ for $p < .05$.

We evaluated the effect of the different modifiers using Linear Mixed-Effect Models (LMMs), implemented in *R* (version 4.4.1) [R-Core-Team, 2015] with the *lme4* package [Bates et al., 2015]. We chose LMMs over traditional ANOVAs because they offer greater statistical power and more accurately model individual-level variability [Mohanathasan et al., 2024]. To decide which factors to include, we performed backward model selection. During model selection, we used Likelihood Ratio tests to compare models by removing individual factors. Factors that did not significantly improve model fit, as determined by the Akaike Information Criterion (AIC) or did reduce it considerably ($\Delta_{AIC} \leq 2$) [Burnham and Anderson, 2004], were excluded (see supplemental material for details). Where applicable, we conducted post-hoc pairwise Bonferroni-corrected comparisons based on estimated marginal means using the *emmeans* package [Lenth, 2024].

For the results of the Listen Phase, we considered the two-way interaction of Speaker Gender (G_{male} vs. G_{female}) and Modifier (M_{Original}, M_{Improve}, $M_{\text{TimeshiftPositive}}$, $M_{\text{OverSmooth}}$, $M_{\text{OverPeakRemove}}$) as fixed factor and participant-ID and text-ID as random (intercept) factors.

For the **ASA-Coherence** ratings, which were rated on a scale from very low (-3) to very high (3), the factor text-ID was eliminated, and all others were kept. Results showed a significant interaction effect ($\chi^2(4) = 17.67, p = .001$) and significant main effects for both Speaker Gender ($\chi^2(1) = 9.62, p = .002$) and Modifier ($\chi^2(4) = 25.15, p < .001$). As we were not interested in comparisons between the two speakers, we conducted Bonferroni-corrected pairwise tests per speaker, yielding significant differences only for the female ECA from $M_{\text{OverPeakRemove}}$ to M_{Original} (t-ratio $= -4.80, p < .001$), $M_{\text{TimeshiftPositive}}$ (t-ratio $= -5.31, p < .001$), and to M_{Improve} (t-ratio $= -4.73, p < .001$) (see Fig. 5.12, left).

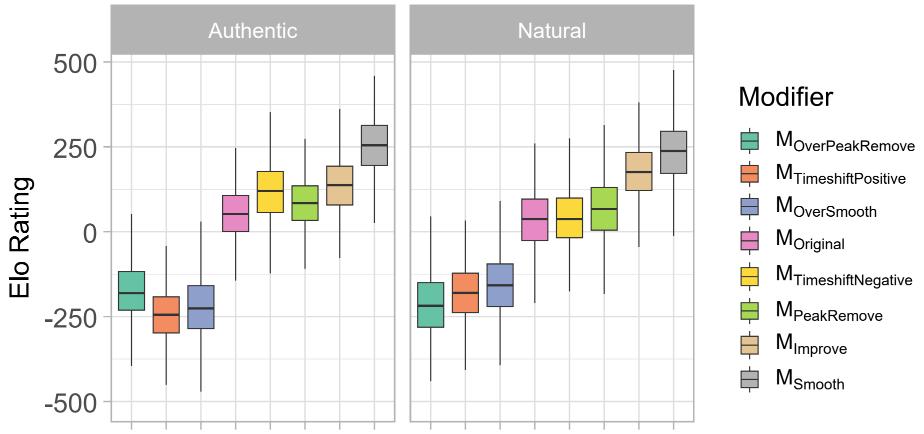


Figure 5.13.: Elo-Ratings of the different gesture manipulation methods one-to-one compared either asking for the most authentic or the most natural gesture. The resulting ratings from 100 randomly ordered runs each are shown as box plots, displaying median, interquartile range, and whiskers for the full range.

Analogously, the **Godspeed-Anthropomorphism** ratings, which were rated on a scale from very low (1) to very high (5), were analyzed. However, here text-ID proved beneficial for the model, so it was kept. Again a significant interaction effect ($\chi^2(4) = 14.51, p = .006$) and main effects for both Speaker Gender ($\chi^2(1) = 7.06, p = .007$) and Modifier ($\chi^2(4) = 40.25, p < .001$) were found. And again, only for the female speaker significant differences were found (see Fig. 5.12, right).

During the Comparison Phase, we gathered 36 ratings per participant comparing two (potentially identical) modifiers with regard to which is more natural/authentic. In total, these were 1188 comparisons. They were not forced-choice, allowing also indifferent responses. So we could specifically use the self-comparison pairs (i.e., where the same modifier was presented twice) as attention checks, expecting indifferent responses for those. Results, however, showed that of these eight checks, participants rated on average 3.24 (SD = 2.36) as non-equal. Therefore, we included an additional random intercept factor for attention in our model.

For evaluation, we used the *EloChoice* package as proposed by Clark et al. [2018]. This evaluation is based on the rating of chess players, which assigns each player a rating from which the probability of winning or losing against any other player, given their rating, can be derived. The ratings are then updated such that a player winning against a player with a higher rating gains more additional rating than winning against a lower-ranked player. However, these chess ratings are dependent on the order of the matches played. Therefore, EloChoice simulates multiple randomized orders of these comparisons, in our case 100. We ran these EloChoice simulations individually for both questions (authentic vs. natural) and test-wise also only for participants successfully identifying at least $min_{\text{Attention}} \in [0, 8]$ of the self-comparison correctly. Hereby, draws had to be excluded as the EloChoice package cannot handle those. However, when computing an LMM with a fixed two-way interaction of the Modifier (all eight levels) and Question (natural vs. authentic) and a random factor for $min_{\text{Attention}}$, we did not find a significant effect of including $min_{\text{Attention}}$, so we removed it again. When we compared this model with one not including Question, we found a significant difference, and kept the two-way interaction in the evaluation (see Fig. 5.13). With this model we found a significant interaction effect ($\chi^2(7) = 1418, p < .001$) and a main effect of the Modifier ($\chi^2(7) = 55881, p < .001$). Pairwise comparisons for the *authentic* attribute showed non-significant differences only when comparing $M_{\text{TimeshiftNegative}}$ to M_{Improve} (t-ratio $= -1.77, p = 1.0$) and to $M_{\text{PeakRemove}}$ (t-ratio $= -1.03, p = 1.0$), between $M_{\text{PeakRemove}}$ and M_{Improve} (t-ratio $= -2.81, p = .61$), and between $M_{\text{TimeshiftPositive}}$ and $M_{\text{OverSmooth}}$ (t-ratio $= -1.54, p = 1.0$). Furthermore, for the *natural* attribute non-significant differences only showed when comparing $M_{\text{TimeshiftNegative}}$ to M_{Original} (t-ratio $= 2.70, p = .85$) and to $M_{\text{PeakRemove}}$ (t-ratio $= -0.34, p = 1.0$), between $M_{\text{PeakRemove}}$ and M_{Original} (t-ratio $= 3.04, p = .29$), and between $M_{\text{TimeshiftPositive}}$ and $M_{\text{OverSmooth}}$ (t-ratio $= -0.04, p = 1.0$). All other pair-wise tests turned out significant (all $p < .001$) and ratings between different attributes were discarded since the absolute EloChoice values are arbitrary and the comparison therefore did not appear to be appropriate.

Discussion

When analyzing the results it is evident that over-smoothing and over-peak-removal performed worst, as expected, both in terms of rated coherence and preference with regard to naturalness/authenticity. $M_{\text{OverPeakRemove}}$ was also rated significantly worse with regard to anthropomorphism. Interestingly, this was only evident for the female ECA. This gender difference may be explained by participant feedback, which indicated that certain movements in the male ECA were perceived as particularly unnatural. Gestures for each ECA were recorded with a gender-matched performer. However, these unexpected movements could have altered ratings and were not altered by our manipulations. More research evaluating different genders but also personality traits [Smith and Neff, 2017] of performed gestures would be required to shed light on this observation. Nevertheless, we will use $M_{\text{OverPeakRemove}}$ as a deliberate social presence degrader when actively trying to reduce the quality of co-verbal gestures in Sec. 6.1.

As to the most promising gesture improvements, the questionnaires did not reveal any significant benefits of the used modifiers. The Elo ratings (see Fig. 5.13) however, show most significant gains by M_{Smooth} and M_{Improve}. Contrary to our expectation, smoothing the gestures was partly even preferred over additionally also removing peaks (here called Improve). This could be due to the limited effectiveness of our peak removal method, which only targeted single-sample peaks and often did not alter the movement at all (see Fig. 5.8). The parameters for this were chosen during development, testing them on individual, exemplary gestures. However, a repeated, more careful evaluation of parameter options could potentially yield superior results. Despite M_{Smooth} sometimes ranking slightly higher, M_{Improve} also removes potential peaks across the broader animation set, which potentially were not part of the evaluation set used in the Comparison Phase. Therefore, we chose to use fully improved gestures for all non-degraded gestures during the study in Sec. 6.1, During the Comparison Phase we used the attributes “authentic” and “natural” to see if they differ in outcome. While the Elo scores differ significantly between those, the general rankings remained similar. We argue that assessing naturalness alone might have been sufficient, given the similarity in outcomes across the two attributes. However, in light of the research by Roth et al. [2019b], perceived authenticity of such virtual entities is a needed prerequisite for pro-social behavior, so for natural interaction with them, as we would expect them from interactions with real humans. Also, it is important to notice that with both attributes the previous results of gesture time shifts [Leonard and Cummins, 2011] could be reproduced. This leads us to believe that manipulation techniques we chose for improving and degrading the recorded gestures can well be used in subsequent studies.

5.4. Back-Channel Generation

A more complex conversational behavior compared to aforementioned uni-modal ones is the production of back-channels. These are signals listeners generate while someone else is speaking to provide information about communicative functions, for example, whether they understand the speaker, potentially disagree with what is being said, need clarification, or simply encourage to go on talking [Bevacqua et al., 2008].

Back-channels can be conveyed by various cues. The most common ones are head nods and head shakes [Kendon, 2002; Bevacqua et al., 2007], where nodding can signal understanding and head shakes potentially the opposite. While this is the most likely interpretation, Leone [2012] also found in student-teacher interactions that nodding during listening sometimes does not represent true understanding but more a quasi-automatic behavior functional towards the fluency of the discourse. Another common back-channel cue are vocalizations, like “m-mh”, short words like “yeah”, or even repetitions of words or phrases just heard [Bevacqua et al., 2010]. Furthermore, also facial expressions like smiling, frowning or raising the eyebrows are important cues [Bevacqua et al., 2007].

However, especially for the above mentioned cues their co-generation, i.e., using signals from different modalities at the same time, can alter the meaning. Heylen et al. [2007], for example, states that a nod alone signals agreeing, while combined with raised eyebrows it is rather perceived as interest. And as observed by Bevacqua et al. [2010], meaning can even be inverted. They found combining the vocalization “really” with a nod signals agreement, while the same vocalization combined with raised eyebrows is perceived as disbelief. Similarly, Etienne et al. [2023] found that smiles were not always rated as positive, especially when being generated in sequence with other potentially negative back-channels. In their study participants, observing an audience generating back-channels, stated that they were not sure that the smiles were genuine and rather classified them as fake smiles. Another important back-channel cue utilized in this study was posture and especially posture changes, likes leaning in etc. (see also [Cassell et al., 2001a]). Further back-channel cues are given by gaze (as already discussed in Sec. 5.2). And also mimicry is often described as back-channeling that happens frequently in human-human conversations and has the potential to positively affect the social interaction and should therefore be implemented in listening ECAs [Roth et al., 2018b]. Mimicry, also described as the “chameleon effect” constitutes a behavior of mimicking the interlocutor’s behavior, delayed by some time. Most often gaze, head gestures, and facial expressions are mimicked [Maatman et al., 2005], for example, with a 4 s delay as in [Bailenson and Yee, 2005]. However, mimicry constitutes a more complex behavior than simply replaying user behavior, for example, Hasler et al. [2017] found that the amount of mimicry in avatar-mediated communication also depends on the visual representation of oneself and the other, in their case whether those avatars belonged to the same racial group.

While for mimicry the timing of the individual signals is given, for other cues this has to be carefully considered. Poppe et al. [2011] found that humans watching an ECA giving back-channel signals are more forgiving to nods being randomly placed than for vocalizations not being performed in accordance with the speaker’s speech act. Similar results were obtained by Inden et al. [2013], who observed that back-channel entrained with prominent vowels produced better results than random signals, albeit still not on par with recorded human behavior. Maatman et al. [2005] developed a set of rules when to produce what kind of cues, for example, performing a nod after a lowered pitch in the speech signal or after a speech disfluency or mimicking gaze behavior when the speaker gazes away for a longer period. This rule-based system was further developed in [Gratch et al., 2006], implementing the *rapport agent*, whose primary goal was to elicit the feeling of rapport between a human speaker and a listening ECA. Rapport describes the feeling of being “in sync” with the conversational partner. In a study they showed that, although the ECA was oblivious to the content of the speech, it was able to positively influence the overall impression of the communication. This system was further improved with regard to perceived naturalness by Huang et al. [2011], for example adding more facial expressions and using a data-driven approach. In general, there exists a large body of work trying to predict the generation of back-channels in human-human conversation. These predictors can then potentially be used to generate back-channels in listening ECAs. Here, Morency et al. [2010] and Huang et al. [2010] found their predictors to outperform rule-based systems. A shortcoming is that, for example, the predictor described by Gurion et al. [2020] is only capable of correctly predicting about a third of the back-channels correctly. However, this can also be attributed to the fact that back-channels are not deterministic and happen rather spontaneously. The prediction accuracy was nevertheless improved in recent years by Ekstedt and Skantze [2022b] using more intricate audio features or by Qian and Skantze [2024] using self-supervised learning.

Back-channels can, however, also be produced in response to speaker behavior beyond the speech itself, so sometimes it is necessary to react to certain back-channel-inviting cues, for example when the speaker is looking for evidence of understanding [Hjalmarsson and Oertel, 2012]. In general it was found that users interacting with a back-channel producing ECA themselves also produced more back-channels when they had the impression that the ECA understood and responded to their back-channels [Buschmeier and Kopp, 2018]. Back-channels can also be used to portray different personalities for the acting ECAs [Sevin et al., 2010]. Furthermore, back-channels can lead to a higher trust and liking towards the ECA [Aburumman et al., 2022] and also improve perceive life-likeness and comprehension [Stevens et al., 2016] and the judged empathy [Lala et al., 2022].

This can be important for cases where attentive listeners should be simulated, for example in therapeutic applications for elderly [Lala et al., 2017]. These models can further help to even out speaking times in conversation, where otherwise single speakers would have taken up an inappropriate amount of the overall time [Cumbal et al., 2022]. A further overview of back-channels with a special focus on head gestures can be found in [Heylen, 2008].

For our studies we implemented rudimentary back-channel capabilities, as the main focus of this work lies on designing speaking ECAs and back-channeling is not a particular focus of this thesis. This implementation comprised the usage of vocal fillers (“m-hm” and “aha”), which we specifically recorded and the capability to dynamically generate head nods. For this we used a simple cosine formula to compute the pitch of the head (Φ_{head}) as:

$$\Phi_{\text{head}} = -A \cdot \cos\left(2\pi \cdot t/d\right) + A$$

where the amplitude (A) is in our case a random number between 1.5° and 4.5° and the duration (d) of a single nod a value between 0.25 s and 0.75 s. Additionally, we chose to have an arbitrary number (n) of nods between 1.5 to 3.0. So t goes from 0 to $n \cdot d$ and additionally the pitch is adapted to linearly fade out the movement as:

$$\Phi'_{\text{head}} = \Phi_{\text{head}} \cdot (n \cdot d - t)/(n \cdot d)$$

This produces one or multiple smooth downward nods and the provided parameters proofed to produce natural-looking nods during internal testing. We also conducted a user study with 18 participants (mean age: 26.6 year, SD = 3.4, 13 male, 4 female, 1 diverse) evaluating these nods, vocalizations and gaze mimicry with regard to naturalness as part of the master thesis of Klara Tyroller, hypothesizing that listening ECAs’ are perceived as more natural when they exhibit such behavior. Participants, situated in a virtual seminar room together with one ECA, had to answer some open questions (e.g., “What is Aachen known for?”) to an ECA who’s back-channels were controlled by the experimenter. Adding nods and vocalizations significantly increased the perceived naturalness ($t(17) = 6.1, p < .001$), asked for by the questions “How realistic did you find the behavior of your counterpart?”. The mimicry, however, could not be reasonable evaluated since there was an implementation bug leading to indifferent behavior.

While in this user study, back-channels were triggered by an human-in-the-loop, we also used this nod generator in automatic contexts, for example in the study described in Sec. 6.1, to randomly generate nods at the end of sentences, combined with brief smiles as shwon in Sec. 5.1.2. Further back-channel signal were not implemented for this work. Additionally, the presented system could benefit from integrated automatic back-channel triggering, for example, base on the results of Qian and Skantze [2024].

As first observed and classifies by Duncan Jr. [1974] a special form of back-channels can be those signaling the willingness to speak up next. The following section will look deeper into this topic, namely turn-taking cues albeit in the following section focusing on the production of turn-taking signals by the speaker. However, Ishii et al. [2021] already stated that those are related and especially that incorporating turn-taking concepts in predictors for back-channel signals can significantly improve those.

5.5. Turn-Taking

The contents of this section are based on and taken in part from work previously published in [Ehret et al., 2023].

After examining speaker and listener behavior individually in the previous, we will now take a closer look at the communicational flow and the social cues involved. The ability to take turns in conversations is a fundamental aspect of human communication. However, it remains often unaddressed in exchanges with virtual interlocutors while leveraging different modalities of turn-taking cues can significantly improve the effectiveness and naturalness of such interactions. This becomes especially important if the natural language processing system generating the speech for the ECA produces longer delays, beyond natural gaps in conversations, to reduce ambiguities in turn-taking. While the use of VR also makes the use of artificial cues like virtual spot lights possible (see [Lee et al., 2024]) these do not leverage the learned social norms used in human-human conversation and potentially reduced the perceived social presence during interaction. Levinson [2016], who analyzed turn-taking from a cognitive science perspective, describes that the average gap between two turns (of around 200 ms) is much smaller than the time to plan the production of just a single word. This means that turn ends are subconsciously predicted, potentially involving more modalities than just speech. Skantze [2021] describes different modalities used in natural conversations to communicate turn-taking. Those are verbal cues (i.e., syntax, semantics, and pragmatics), prosody, breathing, gaze, and gestures. The goal in this section, however, is to derive a system producing co-verbal turn-taking cues, so only using the latter three of the aforementioned modalities. The rationale behind that is, that while the speech signal is often predefined (either scripted or for better naturalness even prerecorded, see Sec. 4.1), the co-verbal behavior of the conversing ECA is frequently generated. Furthermore, according to Skantze [2021], using gestures and breathing has attracted less attention when designing systems to regulate turn-taking. Since human turn-taking signals are ambiguous and sometimes lack clarity, we decided to use a rule-based system, not a data-driven approach. This system should produce clear and intelligible signals, while still leveraging the subconscious processing skills of humans in conversations. We produce co-verbal turn-taking cues specifically for ECAs in VR since the co-presence of the sender and recipient of such signals appears to play a crucial role in their effectiveness (cf. [Skantze, 2021]).

In this section, we will first provide an overview of related work, followed by a description of our derived implementation. Subsequently, we will present a VR study we conducted to evaluate the performance of our system and discuss the insights gained.

Related Work

One modality for giving cues about whether an interactant wants to continue speaking (turn-hold) or is willing to pass the turn on to someone else (turn-yield) is gazing behavior, first described by Duncan Jr. [1974]. There is a multitude of observation studies on how gaze is altered by humans during conversations to signal turn-taking (e.g., [Kendon and Cook, 1969;

Oertel et al., 2012; Dobre et al., 2021]). These observations are then used to predict who is going to speak up next in a conversation, for example using head orientation only [Rienks et al., 2010] or combining it with eye tracking to enhance accuracy [Ding et al., 2017]. Jokinen et al. [2013] found that eye gaze is especially useful to distinguish whether a speaker is taking a pause to think (turn-hold) or wants to yield the turn. Furthermore, this data is also used to derive gaze models which can then be applied to ECAs or socially-aware robots (e.g., [Bohus and Horvitz, 2010; Mutlu et al., 2012; Gillet et al., 2021]) which, for example, leads to fewer interruptions and thereby to a better conversational flow [Heylen et al., 2005].

Beyond that, Wagner et al. [2014] describe gestures as also playing a key role in signaling turn-taking. One important aspect here is that during spontaneous conversations, gestures often terminate before the end of speech when yielding the turn while they extend well beyond the end of the speech when holding the turn [Zellers et al., 2016]. Furthermore, posture shifts occur more frequently at discourse segment boundaries [Cassell et al., 2001a].

Several studies compare how combining different modalities improves the clarity of turn-taking. For example, prosody alone is not sufficient to predict turn-taking [Ruiter et al., 2006; Riest et al., 2015], and combining respiration and gaze yields superior predictions to using gaze alone [Ishii et al., 2015]. These result reproduced also the findings of Chen and Harper [2009] who found multimodal approaches to be superior for prediction. Recent approaches using artificial networks combine even more modalities, e.g., acoustic and linguistic [Roddy et al., 2018] combined with visual features automatically extracted from videos [Ishii et al., 2021]. De Coninck et al. [2019] chose the opposite way, predicting gesture classes and gaze targets from annotated conversational states. More recently Ekstedt and Skantze [2022a] proposed machine learning model named Voice Activity Projection (VAP), which focuses mainly on prosodic features of the speech to predict turn-taking and back-channels. Edlund and Beskow [2009] developed the *MushyPeek* framework, which deliberately manipulated avatar behavior in avatar-mediated communication. Due to these manipulations (e.g., changing gaze behavior or adding raised eyebrows), participants unconsciously changed their communication behavior. On the other end, Mills and Boschker [2022] even removed the speech altogether from a mediated communication and found that participants were able to establish turn-taking just based on observation of gaze behavior.

Furthermore, ECAs can communicate attitude [Ravenet et al., 2015] and personality [Maat et al., 2010] through their behavior when interrupting, which can also be used to shape turn-taking [Cafaro et al., 2016]. Similarly, Cumbal et al. [2024] evaluated different strategies of ECAs to recover from being interrupted by a user. However, turn-taking behavior can also be manually added to communication with an ECA in a Wizard-of-Oz paradigm to effectively influence the turn-taking during the interaction and create more natural intercourse (e.g., [Cassell, 2000; Devault et al., 2015]). Here, for example, Jégou et al. [2015] looked into how to do this by means of manipulating prosody alone. We refer the interested reader to [Ishii et al., 2016] and [Skantze, 2021] for further insights into the intricacies of turn-taking.

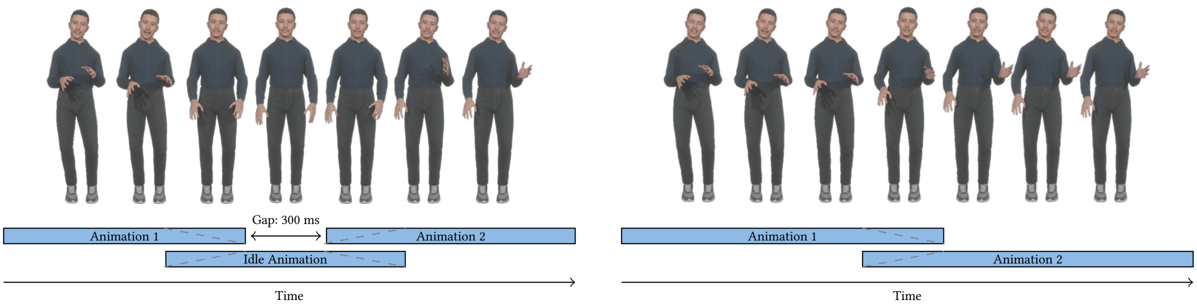


Figure 5.14.: On the left a non-held transition is shown, where we fade to the idle animation during the `Gap`. On the right side both animations are prolonged to perform a gesture holding when fading from one to the other. All poses are 300 ms apart.

5.5.1. Implementation of Non-Verbal Turn-Taking Cues

Following the findings of Skantze [2021], we based our implementation on three non-verbal modalities: *gaze*, *gestures*, and *breathing*. Due to the additive nature of turn-taking cues (cf. [Skantze, 2021]) we combined all three to give cues that are as clear as possible. We deliberately excluded syntactic, semantic, or prosodic turn-taking cues since we strove to implement a system that works with any speech material without a need for adaptation.

We structured each conversational act (i.e., a sentence being uttered by one speaker which might be followed by another sentence by the same speaker or a speaker-switch) in three phases, which will be treated differently when generating non-verbal behavior.

- `DuringUtterance`: From the start of the sentence up to 1 s before the end.
- `CloseToEnd`: The 1 second time frame at the end of the utterance before finishing the sentence. This time frame is chosen in accordance with the evaluation by [Ishii et al., 2015].
- `Gap` between two utterances, which can be uttered by the same speaker (*turn-hold*) or by different speakers (*turn-yield*). The `Gap` between two sentences of the same speaker or by different speakers is chosen to last by default 300 ms, which is approximately the median in real-life conversations (c.f. [Skantze, 2021]).

For each of the three used modalities, we generated behavior according to these phases. Thereby, we aimed for generating behavior patterns that resemble those observed in real conversations. However, since there are large interpersonal differences in these behaviors we tried to derive simple rules to implement a system that is easy to understand, leveraging our trained skills from human-human interactions. At the same time, we deliberately excluded all nuances and possible ambiguities observed in real conversations reducing some of the complexity.

Gazing

To dynamically adapt the ECAs' gaze, we first implemented a general gaze controller as described in Sec. 5.2. The gaze behavior is implemented for the use case of two talkers taking turns telling a story to one `Addressee`. Thereby the talkers always switch roles between `Speaker` and, while the other one is speaking, `Listener`. During the phases of the conversational act, we use different gazing patterns for the phases `DuringUtterance` and `CloseToEnd`. For the latter, we differentiate between holding the turn and yielding the turn to the next talker During the `Gap` the behavior of either `CloseToEnd` realizations is prolonged.
`DuringUtterance`: Following the observations by [Rienks et al., 2010] the `Speaker` divides his/her gazes equally between `Listener` (33%), `Addressee`(33%), and random gaze targets in the environment (33%, see Fig. 5.15). Also following [Rienks et al., 2010], the `Listener` gazes twice as much at the `Speaker` (67%) than at the `Addressee` (33%). Gaze durations are chosen from a normal distribution ($M = 2.27$ s, $SD = 2.4$), following [Ding et al., 2017], with a minimal gaze duration clamped at 1.0 s since smaller gaze lengths tended to look very unnatural.
`CloseToEnd(holding)`: Following the results of [Ishii et al., 2016], the `Speaker` looks at the `Listener` in 25.1% of the cases and breaks the gaze immediately in case the gaze becomes mutual. In our implementation, each mutual gaze is accordingly broken immediately during this phase by averting the gaze towards a gaze target in the environment. In case the previous gaze ends within this phase (it potentially extends further, see gaze duration distribution above), the `Speaker` looks at the `Listener` in 25.1% of the cases and otherwise averts gaze towards an environment gaze target. Heeding to the observations of [Ishii et al., 2016], the `Listener` looks towards the `Speaker` in 62.5% of the cases (if a new gaze target needs to be chosen) and otherwise simply extends the previous gaze during this phase.
`CloseToEnd(yielding)`: To show clear yielding behavior, the `Speaker` always looks at the `Listener`, who in this case is the next speaker. In Ishii et al.'s observation, the `Speaker` looks away in 25% of the cases if the gaze is not mutual [Ishii et al., 2016]. We, however, always keep the gaze at the `Listener` during this phase for clarity (again only changing the gaze once the previous gaze exceeded the minimal gaze duration of 1 s). Also for clarity, the `Listener` always looks at the `Speaker` in this phase (in [Ishii et al., 2016] this was only true in 62.5% of the cases) and averts the gaze immediately into the environment once the gaze is mutual (in [Ishii et al., 2016] this was only observed in 71% of the cases). This is in line with the findings by Oertel et al. [Oertel et al., 2012] in which incoming speakers tended to look away while the previous speaker tried to maintain a mutual gaze. Since the `Addressee` is never expected to take the turn, he/she is never looked at in `CloseToEnd`.

In most cases the `Addressee` is the user him-/herself, so we don't need to generate gazing behavior. However, to also cover cases in which one ECA takes over the role of the `Addressee`, we added a simplified model of always looking at the current `Speaker`, either virtual or human. Following [Wagner et al., 2014] listeners predominately engage using head nods when listening. Therefore, we designed the `Addressee` to produce nods at the end of each sentence of the other ECA, respectively end of turn of the participant, with a chance of 50% to seem more natural and involved.

Gesturing

Following the observations by Zellers et al. [2016], gestures should not terminate in the time frame of 500 ms prior to the speech end if the turn should be held beyond the following gap. Therefore, we manipulated the co-verbal gestures such that in the case of held turns the hands are fixed on the last accent/stroke of the animation before the `Gap` and prolonged by 300 ms into the gap. This way the hands hold the accent while the rest of the body still performs slight movements in a natural way. This animation is then blended together with the animation played after the `Gap`. Accordingly, the first accent/stroke is prolonged 300 ms forward, so that the animation does not blend back to an idle pose during the `Gap` – all co-verbal animations are by default played with 300 ms blend in and out from and to the looped idle animation – and the gesture is held during the `Gap` (see Fig. 5.14).

Breathing

As described in [Ishii et al., 2015] and [Skantze, 2021], respiration can be a helpful cue for initializing a turn but also for holding a turn. To this end, we extracted inhale audio sequences from the used audio material and replay a randomly selected one during the `Gap` for the ECA who is going to speak afterward. This is independent of the fact whether this is a turn-hold or whether the turn is passed on in the break, since – as common in natural conversation – the person speaking after the `Gap` needs to take a breath to have enough air for the following utterance.

5.5.2. Evaluation of Turn-Taking Modalities

To test whether the added non-verbal turn-taking cues are (subconsciously) perceived as intended, we conducted a within-subject VR user study (which constitutes a more realistic setting than the one of de Coninck et al. [2019]). We expected the following hypotheses to be confirmed:

H1 ECAs are rated as more socially present and natural when more modalities of turn-taking cues are shown.

H2 When participants take over an active role in turn-taking, gaps between turns decrease with more modalities of turn-taking cues being shown.

H3 When participants take over an active role in turn-taking, ECAs' behavior is rated less confusing when turn-taking cues are embedded.

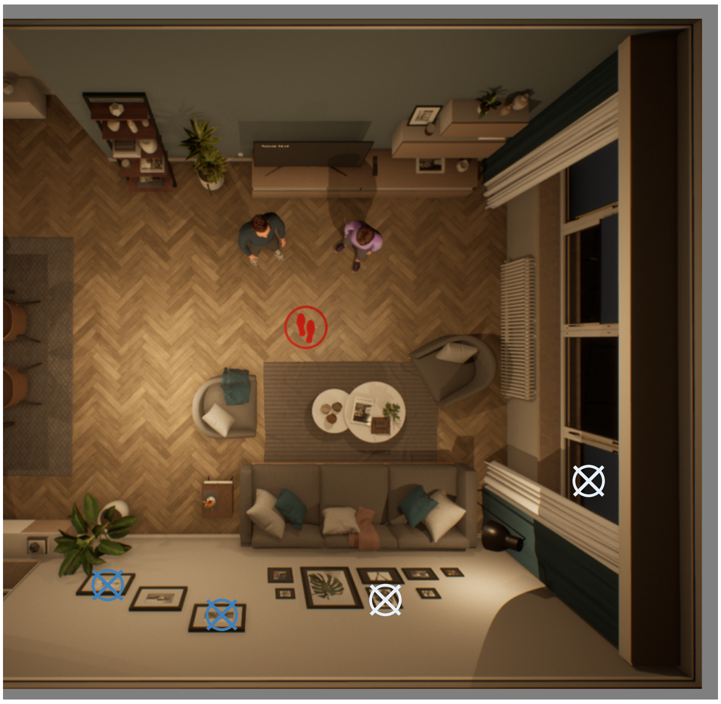


Figure 5.15.: Top view of the study scene. The participant stood on the red footmarks (which were not shown during the study). Environment gaze targets are marked for the female ECA (blue) standing on the right and the male ECA (white) standing on the left of the participant.

Material

The study took place in a virtual living room (see App. A.4.3) which is populated by two MetaHumans[8]. The study was rendered using Unreal Engine 4.27. The ECAs are positioned in front of the participant on both sides of a virtual TV screen, both at 30° and 1.5 m of the participant, facing him/her (see Fig. 5.15). For the gazing implementation, we defined additional environment gazing targets which were placed on sensible objects/locations in the scene (see Fig. 5.15). As speech content, we utilized family stories from the HTR task [Schlittmeier et al., 2023], which can be found in App. A.1. In the database [Ermert et al., 2022], suggestions for turn passes between two speakers are given, yielding 4-5 turn changes per text. The number of sentences spoken by one talker in a row is arbitrary while the sum of sentences spoken by each speaker is balanced. These texts were chosen as they originate from a verified paradigm, featuring compatible content complexity throughout the texts, and provide all the necessary information for this evaluation. Additionally, we posed the questions during the first study part, concealing the true purpose of the study, using attentive listening to the stories and recalling their contents as a plausible cover story. Furthermore, this had participants focus carefully on the conversation and thereby also on the non-verbal behavior. Accompanying the spoken texts, we we utilized the same recorded gestures as for the Gesture Manipulation Study (see Sec. 5.3.4), in this case in their original, non-manipulated version.

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

Apparatus

The study was executed on a desktop PC (Intel Core i9-10900X, 32GB RAM, GeForce RTX 3080 Ti). For the presentation a *Vive Pro Eye* HMD was used, which allowed for eye tracking during the study. Eye tracking was used to identify mutual gaze between the ECAs and the participant and also logged for further analysis. During the study, participants wore the 6-tracker setup described in Sec. 3.4, so that their movement could be transferred onto a gender-matching full-body avatar and additionally be saved for further analysis. The audio was replayed over *Sennheiser HD650* headphones using a *Focusrite Scarlett 2i2 3rd Gen* audio interface. The scene was auralized with Virtual Acoustics[9] using generic binaural rendering. A static directional filter of human speech was assigned dynamically to the speech sound sources (cf. [Ehret et al., 2020]). For simple study control *StudyFramework* (see Sec. 3) for Unreal Engine was utilized.

Study Design

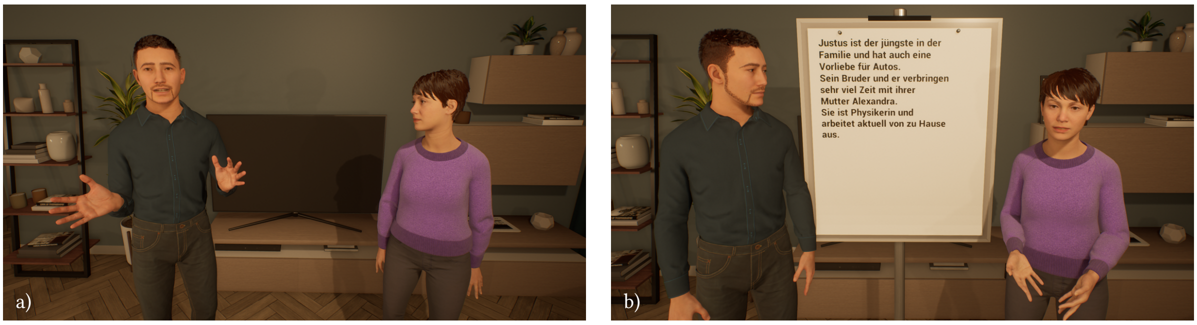
The study was split into two parts: `Listen` and `Act`. In the first part (`Listen`), participants listened to 10 family stories from the HTR being told by the two ECAs. In the second part (`Act`), participants took over a part in telling the stories while one of the ECAs represented the addressee. Thereby participants had to directly react to the turn-taking cues given by the ECA.

In both parts, five levels of the *Turn-Taking Cues* factor (T) are presented:

- T_{None}: no turn-taking cues are given
- T_{Breath}: only the breath cues are audible
- T_{Gesture}: only the gesture cues are shown
- T_{Gaze}: only the gazing cues are shown
- T_{Full}: all of the above are combined

When gazing turn-taking cues are not given we tried to generate similar gaze patterns, which, however, do not carry any turn-taking information. To that end we let the ECAs gaze at the other ECA, the participant, and gaze targets in the environment with equal frequencies, using the same gaze length normal distribution we used in the `DuringUtterance` phase (cf. Sec. 5.5.1). When gestures are not used as turn-taking cues, we still used gesture holding but at random gaps. So, if the ECA did not continue after the `Gap` with a randomly held gesture, that held gesture was interpolated into the idle gesture. The number of held gestures was kept

[9] `https://www.virtualacoustics.org/`



Justus ist der jüngste in der
Familie und hat auch eine
Vorliebe für Autos.
Sein Bruder und er verbringen
sehr viel Zeit mit ihrer
Mutter Alexandra.
Sie ist Physikerin und
arbeitet aktuell von zu Hause
aus.

Figure 5.16.: a) The two ECAs telling a family story as in the *Listen* Phase of the turn-taking study. The TV screen behind them is used to display the questions and further instructions. b) The female ECA taking a turn during the *Act* Phase, the flip chart shows the text the participant has to read out loud once being passed the turn.

approximately the same as in the conditions using them as turn-taking cues. Inhale sounds were omitted entirely when not used as cues.

Study Procedure

After reading a study description for the `Listen` part and giving their informed consent, participants filled out a demographics questionnaire and were equipped with the tracking hardware (HMD, Valve Index Controllers, six Vive Trackers), used for applying their motions onto their gender-matched avatar, and headphones. Once immersed, first a calibration of the avatar (see Sec. 3.4) and the eye tracking was performed. After that, the experimenter adjusted the voice detection threshold such that the HMD's microphone could be used to automatically detect participants starting to speak. Following that, participants undertook one training trial of the `Listen` part (always using T_{Full}). During the **`Listen`** part, a male and a female ECA (see Fig. 5.16 a)) told a family story (see App. A.1) while using different turn-taking cues to signal turn-taking. Participants were instructed to listen carefully to the stories. Once finished nine questions regarding the stories heard (e.g., *"How old is Vincent?"*) were shown on the virtual TV screen, which participants had to answer orally. The correct answer was presented to the experimenter, who had to log whether the right answer was given by the participants by means of button presses. When all nine questions were answered, a Likert-scale questionnaire assessing Social Presence was presented within the virtual environment. The participants had to point and click on the corresponding answer with the controller. The questionnaire included sub-scales from different questionnaires which we expected, if anything, to change due to the used manipulation. The underlying hypothesis is based on the observations in [Pütten et al., 2010] that higher social presence was found for ECAs exhibiting richer non-verbal behavior For *Anthropomorphism* the first construct of the Godspeed questionnaire [Bartneck et al., 2009] was presented where participants have to pick values on 5-point bipolar scales (e.g., between *Fake* and *Natural*) (see App. B.7). After that the constructs *Human-Like Behavior (HLB)* and *Agent's Coherence (COH)* from the ASA questionnaire [Fitrianie et al., 2022] were utilized, which had to be answered on a 7-point Likert scale (see App. B.11). Once answering those, the actual `Listen` phase started, repeating the same task as in the training trial 10 times. During these 10 trials, each of the five levels of T was presented twice. The presentation order of the turn-taking levels and the presented texts is counterbalanced using the Balanced Latin Square method. Participants were asked after each trial whether they wanted to have a break (as an additional field in the last questionnaire) and had to take a break of at least 5 min after completing all 10 trials. At the beginning of the break, a short questionnaire had to be filled out (at a desktop computer) asking for their general experience during the `Listen` part.

When feeling ready for the next part, participants had to read the study description for the `Act` part and conduct 10 trials of the `Act` part which were again foregone by a training trial. During the `Act` part, the spatial layout remained the same apart from a flip chart being placed between both ECAs. This virtual flip chart was used to present the text that had to be spoken by the participant, since in this study part the participants took over one part in telling the stories (see Fig. 5.16 b)). In this part 10 different stories were used than in the `Listen` part. While participants told the story with the ECA of opposite gender to their own, the ECA with the same gender took over the role of the `Addressee`. Participants were shown whether they take the first turn at telling the story and the sentences they have to speak next. However, when the ECA speaks they have no information on when to start and are therefore told to carefully look at the ECA to find out when to speak and then start speaking as quickly as possible. Using the HMD's microphone and a calibrated speech detection threshold, the start of a participant's utterance is recognized and the gap length since the end of the ECA's speech is logged. Once participants are done with their turn (i.e., they read the entire text currently displayed on the flip chart), the experimenter triggers the ECA to continue by means of pressing a button. Additionally, the experimenter logs any attempts to speak during the ECA's turn. If the participant does not start speaking for 3 s after the ECA is done, the ECA performs a dedicated turn-yielding gesture towards the participant. Once the full story was told we did not ask the related HTR questions but showed a virtual Likert scale questionnaire asking whether it was *easy* to understand when to speak up, whether the behavior of the partner was *confusing* or *ambivalent*, and whether the task was *frustrating*. All of the above were answered on 7-point Likert scales from -3 (*Disagree*) to 3 (*Agree*). Again, the 10 trials were counterbalanced. After finishing this part, participants had to answer a final desktop-based questionnaire and were compensated 15 € for their time. On average the study took 75 min, of which the immersed time for the `Listen` part was 31.9 min and 11.6 min for the `Act` part.

Participants

32 persons (21 male, 11 female) took part in our study. One female participant felt unwell during the execution and had to cancel the study. The remaining participants had a mean age of 25.9 years ($SD = 5.0$) and all self-reported normal hearing and normal or corrected to normal vision. Four participants (12.5%) were fluent in German while the others had German as their mother tongue (the whole study was conducted in German). Three of the participants (12.5%) never used VR before, seven (21.9%) only once before, 14 (43.7%) less than 10 times, and the rest (21.9%) more frequently.

5.5.3. Results

Data that is recorded per trial is analyzed using one-way repeated-measure ANOVAs with the single factor T (levels: T_{None}, T_{Breath}, T_{Gesture}, T_{Gaze}, T_{Full}). Data is checked before on normality using *Shapiro-Wilk tests*. Where the assumption of sphericity (evaluated with *Mauchly's test*) is violated *Greenhouse-Geisser Correction* is used when interpreting the ANOVA. When applicable paired-sample t-tests with *Bonferroni* correction are used as post-hoc tests.

Analyzing the questionnaires posed after each trial in the `Listen` part, we first confirmed the internal validity of the questionnaires by computing their *Cronbach's Alpha*, which were $\alpha = .95$ (*Godspeed*), $\alpha = .93$ (*HLB*) and $\alpha = .77$ (*COH*). Averaging the scores per turn-taking level for each participant and computing ANOVAs did not reveal any significant effects (all $F \leq 1.12$ and $p \geq .33$). On average the ratings for anthropomorphism (Godspeed) were $M = 2.7$ ($SD = 1.1$; from scale $[1, 5]$), for human-like behavior (HLB) $M = 0.6$ ($SD = 1.5$; from scale $[-3, 3]$), and for coherence (COH) $M = 2.3$ ($SD = 0.9$; from scale $[-3, 3]$).

Due to the fact that the number of texts used is a multiple of the numbers of levels of T, the balanced Latin Square counterbalancing always matched the same text to the same level of T. Although the HTR questions were primarily used as a disguise, we still planned to evaluate the performance in the HTR task. However, due to the above-mentioned shortcoming, it is not feasible to evaluate the answers given, since the texts and their questions might vary in difficulty, which might be confounded with experimental variation. In the questionnaire following the `Listen` part participants rated on a scale from -3 (*'Strongly Disagree'*) to 3 (*'Strongly Agree'*) that the ECAs sounded like humans in the real world ($M = 1.6$, $SD = 1.7$) but did not look as alike to humans in the real world ($M = 0.3$, $SD = 1.7$). Participants on average also stated that they noticed the ECAs signaling to yield or keep the turn ($M = 0.6$, $SD = 1.8$). However, also 19.4% rated this below or equal to -2.

A repeated-measures ANOVA (with Greenhouse-Geisser correction) revealed a significant effect of T on the gap participants left before starting to speak once the ECA finished speaking during the `Act` part, $F(3.04, 91.4) = 4.93, p = .003$. Post-hoc tests revealed a significant difference between T_{Breath} and T_{Gesture} ($p = .03$) and between T_{Breath} and T_{Full} ($p = .002$). There were also two non-significant trends between T_{None} and T_{Gesture} ($p = .10$) and between T_{None} and T_{Full} ($p = .10$), all other $p > .44$ (see Fig. 5.17).

We analyzed the four questions asked after each `Act` trial for internal consistency. We concluded to analyze the questions for *easiness* and the inverted answers to the questions whether the ECA's behavior was *ambivalent* or *confusing* together (Cronbach's $\alpha = .81$). This is called *Clarity* from here on and is the mean of the three aforementioned scales (*ambivalent* and *confusing* inverted). The question regarding *frustration* is evaluated separately since it would have reduced the Cronbach's Alpha score to $\alpha = .79$ and is differently framed. A repeated-measures ANOVA revealed a significant effect of T on Clarity ($F(4, 120) = 5.42, p < .001$). Post-hoc tests showed significant difference between T_{None} and T_{Gesture} ($p = .04$) and between T_{None} and T_{Full} ($p = .01$), all other $p > .18$ (see Fig. 5.17). For the *frustrating* questions, no significant effect was found ($F < 1.92, p = .14$), with the means per turn-taking level all

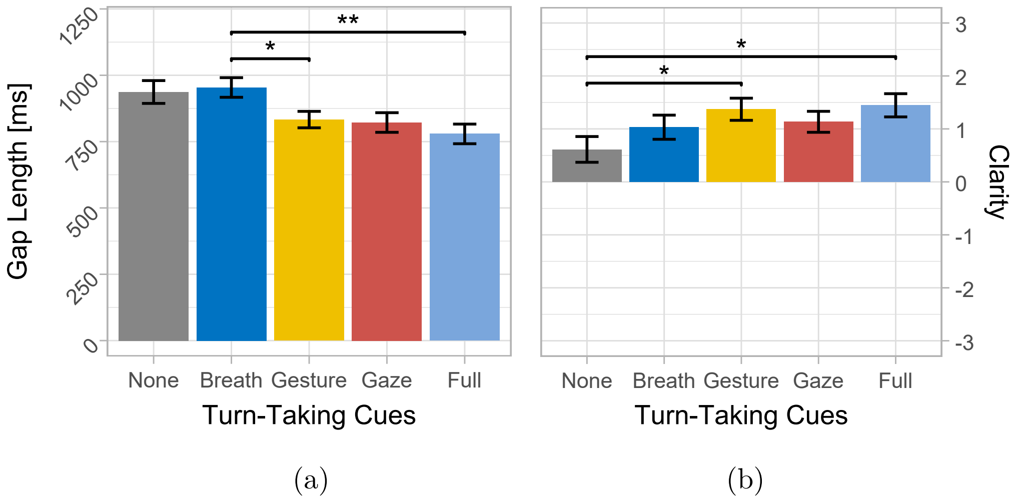


Figure 5.17.: Gap length (left) in ms and Clarity ratings (right) from a scale $[-3, 3]$ during the `Act` part. Error bars indicate standard error. Significant pairwise differences are shown as $**$ for $p < .01$ and $*$ for $p < .05$, all other differences are non-significant.

between -2.66 and -2.36. We also tracked whether participants tried to speak in a `Gap` when they should not. In sum this happened 21 times during T_{None}, eight times during T_{Breath}, 13 times while in T_{Gesture}, two times in T_{Gaze} and six times when all cues are shown in T_{Full} (of 651 gaps in total). However, a Friedman test (which is the non-parametric equivalent to a repeated-measures ANOVA and had to be used since the assumption of normality was violated), did not show a significant effect of T ($p = .20$). Explicit yield gestures (played after 3 s of silence) were in sum only triggered five times for different participants, so we did not analyze them further.

In the questionnaire following the `Act` part participants rated on a scale from -3 (*'Strongly Disagree'*) to 3 (*'Strongly Agree'*) that reading the texts was easy ($M = 2.4$, $SD = 0.7$) but, as expected, understanding when to speak was not as clear ($M = 0.7$, $SD = 1.3$). Furthermore, participants felt that the ECA in general reacted on them ($M = 1.3$, $SD = 1.86$), however, with a large inter-personal variability. Additionally, we gave a list of potential turn-taking cues from which participants had to select those they noticed. 80.6% noticed changes in gaze behavior, 51.6% in gesticulation and only one participant (3.2%) noticed audible inhalation. 25.8% noticed special gestures used by the ECAs. However, also 61.3% reported that they noticed changes in prosody, in speech speed (35.5%), or text content (41.9%), which we explicitly did not alter. Additionally to the options we provided, two participants (6.5%) reported focusing on the behavior of the `Addressee` and three participants (9.7%) that they looked out for long pauses. When asked which additional cues would have helped, the most prominent were mimics (19.4%), like raising the eyebrows, and special turn-yielding gestures (25.8%).

5.5.4. Discussion

When participants were only listening to the ECAs taking turns, we were not able to measure any differences between the different turn-taking cues given. While participants gave in general positive feedback, they also complained about the hardness of listening to and remembering the family stories which had a very high information density. This difficulty potentially reduced their attention to the turn-taking cues given. Especially *Coherence* (invertedly evaluated with questions like “The persons’ behavior does not make sense”) was rated very high, however, similarly in all conditions (means per turn-taking cue level ranging between 2.16 and 2.32). Therefore, we have to discard hypothesis **H1** as no differences in the evaluated sub-dimensions of social presence were found.

During the `Act` part participants had to specifically focus on the turn-taking cues to decide when to start speaking. When evaluating the gap length, we found evidence that adapting the gestures is the most effective cue in our system. We were not able to show that adding more modalities is beneficial for gap lengths, although there might be a tendency (cf. Fig. 5.17). We nevertheless partly accept hypothesis **H2**. Furthermore, *Clarity* seems to improve with added cues, albeit only significantly again for manipulating gestures. This again leads us to partly accept hypothesis **H3**. What is interesting to notice is that while gesture manipulation had the only significant effect, it was only noticed by half of the participants when having to state what they focused on for turn-taking. Gaze manipulations on the other hand were noticed by more than 80%. Interestingly the majority of participants also reported focusing on modalities we explicitly did not change, like prosody. Breathing, however, went fairly unnoticed and also did not show any effects.

Limitations

While the inhalation sound was played at the identical volume as the speech, this modality could still be improved especially for showing the willingness to take over the turn, for example, by a sharp inhalation during another speaker’s turn (we only played inhalation sounds during the gaps) and was therefore slightly increased for all following studies. Furthermore, the gaps during the `Listen` part were static and rather short (all lasting 300 ms) which might have had a negative influence, since the additional modalities might especially play a role in prolonged gaps, e.g., due to thinking. A further aspect we noticed is that the environment gaze targets were not optimally placed often leading to “averted” gazes which are only slightly off from looking at the participant, which some commented on negatively. This was adapted for following studies moving the gaze targets farther away from direct line of sight to the user. Furthermore, since most of our participants came from the same cultural background (German), the presented results might only be applicable to this cultural group. Another observed behavior we did not consider is that of posture shifts, which, following [Cassell et al., 2001a], appear more frequent at turn shifts.

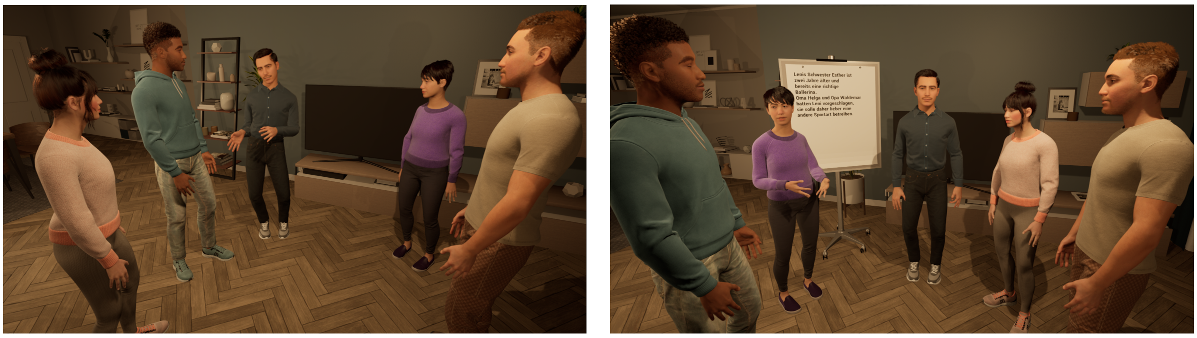

Lenis Schwester Esther ist
zwei Jahre älter und
bereits eine richtige
Ballerina.
Oma Helga und Opa Waldemar
hatten Leni vorgeschlagen,
sie solle daher lieber eine
andere Sportart betreiben.

Figure 5.18.: Listening Agents Study: The study scene was enhanced by three additional ECAs functioning as listening agents. During the listening phase of the study these listening agents followed the conversation between the two ECAs telling the HTR stories (left) and also between the participant and one ECA in case the participant took over in telling the story in the second part of the study (right).

5.5.5. Preliminary Investigations: Integration of Listening Agents

The contents of this section are based on and taken in part from work previously published in [Ehret et al., 2025a].

While the direct cues given by the ECAs were further improved, e.g., repositioning gaze targets and increasing the loudness of breathing, we also looked into another possible direction to increase the clearness of turn-taking cues. To this end, we implemented and evaluated the use of additional listening agents, as listeners in real life conversations are able to anticipate turn changes and indicate this by gaze shifts [Holler and Kendrick, 2015]. They only engage in listening to the conversation and potentially enhance the turn-taking cues by following with their attention those given by the speaking ECAs. The listening agents' main interaction is thereby the gazing behavior [Oertel et al., 2021] and they also occasionally give back-channel signals like nodding, small vocalizations, and smiling on mutual gaze with the speaking ECA [Skantze, 2021] for an enhanced rapport and understanding. The presented listening agents were implemented and evaluated during the master thesis of Valentin Dasbach and therefore only briefly discussed here for the sake of completeness. The evaluation was done using a very similar study design as used in Sec. 5.5.2. The distribution of gazes for the speaking ECAs, the other listeners and the environment were designed following the descriptions in [Vertegaal et al., 2001], [Rienks et al., 2010], and [Oertel et al., 2021]. Thereby at turn changes, listening agents gazed with a 60% chance 100 ms prior to the change already to the next speaker, to show the anticipation of them speaking next as additional turn-taking cue [Holler and Kendrick, 2015]. We chose minimal variances in this timing to enhance the believability of the interaction and ensure a more natural conversational flow, specifically $M = 0.1$ s and $SD = 0.03$, following a normal distribution.

The study was conducted with $n = 25$ participants (mean age: 26.2 years, SD: 3.5; 13 male, 12 female). To this end the within-subject study from Sec. 5.5.2 was enhanced by three

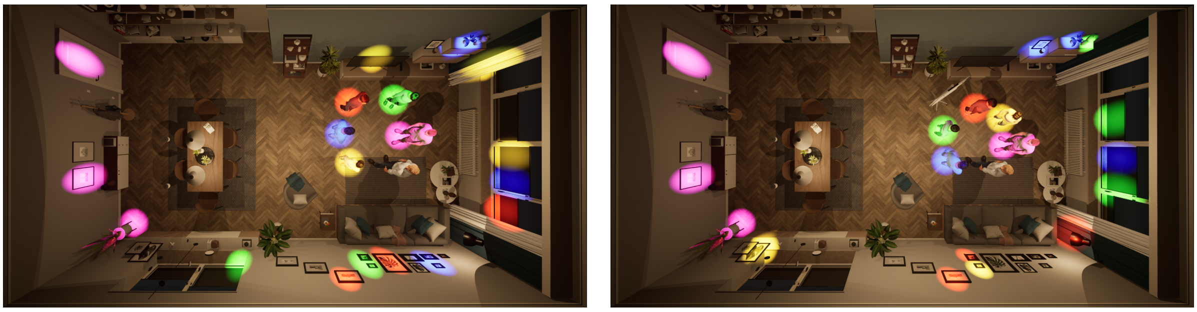


Figure 5.19.: Besides the user visualized as male avatar, the social group includes (left) two speakers (green, red) and three bystanders during the Listen Phase, and (right) one speaker (green) and four bystanders in Act Phase. Agents are consistently colored across scenarios, with environment-based gaze aversion targets per agent marked accordingly.

listening agents (see Fig. 5.18), keeping the overall study procedure. See Fig. 5.19 for the circular formation of the involved interactants, a common setup for stationary groups during conversations [Bönsch et al., 2020b; Ennis and O'Sullivan, 2012], which not only facilitates engagement but also provides participants equal visual access to all group members. This figure also shows their respective environment gaze targets when averting gaze. As in the turn-taking study explained in Sec. 5.5.2 participants again had to listen in the first phase (*Listen Phase*) to a story been told by two ECAs and took over the part of one ECA in telling the story in the second phase (*Act Phase*) while having to react to all available cues to estimate when it was their time to speak. The study had one main factor *Bystanders* (B), which had three levels: In B_{None} no additional listening agents were added, while the three added listening agents in B_{Random} showed random behavior and acted according to the social model explained above in B_{Social}. In each phase each level was repeated four times per participant. As originally also another factor was assessed, namely whether participants are treated in the gaze model as addressees or only as bystanders overhearing the conversation. Analysis, however, showed that there were no differences and this factor is therefore ignored here. However, this led to two blocks for each condition with two repetitions each, of which the order was counter-balanced between participants using Balanced Latin Squares (see Sec. 3.2.1).

Results

In the Listen Phase, we evaluated whether bystanders had an influence on social presence (by means of the *human-like behavior* and the *social presence* construct from the ASA questionnaire, see App. B.11). We were not able to show a difference in social presence between the three levels of Bystanders. Additionally, we looked at the gaze behavior of participants. More specifically, we firstly evaluated at which gaze targets (the speaking agents, the other bystanders, and the environment) participants looked and further examined the gaze switching behavior between sentences of the ECAs towards the next speaker. Fig. 5.20 (left) shows the gaze proportions and Fig. 5.20 (middle) shows the mean time it took participants to focus on the next speaker when the turn was yielded, so the speaker changed. Therefore we evaluated

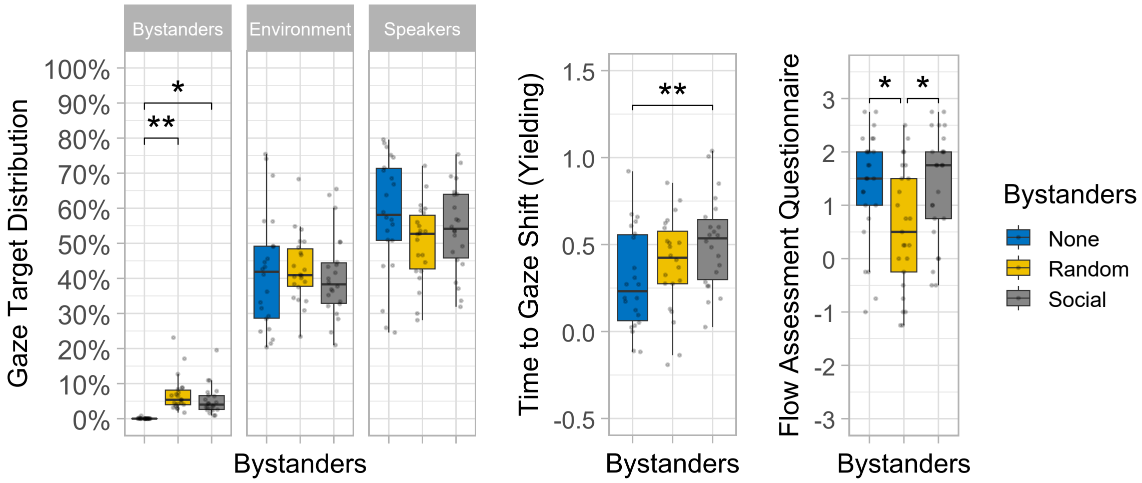


Figure 5.20.: Results of the Listening Agents Study: The gaze target distribution during the story presentation (left). The time in seconds it takes participants to look at the next speaking agent in case of a turn yield (middle) and the mean answers to the questionnaire asked after each block in the Act Phase. Significant pairwise differences are shown as $**$ for $p < .01$ and $*$ for $p < .05$, all other differences are non-significant.

the time it took to look at the next speaker after the last speaker ended its turn in a time frame of 0.5 s before the sentence ended and up to 3 s after the turn change. If during this time frame the next speaker was never looked at, this turn change was ignored for the evaluation. Three participants were excluded from the gaze analysis since there were technical problems with the gaze tracker.

Due to violations of normality in the gaze target distribution data, we employed the Aligned Rank Transform (ART) [Elkin et al., 2021] method for non-parametric factorial analysis which allows analysis analogous to a two-way repeated-measures ANOVA while respecting the non-parametric nature of the data. We assessed the effects of the three listening bystander conditions on the participants' gaze allocation among the three gaze target categories. The analysis indicated significant main effects for targets ($F(2, 168) = 253.46, p < .001$), as well as interaction effects between bystanders and targets ($F(4, 168) = 5.66, p < .001$), showing that the presence of bystanders significantly influences where participants direct their attention during conversations (Fig. 5.20 (left)). Tukey-corrected paired ART-C tests revealed several significant differences in the participants' gaze allocation when comparing the bystander conditions among the gaze targets (Fig. 5.20 (left)): Unsurprisingly, gaze allocation towards the *Listening Bystanders* was significantly higher when bystanders were present (B_{Random} and B_{Social}) compared to B_{None} (both $p's < .021$). However, there was no significant difference in gaze allocation between B_{Random} and B_{Social} ($p > .99$). No significant differences were observed in gaze allocation towards the *Environment* across the bystander conditions (all $p's > .99$) as well as towards the *Speakers* (all $p's > .60$).

A repeated-measures ANOVA was performed for the gaze switch times, as normality was confirmed for the data by means of a Shapiro-Wilk test. It revealed a statistically significant difference in gaze switch times in the Listen Phase across the three different bystander conditions ($F(2, 44) = 7.95, p < .001$). Pairwise t-tests indicated that this effect was due to a significant difference in switch times between B_{None} ($M = 0.29$ s, $SD = 0.28$) and B_{Social} ($M = 0.51$ s, $SD = 0.26$) ($p = .003$), while no significant difference ($p's > .08$) was found in the other two pairs with B_{Random} ($M = 0.44$ s, $SD = 0.27$). (Fig. 5.20 (middle))

During the Act Phase, we again evaluated gap times before participants started speaking, which did not show significant differences. We further posed a custom questionnaire after two repetitions of the same condition evaluated how well participants assess the flow of the cooperative story telling. This questionnaire contained 4 questions, namely "*It was easy to comprehend when I should speak.*", "*The behavior of the other persons was ambiguous.*", "*The behavior of the other persons confused me.*", and "*The other persons followed the conversation.*", which were rated on a 7-point Likert scale between -3 ("*Do not agree*") and 3 ("*Agree*"). Answers to the second and third question were reversed, so that positive values represent a good comprehension. Cronbach's α was acceptable with a value of .68 so we evaluated the mean of those questions. A repeated-measures ART ANOVA showed a significant difference between the bystander conditions ($F(2, 48) = 5.47, p = .007$). Tukey-corrected paired ART-C tests revealed significant differences in mean ratings between B_{Random} and B_{None} ($t(48) = 2.86), p = .017$) as well as between B_{Random} and B_{Social} ($t(48) = -2.87), p = .016$), with B_{Random} receiving the lowest scores, as shown in Fig. 5.20 (right).

Furthermore, participants provided qualitative insights through written feedback at the end of the study. Verbal comments throughout the study were noted by the experimenter together with the related condition. All qualitative feedback was grouped manually, and key insights are reported anecdotally. Feedback indicated that while many found it generally easy to follow conversations and anticipate who would speak next, some expressed difficulty when listening bystanders were present. Three participants, for example, explicitly stated they recognized the next speaker based on listening agents' behavior, while six participants reported that they identified turn changes primarily through cues from the current speakers. Overall responses suggested that while engagement levels varied based on agent configurations (particularly with random gazing), many participants felt more comfortable following conversations without bystanders as potential visual distractors.

Discussion

The findings of this study provide valuable insight into how virtual listening bystanders influence participants' gaze behavior and perception of turn-taking during interactions with ECAs.

The analysis revealed that the presence of social bystanders negatively impacted participants' ability to quickly switch their gaze to new speakers during conversations. Participants exhibited longer gaze switch times in conditions with social bystanders compared to no bystanders, suggesting that additional visual stimuli may have distracted users from identifying turn changes effectively. This aligns with previous research indicating that clarity in visual cues is essential for recognizing conversational dynamics ([Wang et al., 2013; Mutlu et al., 2009; Oertel et al., 2021]). These results challenge our hypothesis that social bystanders would enhance participants' detection of turn changes compared to only speaking ECAs. However, they show that unsocial behavior will worsen the conversation flow assessment. So, while social bystanders may improve the naturalness of interactions, they may also be distractors.

The results also indicated a shift in attention away from speakers when bystanders were present, highlighting how additional agents can dilute the focus on primary conversational partners. Interestingly, while random bystanders tended to look more frequently at the environment, this behavior did not effectively redirect users' focus toward those areas. Instead, both random gazing and social bystanders primarily changed gaze distribution within the social group. While it is not surprising that the presence of other agents can draw attention—especially when they look at the user—this finding remains relevant as it underscores potential challenges in maintaining engagement within multi-agent interactions. The decreased gaze toward speakers in both random and social bystander conditions points to potential challenges in maintaining engagement within multi-agent interactions.

These findings suggest that while incorporating listening agents may aim to enhance social presence, it can inadvertently lead to confusion regarding or ignorance of turn-taking cues, even in such a simple setting with a limited amount of speaking agents. It is crucial to critically consider that if no listeners are present, users cannot engage visually with them, thus, their absence does not detract from interaction quality. However, when listeners are included, their behaviors must be meaningful. Otherwise, it may be better to exclude them entirely. As such, careful consideration must be given to agent behaviors and their implications for user experience. Furthermore, feedback indicated that the unsocial bystanders particularly obscured the conversation flow due to their random gazing behavior, which may have detracted from the overall interaction quality.

There are a few **shortcomings** in our study. First, the sample size is relatively small, which may affect the generalizability of our findings. Second, we employed a challenging standardized psychology task designed specifically to focus participants on the ongoing conversation rather than the ECAs' behavior. While this approach aimed to elucidate subconscious social dynamics, it might limit broader applicability. Finally, exploring alternative social models could provide further insights into optimizing ECA configurations for enhancing communication. Still, this study highlights an important area for future research: understanding how various configurations of ECAs can be optimized to support clearer communication and enhance user interaction without overwhelming them with extraneous visual information or introducing distracting behaviors.

5.6. General Discussion

In this chapter we discussed the different co-verbal modalities that need to be implemented to generate believable ECAs. These comprise facial animations, including gazing, and gestures. We also looked into communicative functions potentially conveyed by these behaviors, for example, back-channels and turn-taking signals. We deliberately excluded the topic of proxemics [Bönsch et al., 2016], so the distance ECAs keep from each other and human interlocutors. While this is a very relevant topic for ECAs that move through a virtual environment together with a user, it is not as relevant for the main topics discussed here, where the ECAs stayed in one place and only the user moves around. We refer the interested reader to our other publications regarding this topic, for example, [Bönsch et al., 2018a].

One key observation during the creation of this work was, that ECAs always have to be regarded holistically. If one of these modalities is falsely or not implemented at all, this can have a significant negative effect that has the potential to mask improvements that should be evaluated for other modalities. One example of this will be shown in the following chapter – were we omitted facial animations altogether and evaluated the effect on social presence. Therefore, it is important to cater for all of these modalities at least to some degree. With the presented implementation, which are all part of the `CharacterLib` plugin (see App. C.3) for Unreal Engine, we tried to address this challenge by giving a toolbox which can be used to easily generate all of the aforementioned behavior. However, providing out-of-the-box solutions for all co-verbal behaviors is a complex endeavor since, as discussed before, context matters a lot for the actual behavior to be produced. While it seems easy to implement a natural blinking pattern that could be applied to ECAs in all contexts, it becomes already far more complex when turn-taking signals should be produced. And even for the simple case of blinking, in reality it is very dependent on the activity and the emotional state of the speaker [Willett et al., 2023]. This can be explained by the fact that the co-verbal behavior is a result of a plethora of cognitive processes happening within a human performing this behavior, which are obviously not reproduced entirely in an ECA. Therefore, there will always be a necessity to carefully tune and evaluated co-verbal behavior generated by ECAs based on the actual context they are used in as, for example, Han et al. [2024] found people to change their co-verbal behavior strongly between different spaces, e.g., being in public spaces or rather alone.

Another challenge along these lines is that the same behavior can be judged differently based on its co-occurrence with other behavior. For example, in the observations of Etienne et al. [2023], smiles of audience members were perceived as non-genuine “fake smiles”, because the same virtual audience member performed a dismissive gesture some times before. The same visual representation of a smile, would, however, be rated more positively in another context or if the behavior of the ECA would have been more consistently positive. One possible explanation for this is that humans form a theory of mind about others, including anthropomorphic ECAs. This leads us to imply intents for the sum of co-verbal behavior perceived and ascribe feelings or motives to ECAs which were potentially not intended during their design and just happened due to an unfortunate combination of produced signals.

One tendency that could be seen for all described modalities is the trend towards using more data-driven generative approaches relying on machine-learning. These could be found for all presented behaviors and have the possibility to produce more natural behavior, especially avoiding repetitions of the same gestures or sequences over and over again. This comes, however, at the cost of lower predictability and reproducibility. Another potential shortcoming is that these models (if not explicitly incorporating different speaker ids [Yoon et al., 2020]) try to recreate the mean behavior of the exemplary human data they were trained on. However, for co-verbal behavior their is a large inter-personal difference of the way gestures etc. are used. So these models potentially produce something less expressive than the performances of the individuals they were learning from. This, however, also poses a major challenge for rule-based approaches, which require someone to judge how natural/realistic an implemented behavior is, which can be extremely demanding due to inter-personal and potentially also inter-cultural [Wagner et al., 2014] differences which lead to the non-existence of a single correct/realistic realization. In general, the results that are created by those data-driven approaches have massively improved over the recent years, so their usage becomes even more attractive. Furthermore, these approach provide the capability to produce behavior for all modalities together, reducing the aforementioned risks of detrimental co-occurrence. In this work merely rule-based/recorded behavior was used, due to its reproducibility and the lack of available easy-to-use models at the time of implementing. However, we recommend to re-evaluate the usage at a given time. Especially the generation of co-verbal gestures can be a very interesting venue, since actors being asked to reenact a speech also tend to use repetitive gestures, and the use of recorded gestures strongly limits flexibility both in adapting the scenario and also during implementing the actual animation, which requires blending between different gestures and potentially adapting gestures to other idle poses etc.

In summary, we presented related work to the most important co-verbal cues throughout this chapter and introduced our *Character Plugin*, which provides the basics needed to implement convincing co-verbal behavior for ECAs. While basal behavior can be used as is, we also developed the plugin with the need of developers in mind to extend it, based on the specific needs of their research objectives, and therefore aimed for easy extendability during software design.

Chapter 6

Evaluating Embodied Conversational Agents

In the beginning of this thesis (Sec. 2) we discussed the concept of social presence and existing metrics to measure it, both subjectively with questionnaires as well as objectively using other measures like proximity or gaze behavior. Throughout the course of this thesis, we, however, discovered that measuring social presence through questionnaires is particularly challenging, as participants often struggle to rate interactions retrospectively, which can lead to indifferent results (see, e.g., Sec. 4.2.1). In this chapter we aim to elaborate further on the approach of objectively measuring social presence by means of using other metrics as proxy and comparing those to questionnaire-based measures. Although the insights from questionnaires are limited themselves (see, e.g., [Slater, 2004]), this comparison can still provide valuable information about the proxies' relative effectiveness and help identify strengths and weaknesses in measuring social presence overall. To address this question, we will, among others, use the deliberately degraded gestures from Sec. 5.3.4. Improving gesture quality is essential, as the realism of human motion significantly influences user experience in VR, demonstrated, e.g., by Patotskaya et al. [2023]. The presence of degraded co-verbal motions therefore offers an opportunity to deliberately alter the fidelity of ECA performance, impacting the perceived social presence (see [Xenakis et al., 2023]). Other possibilities, looked at in this chapter, are alteration of the verbal content using synthetic voices (see Sec. 4.1) and further degraders of the co-verbal behavior, like omitting lip sync or co-verbal gestures altogether.

This approach goes beyond the contributions of the previous chapters, where we evaluated isolated aspects of ECA behavior, like the voice or the auralization. Here, we will look at ECA performances in their entirety and will therefore use the results and behaviors described in the previous chapters, like believable gazing (Sec. 5.2), recorded speech (Sec. 4.1), auralization using static directivity filters (Sec. 4.2), turn-taking (Sec. 5.5), and minimal back-channeling (Sec. 5.4) signals.

In the first section (Sec. 6.1) we will evaluate cognitive spare capacity as measured by the HTR paradigm [Schlittmeier et al., 2023] as a potential proxy for perceived social presence. After primarily looking at the ECA's performance in the first study, we will consecutively present a study, where we evaluated the influence of audio-visual coherence of the surroundings in Sec. 6.2 on (social) presence and also again on cognitive performance. Finally, we will conclude this chapter with a brief discussion in Sec. 6.3.

6.1. Objective Social Presence Evaluation for Deliberately Degraded ECA Performance

The primary objective of this section is to evaluate whether the HTR paradigm can serve as an objective proxy for perceived social presence. The HTR task primarily measures memory performance, but when combined with a secondary task in a dual-task paradigm (see [Mohanathasan et al., 2022, 2024]), it can also assess cognitive spare capacity. This dual focus allows us to gain deeper insights into how various ECAs influence users' cognitive load and overall engagement during interactions, which are critical factors in determining perceived social presence.

To effectively observe changes in the HTR paradigm, it is essential to use degraders that are likely to produce significant effects. To this end, we deliberately degraded the speech performance of ECAs presenting the HTR family stories (see App. A.1) using various behavioral degraders, classified into gesture-related and speech-related degraders:

One critical area we identified is the degradation of co-verbal gestures, as our research in Sec. 5.3.4 demonstrated their strong influence on perceived naturalness, a key contributor to social presence. Consequently, we selected the worst performing condition of that study, namely the $M_{\text{OverPeakRemove}}$ modifier, referred to here as condition D_{BadGest}, which makes co-verbal gestures while speaking look more robotic, resulting in discontinuous movement patters. To complement these bad gestures, we added a condition termed D_{NoGest}, only playing idle animations during speech, so omitting co-verbal gestures entirely, as we expect missing co-verbal gestures to have a detrimental effect on naturalness as well. In addition, since we have two conversing ECAs that require effective turn-taking cues for a smooth conversational flow, we utilize our turn-taking model described in Sec. 5.5. This model incorporates gaze behavior, breathing patterns, and holding gestures to facilitate turn-taking. As part of our degradation strategy, we intentionally introduce incorrect turn-taking signals, referred to as D_{WrongTT}. These signals provide random indications of whether the turn is being held or yielded after each sentence, regardless of the actual narrative progression. Importantly however, these signals remain consistent across all three modalities.

To validate the general approach, we also added two more speech-related degraders potentially showing even stronger effects: As a more realistic presentation of ECAs is associated with higher social presence (e.g., [Zibrek and McDonnell, 2019; Kimmel et al., 2023; Xenakis et al., 2023]), we omitted lip sync animation as one additional condition in $D_{\text{NoLipSync}}$. Following the experiments of Kimmel et al. [2023], we expect this to decrease perceived social presence. Furthermore, we found that synthetic voices, generated by Text-to-Speech (TTS) engines, can have a detrimental effect (see Sec. 4.1). Therefore, we added as last degrader also synthetic voices instead of the recorded ones used otherwise (D_{TTS}).

Finally, as baseline, we have a condition where no degrader is applied (D_{None}). All in all, this results in six levels for the factor Degrader: D_{BadGest}, D_{NoGest}, D_{WrongTT}, $D_{\text{NoLipSync}}$, D_{TTS}, and D_{None}.

Hypotheses

We conducted a within-subject study with the degrader as the only factor addressing the following hypotheses:

H1 Social presence will be lower for degraders than D_{None}, especially $D_{\text{NoLipSync}}$ and D_{TTS}.

H2 Cognitive spare capacity, measured through the HTR paradigm combined with a secondary task, correlates with perceived social presence.

H1 is motivated by the results from our pre-study (Sec. 5.3.4) as well as the results by Kimmel et al. [2023] and Zibrek and McDonnell [2019] hinting at higher social presence for more natural presentation. The idea for **H2** is that unexpected ECA behavior binds cognitive resources, which are then no longer available for performance in the dual-task and therefore its measured cognitive spare capacity decreases. Additionally, we exploratively evaluate how HTR dual-task performance as an objective proxy potentially correlates with other potential indicators like gaze behavior [Cañigueral et al., 2021] and the physiological cognitive load indicator measured by the *HP Reverb G2 Omnicept* [Siegel et al., 2021] based on heart rate, pupilometry, and gaze tracking.

Study Material

We again used the same stories from the HTR paradigm, including content questions being verbally answered (see App. A.1). The stories are again presented by two ECAs to incorporate turn-taking, for which the same full implementation as in Sec. 5.5 is used. In all conditions, the speaking ECA prolongs and holds gestures during the gap (500 ms[1]) between two sentences when signaling to continue speaking and breathes in audibly before starting a sentence. Additionally, ECAs should engage in mutual gaze when yielding and break any mutual gaze when holding the turn (see [Ehret et al., 2023]). Unfortunately, an implementation error was only discovered after the study was completed, resulting in the gazing at the end of the turn always

[1] The gap between sentences is increased to 500 ms which is slightly longer than in Sec. 5.5 for increased naturalness, as the 350 ms used there in hindsight appeared too rushed.

being executed invertedly, so mutual gaze was held when the turn was held and not when it was yielded. However, since this was only one of three components of turn-taking signals and gestures were estimated by us to be the the most important (see Sec. 5.5), we assume that the effect was limited, but should nevertheless be considered when interpreting the results.

As proposed in [Mohanathasan et al., 2022], we used a vibrotactile task as the secondary task in the dual-task paradigm [Gagné et al., 2017]. This vibrotactile secondary tasks yields a better measure for cognitive load compared to a visual secondary task [Mohanathasan et al., 2022] and avoids participants focusing to consciously on the ECA behavior as they were occupied by the secondary task. Therefore, while listening to the stories being told, participants also had to respond to vibrational patterns being presented via the controllers (see App. A.1 for details). We utilized the gaze tracker of the *HP Reverb G2 Omnicept*, to track where participants looked during the presentation of the stories. Therefore, we logged the amount of time participants gazed at the eyes, mouth, or body of the currently speaking and also the currently not speaking ECA. *HP Reverb G2 Omnicept SDK* also provides a machine-learning-based interference model for cognitive load based on physical markers sensed by the headset [Siegel et al., 2021]. We stored the average value of this cognitive load measure during listening to the story for each condition. Since only one of the two ECAs (see Fig. 6.1) is speaking at a time while the other one is listening, we added listener reactions. These were implemented with a nod or a brief smile at the end of the other's sentence with a probability of 50% each (see Sec. 5.4).

During the D_{None} condition, this behavior was shown as described above. In the D_{BadGest} condition, however, all gestures were replaced by the ones using the $M_{\text{OverPeakRemove}}$ modifier as described in Sec. 5.3.3. When presenting during the D_{NoGest} block, the ECAs did not perform any co-verbal gestures, but just stayed in their idle movement all the time, exhibiting a slight body sway only. For the D_{WrongTT} degrader, the ECAs did not perform random behavior independently for each of the three modalities (gesture, gaze, and breathing) as outlined in Sec. 5.5. Instead, they demonstrated random turn-taking behavior that was consistent across all three modalities. However, it is important to note that the gaze behavior was not executed correctly due to an implementation error, as stated previously. This means that for each gap between two sentences it was randomly decided whether to perform turn-hold or turn-yield behavior independent of which ECA would speak the next sentence. For D_{TTS}, synthetic versions of the sentences were generated using *Google Text-to-Speech* with the voices *en-AU-Wavenet-D* and *en-US-Wavenet-C*. To approximately match these sentences with the gestures recorded for the recorded speech, each sentence was scaled to the length of the recorded sentences using *pyrubberband*. This yielded a mean speed-up of 34% (min = 5%, max = 73%). To generate lip movement for the synthetic speech, *Oculus Lip Sync* was used on these scaled files. And lastly, during the $D_{\text{NoLipSync}}$ block, simply the recorded lip sync was not shown. The different conditions and the task procedure can be seen in the supplemental video of [Ehret et al., 2025b].

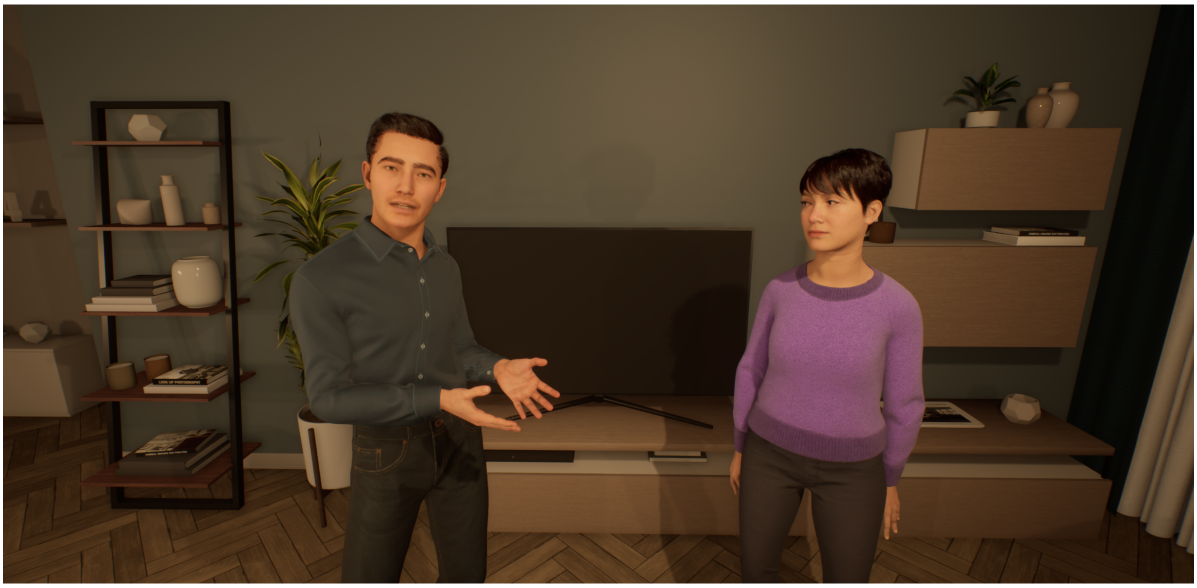


Figure 6.1.: The two ECAs telling one HTR story during the study without degradation (D_{None}). The male ECA is gesturing, while the female uses the idle movement with gaze behavior superimposed.

Study Procedure

The study was conducted following the Declaration of Helsinki and using the *StudyFramework* (see Sec. 3). After reading the study description and giving their informed consent, participants filled out a demographics questionnaire. They were required to be fluent in German. Afterwards, they were equipped with the tracking hardware (*HP Reverb G2 Omnicept* and two *Valve Index Controllers*) and *Sennheiser HD650* headphones connected to a *Focusrite Scarlett 2i2 3rd Gen* audio interface. Once immersed in the virtual living room scene (see App. A.4.3), first a calibration of the gender-matched body-avatar (see Sec. 3.4) was performed and the eye tracking was calibrated. After that, participants had time to practice the secondary (vibrotactile) task (see App. A.1 with the pattern short-short, short-long, long-short, and long-long) until they felt comfortable doing it, and afterwards performed a 15-trial single task baseline measurement of it in the empty living room. This was followed by a single-task practice and baseline block of the HTR task, with both ECAs present, without a degrader being applied to their performance. After the presentation participants had to verbally answer nine questions related to the content of the story. They were explicitly allowed to answer “I don’t know”, rather than guess, in case they had no idea about the correct answer. This was counted as incorrect, nevertheless. The correct answer was displayed for the experimenter, who had to log whether the right answer was given by the participants by means of keyboard buttons. The first block of the study was then finished with a practice block, performing both tasks at the same time. Participants were instructed to listen carefully to the stories, prioritizing the primary (listening) task, while paying attention and looking at the ECAs.

After a break, six blocks with three repetitions of listening to a story and answering the nine questions for each text followed, one block per Degrader. The order of the blocks was balanced between participants using Balanced Latin Squares (see Sec. 3.2.1). Each of these blocks ended with two Likert scale questionnaires regarding the just-finished block being presented within the virtual scene, which participants had to answer using the controller. The first questionnaire was the *Human-Like Behavior (HLB)* construct from the ASA questionnaire [Fitrianie et al., 2022] in its German translation [Albers et al., 2024], which had to be answered on a 7-point Likert scale (see App. B.11). We only picked two constructs from the full questionnaire, to not overwhelm participants with to many questions and focusing on the dimensions most interesting to our primary research focus here. Secondly, we also posed the *Social Presence (SP)* items of the Multimodal Presence Scale (MPS) [Makransky et al., 2017], which are answered on a 5-point Likert scale, again using a German translation [Volkmann et al., 2018] (see App. B.8). After this, a short break in which refreshments for the participants were provided followed, and afterwards the next block started.

After finishing all six study blocks, participants had to fill out a final desktop questionnaire using *SoSciSurvey*, were debriefed, and compensated with 15€. The study lasted 60 to 90 minutes, with the primary factor for variation being the time each participant took to reflect on the content questions following the stories and the time participants took for breaks between the blocks. Of this time participants spend on average 37.7 minutes (SD = 7.3) immersed in VR, listening to the stories or answering questions about them during the Study Phase.

6.1.1. Results

In this study, $n = 30$ persons (19 male, 10 female, 1 preferred not to say) participated, with a mean age of 25.4 years (SD = 4.8, 20-46 years). All of them reported (corrected-to) normal vision and normal hearing. While 19 were native German speakers, the remaining 11 self-reported to be at least fluent in German as the study was conducted in German. Ten participants reported to interact with virtual humans (e.g., in computer games) multiple times a week, four stated to do so at least several times a month, while 12 answered “less often” and four responded with “never”. A simulation-based a priori power analysis (using *simr*) [Green and Macleod, 2016] showed for the vibrotactile task a power of $> 96\%$ (95% CI [96.38%, 100%]) assuming a sample size of $n_2 = 30$ and a performance spread of 60% to 80% correct responses across the six degrader conditions. Regarding the questionnaire results, it predicted a power of 86.90% (95% CI [84.65%, 88.93%]) assuming a medium effect (Cohen’s $f = 0.25$). This suggests that the chosen sample size was adequate for detecting medium-sized effect in this study.

We conducted statistical tests using Generalized Linear Mixed-Effect Models (GLMMs) in *R* (v4.4.1) [R-Core-Team, 2015] with the *lme4* package [Bates et al., 2015]. Post-hoc pairwise comparisons were Bonferroni-corrected and based on estimated marginal means, computed using the *emmeans* package [Lenth, 2024]. Following the analysis in [Mohanathasan et al., 2024], for binary data (i.e., vibrotactile and HTR question performance) binomial distributions with logit link function were used and for strictly positive continuous data (i.e., reaction times),

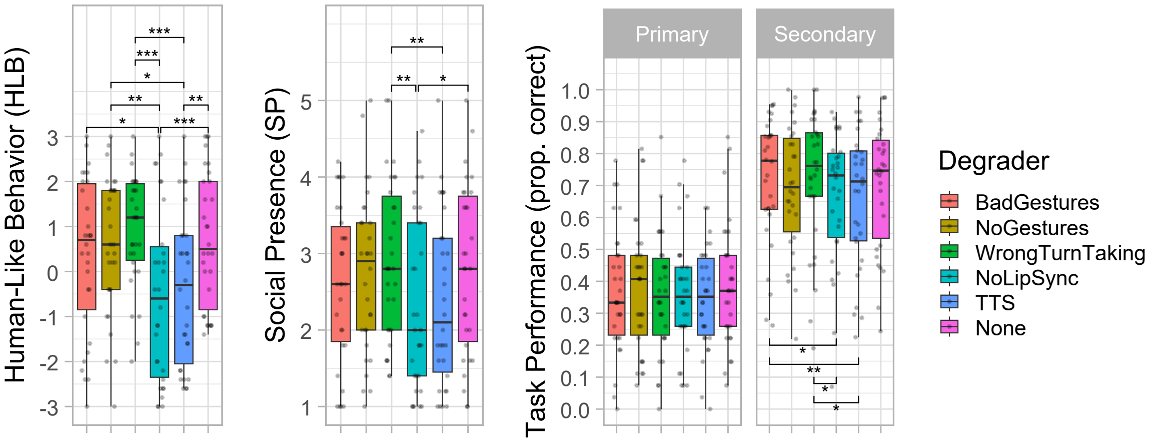


Figure 6.2.: Subjective rating of human-like behavior (HLB) construct of the ASA questionnaire [Fitrianie et al., 2022] (left) and the social presence (SP) dimension of the MPS questionnaire [Makransky et al., 2017] (middle). Objective rating in form of proportion of correct answers in the primary (memory) and secondary (vibrotactile) task (right). All shown as box plots, displaying median, interquartile range, and potential outliers, with the rating or mean performance per participant shown as semi-transparent points. Significant pairwise differences are shown as $***$ for $p < .001$, $**$ for $p < .01$, and $*$ for $p < .05$.

a Gamma distribution with log link function was utilized. In case of potentially negative or zero data, again, LMMs instead of GLMMs were employed. Potential random effects considered during backward model selection were self-reported experience with virtual humans and self-reported focus on the virtual humans during conversation, age, text, and participant ID, repetition during block, and, where applicable, vibrational pattern or question ID. The only fixed effect was the Degrader. To decide which factors to include, we again performed backward model selection (see Sec. 5.3.4).

Looking into the behavioral results of the **primary (listening) and secondary (vibrotactile) task**, the Degrader had no influence on the performance in the primary listening task ($\chi^2(5) = 5.21, p = .39$) as can be seen in Fig. 6.2 (right). This figure also shows an overall low recall rate of on average only 36.8% (SD = 48.2), spanning almost the full 0–100 % range——demonstrating vast inter-participant variability. However, performance in the secondary (vibrotactile) task was significantly influenced by the degraders ($\chi^2(5) = 18.38, p = .003$). This was assessed using a model that included the Degrader as a fixed effect, along with the pattern used (short-short, short-long, long-short, long-long) and the text and participant identifier as random effects. Bonferroni-corrected pairwise comparisons of the number of correct responses to vibrotactile patterns showed a significant difference between D_{TTS} and D_{WrongTT} (z-ratio $= -3.37, p = .011$), D_{BadGest} and $D_{\text{NoLipSync}}$ (z-ratio $= 3.30, p = .015$), D_{BadGest} and D_{TTs} (z-ratio $= 3.52, p = .007$), $D_{\text{NoLipSync}}$ and D_{WrongTT} (z-ratio $= -3.16, p = .024$), see Fig. 6.2 (right). There were also non-significant trends for D_{TTS} vs. D_{BadGest} (z-ratio $= -2.80, p = .076$) and $D_{\text{NoLipSync}}$ vs. D_{WrongTT} (z-ratio $= -2.80, p = .077$). Following the procedure in [Mohanathasan et al., 2024] when analyzing response times for correctly answered vibrotactile trials, trials below 200 ms were removed (see [Whelan, 2008]) as well as times outside a 2-SD margin around the mean (see [Berger and Kiefer, 2021]). The remaining response times did not show a significant dependence on the degraders ($\chi^2(5) = 5.78, p = .33$).

Also, no difference in the **physiological cognitive load** indicator directly measured on the HP Reverb was evident between the blocks ($\chi^2(5) = 4.49, p = .48$). When looking at the fraction of time participants **gazed** at different regions of the ECAs during presentation of the story, we did not find a significant influence of Degrader ($\chi^2(5) = 3.38, p = .64$). However, suprisingly, participants looked throughout all blocks only 25.3% (SD = 37.1%) of the time at the speaking ECA and 3.8% (SD = 6.8%) at the listing one.

After each block with a given Degrader, participants completed two questionnaires. They rated perceived **social presence (SP)** and **human-like behavior (HLB)**. For SP, which was rated on a scale from very low (1) to very high (5), the LMM model fitting to the data best just contained Degrader as fixed effect, which showed a significant effect ($\chi^2(5) = 23.6, p < .001$). This was due to a significant difference from D_{WrongTT} to $D_{\text{NoLipSync}}$ (t-ratio $= 3.87, p = .002$) and to D_{TTS} (t-ratio $= -3.56, p = .008$). Furthermore, D_{None} was also significantly different from $D_{\text{NoLipSync}}$ (t-ratio $= 3.87, p = .05$). All other pairs were non-significant (see Fig. 6.2 (middle) and Tab. 6.1). HLB, rated from very low (-3) to very high (3), analogously also showed an effect of Degrader ($\chi^2(5) = 43.7, p < .001$). The significantly different pairs can be found in Fig. 6.2 (left) and Tab. 6.1.

HLB **SP**	Bad Gest	No Gest	Wrong TT	No LipSync	TTS	None
Bad Gest		1.0	.74	**.03**	.21	1.0
No Gest	1.0		1.0	**.001**	**.01**	1.0
Wrong TT	.71	1.0		< **.001**	< **.001**	1.0
No LipSync	.95	.13	**.002**		1.0	< **.001**
TTS	1.0	.29	**.008**	1.0		**.004**
None	1.0	1.0	1.0	**.05**	.13	

Table 6.1.: p-values for Bonferroni-corrected pairwise comparisons of the questionnaires asked after each block. The lower left half shows values of the social presence (SP) dimension of the MPS questionnaire [Makransky et al., 2017] and the top right half shows ratings with regard to the human-like behavior (HLB) of the ASA questionnaire [Fitrianie et al., 2022]. Significant p-values are set in bold.

Participants were asked in the **post-study questionnaire** to rate several questions about their experience throughout all conditions on a 7-point Likert scale from "Strongly disagree" (-3) to "Strongly agree" (3). They responded that answering questions to the family stories was not easy ($M = -2.37$, $SD = 0.96$). When asked whether the ECAs looked, sounded, and gestured like real humans, the responses were slightly above average (looked: $M = 0.43$, $SD = 1.43$;

sounded $M = 0.97$, $SD = 1.69$; gestured: $M = 0.56$, $SD = 1.59$). The gaze behavior was rated as less realistic ($M = -0.20$, $SD = 1.73$), however, again with a large variance in the responses. When asked whether they carefully focused on the virtual person, once more a large spread is obvious ($M = 0.97$, $SD = 1.65$, $min = -2$, $max = 3$), albeit in general in a positive range.

Furthermore, participants were asked to indicate the block that they (least) preferred. Since they were not told which degraders were employed during the study, they should refer to this block by stating its sequence number or remarkable differences from the other block. The results can be found in Tab. 6.2

	Bad Gest	No Gest	Wrong TT	No LS	TTS	None	unclear	misinterp
most preferred	0	0	3	1	1	1	13	11
least preferred	4	2	1	6	8	1	2	9

Table 6.2.: Answers to the question which block was most/least preferred. “unclear” means that participant responses could not be clearly matched with an actual degrader while “misinterpreted” means that participants answered something very different like “names were to complicated”. For the least preference three participants gave two answers.

When asked what could be improved about the presentation, answers were very diverse. Three participants answered that the ECAs should make more eye contact and another three found the gesture to not be natural (while it is unclear to which block they were referring to). Furthermore, two participants found the content of the stories could be made more natural, as people would normally not tell such information-dense stories in a casual setting and two participants would have liked reactions of the ECAs to their answers given. Also, three participants mentioned that the repetitive gestures were not realistic. While it is true that idle gestures were pooled from four gestures per ECA and therefore repeated, this was not true for the co-verbal gestures, which had been individually recorded for each sentence [Ehret et al., 2023]. In the general open-ended feedback questions, four participants positively mentioned the eye contact made by the ECAs while three positively mentioned the gestures. Also two participants mentioned that the visual representations could have been better and ten participants mentioned that they found it particularly hard to follow the stories while reacting to the vibrotactile patterns in parallel.

6.1.2. Discussion

To evaluate whether cognitive spare capacity assessed via the dual-task implementation of the HTR paradigm provides an indirect indicator of social presence, we first have to look at the influence of the examined degraders on social presence by means of established questionnaires. This study indicates that, under the conditions tested, synthetic voices were rated worse with regard to social presence and human-like behavior, and also the least preferred. This is in line with the results of [Choi et al., 2023], finding decreased human-likeness ratings for TTS. Zibrek et al. [2021] did not find a significant detrimental effect of synthetic voices on social presence, however, in our implementation also the lip syncing was of lower quality during the TTS condition, using *Oculus Lip Sync* instead of recorded face movement. Furthermore, due to the temporal matching to the recorded sentences, some sentences potentially appeared rushed. While Luo et al. [2023] did not find a significant influence of omitting articulation movement on social presence, we found that a clear detrimental effect was evident. This is more in line with the results of Kimmel et al. [2023], who found that mouth movement significantly improved social presence during avatar-mediated communication, solving an explanation task. Our results also go beyond those of Easley et al. [2024], who did not find an increase in cognitive load with omitted lip sync as they originally hypothesized. The other degraders, however, had no clear influence on social presence, so **H1**, that social presence is rated lower for all degraders, can only be partially accepted.

With these results, we are finally in the position to look at our primary research objective, using cognitive spare capacity measured by the dual-task HTR paradigm [Mohanathasan et al., 2024] as an objective proxy for perceived social presence. Cognitive spare capacity decreased during the TTS conditions and also for missing lip sync. For the latter, an alternative explanation to subjectively decreased social presence and resulting decreased cognitive spare capacity, could also be that decreased speech intelligibility due to missing lip movement increased the listening effort, as observed in [Sewell et al., 2023]. Nevertheless, the sensitivity of our cognitive spare capacity measure to these changes suggests it is aligned with the questionnaire-based social presence metric used, leading us to cautiously support **H2**.

However, both methods—questionnaire-based and HTR dual-task—detected only major degradations. For instance, although our gesture manipulation study (see Sec. 5.3.4) showed degraded gestures (D_{BadGest}) to be perceived as less natural, these changes did not significantly affect social presence ratings or HTR performance. The same is true for omitting co-verbal gestures all together (D_{NoGest}) and for turn-taking cues being performed at random sentence endings (D_{WrongTT}). Looking into non-significant trends, it can be noted that the degraded gestures (D_{BadGest}) were not rated as strongly better compared to omitting lip sync and TTS when comparing them with the other three degraders (see Tab. 6.1). This trend is, however, not evident in the behavioral performance in the secondary task (see Fig. 6.2). Although these patterns were not statistically significant, they suggest that poorly fitting co-verbal gestures are not perceived as beneficial compared to omitting them altogether. This warrants further investigation before drawing firm conclusions about whether omitting co-verbal gestures is preferable to using degraded ones during ECA design.

It should be noted that during the HTR task only cognitive load differences were observed, indicated by performance variations in the secondary (vibrotactile) task, whereas memory performance in the primary (listening) task remained stable in contrast to, e.g., [Bailenson et al., 2005]. This is, however, consistent with the logic of using the HTR task in a dual-task paradigm, which assumes that participants prioritize the primary task and maintain performance levels, while cognitive spare capacity is allocated to the secondary task. The underlying expectation is that trying to memorize as much information as possible from the family stories, although demanding, does not fully occupy participants' cognitive resources. As a result, cognitive spare capacity remains available and can be allocated to the secondary task [Mohanathasan et al., 2025]. When additional cognitive load is introduced, for example through conversational degraders, this remaining capacity is further reduced, which can then be observed in the performance on the secondary task. This rationale underlies our use of the HTR task within a dual-task paradigm to examine cognitive spare capacity as an indirect proxy of social presence. Another potential explanation for not measuring degradation in memory performance is that small differences in the primary task went unnoticed due to the large inter-subject variability (see Fig. 6.2, right).

Furthermore, the employed cognitive spare capacity measure using the HTR paradigm was in our study more sensitive than the machine-learning-based physiological indicator, out-of-the-box available on the *HP Reverb G2 Omnicept*. Surprisingly this can also be said for using gaze behavior as a proxy for social presence. Although prior work [Doherty-Sneddon and Phelps, 2005] shows cognitive load can alter gaze behavior, we observed no gaze differences across our conditions. This contrasts with He et al. [2022], who found gaze varied between idle and generated co-verbal gestures. However, participants on average gazed only a quarter of the time at the speaker during story presentation and the remainder primarily away from the ECAs, which can again be a consequence of the high cognitive demand of the dual-task. This demand could have potentially prevented participants from carefully looking at the presented ECA behavior, which in turn potentially distorted the gathered results. This should be carefully considered in future research when employing gaze behavior as an objective metric. Nevertheless, the intention of this research was to evaluate perceived social presence as a holistic concept, without having participants focus too closely on specific modalities. Therefore, we deliberately decided on a demanding task to draw the participants' focus to the content presented by the ECAs, ensuring that their focus remained on processing the information rather than the characteristics of the ECAs.

There are some more **limitations** to consider with regard to the presented work. The recorded gestures reflect the interpretations of only two performers, limiting generalizability, which potentially has caused the gender difference in the gesture manipulation study (Sec. 5.3.4), which could also be not based on gender but simply on the different personality traits of the performers. Unfortunately, an implementation error in the turn-taking behavior was only discovered after the study was completed, resulting in the turn-taking gazing at the end of the turn always being executed invertedly, so mutual gaze was held when the turn was maintained, not when it was yielded. However, since this affected all conditions equally and concerned only one of three turn-taking signals, with gestures being the most important [Ehret et al., 2023], we assume that the effect was limited, but should nevertheless be considered when interpreting the results. Still, this error might have inadvertently improved perceived plausibility in the "incorrect" condition, as the gaze behavior was at least occasionally appropriate due to its random character. This may partially explain why this condition was not rated more negatively, though we consider the more plausible explanation to be the subtlety of turn-taking cues in general.

A general consideration whenever working with virtual humans is to not fall into the so-called “uncanny valley”, which can have generally detrimental effects if the virtual representations become very good but some aspects/modalities don’t quite fit the general realism level (cf. [Kätsyri et al., 2015]). While some participants reported feeling uneasy around the virtual humans, others made very positive statements about the realism of the virtual setting and their interlocutors. This brings us to believe that there was a wide range of expectations on how these virtual humans should look and behave, which potentially partly masked the effects we were looking for.

After looking at applicability of the HTR paradigm to evaluate the performance of ECAs, we will next present a study focusing on the surroundings of such an interaction, again utilizing the HTR paradigm.

6.2. Evaluation of Audio-Visual Mismatch

The contents of this section are based on and taken in part from work previously published in [Ehret et al., 2024b].

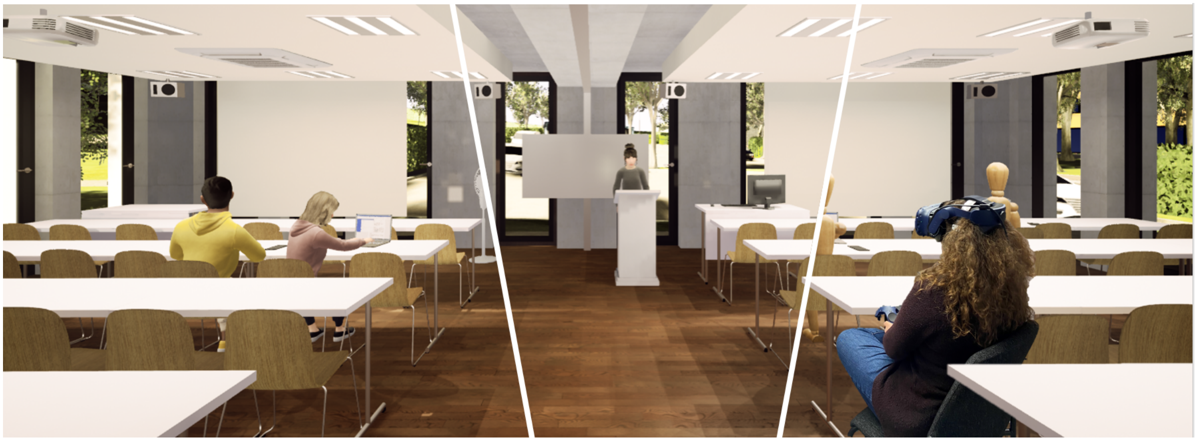


Figure 6.3.: Three levels of background sound source visualization fidelity. Left: Animated virtual characters and, e.g., a moving fan (condition `Animated`); Center: No Visualization of background sound sources (condition `None`); Right: Peers visualized as non-moving wooden mannequins and other sources as static objects (condition `Static`). The female speaker in the center is identically visualized at all levels. A participant wearing a head-mounted display is embedded to show the seating position of participants during the study.

In VR applications, strategically designing visual and acoustic features plays a crucial role in enhancing (social) presence and perceived realism [Kern and Ellermeier, 2020]. Consequently, such design elements also contribute to improved user engagement [Mantovani et al., 2003], encompassing factors like the listening experience and cognitive performance. This strategic design can be implemented through various means. For instance, optimizing visual signals such as using higher-quality renderings [Hvass et al., 2017] or allowing user interactions within the IVE [Witmer and Singer, 1998] have demonstrated efficiency. Moreover, Kim et al. [2018] found that the visual embodiment of ECAs as the user's interaction partner significantly enhances the perceived social presence compared to audio-only interactions. Additionally, integrating animated behavior indicating social cues like gestures and facial expressions during user-commanded actions enhances user engagement more effectively than less interpretable ECA behavior where users may not readily discern the ECA's actions or intentions [Kim et al., 2018]. In the acoustic domain, integrating stimuli coherent with the virtual scene and actions taking place, e.g., tailored soundscapes or footstep sounds [Kern and Ellermeier, 2020], contribute to a more immersive experience [Hendrix and Barfield, 1995].

When visual and acoustic signals closely align semantically, despite minor temporal or spatial differences, they synergize into an integrated audio-visual signal [Laurienti et al., 2004; Spence, 2007]. This phenomenon raises the question of how different visual representations for the same sound influence audio-visual integration and, more importantly, affect the perceived (social) presence and, thus, user engagement. We address this question here specifically concerning background sounds, an integral part of tailored soundscapes, intended to enhance IVE vibrancy without disturbing users or depleting their cognitive resources.

In this study, we examine whether there is a requirement to visually depict background sound sources in VR. We look primarily at background sounds emitted from ECAs populating the IVE but also at those being emitted by non-human scene elements. Thereby we aim to determine the required **audio-visual coherence**, particularly in synchronizing the audio and visual elements, to strike a delicate balance: enhancing (social) presence while mitigating disturbances arising from the representation style of the background noise sources, thereby ensuring the user's optimal performance in the cognitive task at hand, such as attentive listening and efficient processing of speech content. Furthermore, we aimed to explore whether there are differences in the subjective perception of background sounds emanating from ECAs compared to non-human sources. To this end, we compared three distinct **visualization fidelities** in terms of the accuracy of the visual elements (see Fig. 6.3), based on Kim et al. [2018]'s approach, in a within-subject study: (i) without visualizing background sound sources (`None`), (ii) non-animated placeholders without illustrating what is causing the sound (`Static`), and (iii) animated visuals showing the precise origin of the sound (`Animated`).

To prevent participants from focusing directly on the specific audio-visual signals, we utilized a demanding dual-task paradigm [Gagné et al., 2017; Schiller et al., 2023b, 2024], namely again the HTR task performed as dual-task paradigm (see App. A.1). This paradigm was carefully designed to evaluate participants' ability to simultaneously maintain their performance on a primary (listening) task — attentive listening to an ECA's speech content — while engaging in a secondary (vibrotactile) task within the IVE [Schiller et al., 2024], before recalling the memorized information to answer questions about its content. Besides cognitively challenging the user, this dual-task paradigm was instrumental in objectively assessing (i) participants‘ memory performance in the listening task and (ii) participants' accuracy and response times in the secondary task as behavioral indicators for listening effort (LE), in the following referred to as behavioral LE. This is named here differently from the previous section, where we termed it to measure cognitive spare capacity, as this study was a successor to a study published in [Schiller et al., 2024] and the terms should be consistent. In an explorative fashion, following the results of Sec. 6.1, we investigated whether there is a potential correlation between visual fidelity and participants' behavioral LE. We aimed to determine if behavioral LE could serve here as well as a viable objective metric for assessing the optimal audio-visual coherence of an IVE. Complementing this, we collected subjective measures, including user ratings on perceived LE and (social) presence, to gain a nuanced understanding of participants' experiences and the impact of audio-visual coherence on their (social) presence and engagement in an IVE.

The remainder of this section comprises details of our user study (Sec. 6.2.1), results (Sec. 6.2.2), and a discussion of findings (Sec. 6.2.3).

6.2.1. Method

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

Virtual Environment

As the visual setting, a seminar room [Llorca-Bofí and Vorländer, 2021] (see also App. A.5) was chosen and participants consistently occupied a specific desk in the third row (see Fig. 6.3). Six *MetaHuman*[2] models were seated in the room to simulate fellow students, creating a more realistic scenario. The peers were strategically positioned, with some directly in the participants' field of view and two peers in the back. An ECA, representing a female university professor, stood in front of the class at a lectern and was also visualized using a MetaHuman, animated with an idle animation and recorded facial movement when speaking. To fit the speech sound featuring a read-out style, we implemented a gazing schedule such that the ECA looked down towards the lectern at the beginning of each sentence and then alternated her gaze between the virtual peers and the participant, using the gaze dynamics described in Sec. 5.2.

For the IVE's soundscape, we incorporated three classes of **background noise sources**, employed via binaural acoustics as suggested in [Dicke et al., 2010]: (i) Human sounds produced by the peers (i.e., coughing, whispered conversation, laptop typing, or yawning) and non-human sounds originating from sources (ii) within the seminar room (i.e., a fan, window blinds closing, or phones ringing) and (iii) outside (i.e., a car passing by the window or a dog barking). Examples of some background noise source representations can be seen in Fig. 6.4. While animations for objects (e.g., the fan rotating and turning left to right) were simple to implement, animations for the sounds produced by the virtual peers were not readily available. They had to be coordinated with the sounds, which were acoustically recorded under controlled

[2] `https://www.unrealengine.com/metahuman`

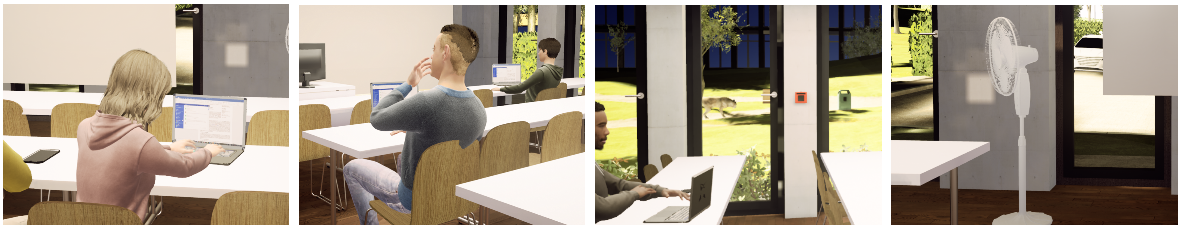


Figure 6.4.: Examples of high fidelity background sound representations. From left to right: A virtual peer typing and a vibrating mobile phone on the table; a peer coughing; a barking dog crossing by outside the window; a fan in the front turning left to right with a spinning rotor.

conditions. To this end, we used the simple motion capture setup described in Sec. 5.3.2 These animations were then manually post-processed to eliminate tracking errors. Additionally, we recorded several seated idle animations to introduce diversity in the movements among the virtual peers. While the fan in front of the seminar room was continuously operating throughout the entire study, we ensured a balanced distribution of both the quantity and class of the remaining background sounds between different runs in the study. To this end, we manually created 22 schedules for the execution of the sounds while the lecturer was speaking, a number derived from the study procedure outlined below.

Study Task

For the primary task, we asked participants to listen to family stories narrated by the ECA standing in front of the class (see Fig. 6.3). We utilized texts from the established Heard Text Recall (HTR) paradigm [Schlittmeier et al., 2023] (see App. A.1). Although there were already recordings available, we recorded 20 (derived from the study procedure detailed below) of these stories in a hemi-anechoic chamber at the Institute for Hearing Technology and Acoustics. A female voice expert (a speech-language pathologist and voice researcher) read the texts, each lasting between 53 and 62 s. This was done since for another experiment in the same setting, published in [Schiller et al., 2024], we needed the text spoken both with a normal and a hoarse voice. In addition to the voice recordings, we captured facial movements, using an iPhone XR and ARKit, to later animate the virtual speaker's face (see Sec. 5.1). Following the text presentation, participants sequentially answered the nine questions per text (see App. A.1) displayed on projection screens located to the left and right of the ECA (see Fig. 6.3). Participants answered these questions verbally and the correctness of their responses was logged by the experimenter.

In order to quantify participants' LE, we employed a dual-task paradigm that comprised the HTR as the primary listening task and a vibrotactile secondary task (see App. A.1). Both were conducted alone (`Single-Task` baseline) and in parallel (`Dual-Task` condition). Specifically, while listening to the ECA's speech, participants reacted to vibration patterns presented via two handheld controllers by clicking a button on either the right or left controller. Based on the cognitive load theory [Paas and Ayres, 2014], a decrease in task performance (more errors or increased response time) was taken as an indicator of higher listening effort in the respective listening condition.

Study Design

We conducted a within-subject study evaluating the influence of audio-visual coherence of background sound sources on perceived (social) presence and user engagement. While the sound itself was kept identical, we varied the visual fidelity across three levels: `None`, `Static`, and `Animated`. In the first level, no representation for the origin of the sounds was shown (see fan and peer missing in the center of Fig. 6.3). In the `Static` condition, objects were placed as placeholders at every source of a sound but they, for example, did not move in the case of the fan or the car outside. Furthermore, we replaced the virtual peers with static, non-animated wooden mannequins to avoid eeriness effects of static highly-detailed ECAs. The `Animated` condition featured representations of background sounds in high fidelity, as described in the previous section.

We expected the following hypotheses to be confirmed:

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

H2 (Social) Presence positively correlates with higher fidelity.

These two hypotheses are motivated, e.g., by the results of Kim et al. [2018]. They compared users' perceptions of three types of virtual interaction partners. These partners were a disembodied voice, an ECA with embodied gestures, and an ECA with both embodied gestures and locomotion. They found that visual embodiment and plausible social behavior, encompassing gestures and locomotion and thus a fully animated ECA, can significantly enhance users' perception of ECAs in terms of social presence, comfort, and engagement, creating a more natural and intuitive interaction experience. Furthermore, also the difference between background noise being produced by other (virtual) humans in comparison to other, non-human sources should be explored. Additionally, we aimed to exploratorily assess, whether behavioral LE is correlated with the fidelity level of background sound source visualizations and thus the audio-visual coherence. We carefully suggested a potential negative correlation, implying that high visual fidelity (`Animated`) with dynamic motions might induce attention capture, potentially disturbing users, and diverting them from their primary cognitive task at hand.

Apparatus

The experiment was implemented using the *Unreal Engine* (version 4.27) and the *StudyFramework* plugin (see Sec. 3). Participants wore an *HTC Vive Pro Eye* while being seated in a sound-proof hearing test booth (A:BOX, Desone Modulare Akustik, Berlin, Germany) with the dimensions 2.3 m × 2.3 m × 1.98 m ($w \times d \times h$) and a room volume of approximately 10.5 m^3. The audio was played over *Sennheiser HD 650* headphones and the binaural dynamic live-rendering using an artificial head HRTF in a 1x1° resolution [Schmitz, 1995] was done using *Virtual Acoustics* [Schäfer et al., 2023] including *RAVEN* [Schröder and Vorländer, 2011] for acoustic room simulation. All background sounds being made by humans used a human singer directivity filter and the sound of outside sound sources was combined with a transmission filter for the windows and played at an appropriate window.

Study Procedure

Upon written informed consent and eligibility check by means of an audiometry-screening ($\leq$ 20 dB HL between 500 Hz and 4 kHz according to a pure-tone audiometry screening, performed with an *Auritec ear 3.0* audiometer), participants were allowed to take part in the study. They were seated in the soundproof booth at a table, position-wise exactly matching the virtual desk in the seminar room. They were equipped with headphones (Sennheiser HD 650) and a head-mounted display (HTC Vive Pro Eye) with two controllers. First, participants completed a practice block of the vibrotactile task (no HTR text, 1 sound schedule), followed by a single vibrotactile baseline block (no HTR text, 1 sound schedule). Next, we presented two HTR texts (see App. A.1) to practice the primary (listening) task (2 HTR texts, 2 sound schedules). All of the above were conducted in the `None` condition. This was followed by the baseline block of the listening task, containing three texts, one for each visual fidelity level in counterbalanced order (3 HTR texts, 3 sound schedules). Afterward, there were also three texts for practicing the dual-task paradigm, counterbalanced in all three conditions (3 HTR texts, 3 sound schedules), followed by a short break. After that, three experimental blocks followed. Each block contained four repetitions of a text being presented with parallel vibrotactile tasks and questions being asked, all using the same visualization fidelity (3 × 4 HTR texts, 3 × 4 sound schedules). The order of these blocks was counterbalanced, and the assignment of texts and background sound schedules were randomized between participants. This procedure resulted in 22 trials in total (6 for practice, 4 for single-task baseline, 12 for dual-task), requiring 20 HTR texts and 22 sound schedules.

After each dual-task experimental block, participants were asked to fill out an intermediate questionnaire, rating their perceived presence using the igroup presence questionnaire [Schubert et al., 2001] (see App. B.14). Social presence of the ECA was rated using the anthropomorphism construct of the Godspeed questionnaire [Bartneck et al., 2009] (see App. B.7) accompanied by the question “The speaker appeared to be sentient (conscious and alive) to me” (German: “Die Sprecherin wirkte auf mich wie ein fühlendes Wesen (mit Bewusstsein, lebendig)”). This last question is one of five items of the Social Presence Survey (SPS) [Bailenson et al., 2001] (see App. B.2) and was used in isolation as in [Ehret et al., 2021] to enhance measuring the perceived anthropomorphism with a further social dimension. Social presence was only evaluated for the speaker in front, as the virtual peers were not visually present at all visual fidelity levels. Additionally, six questions assessing participant’s subjective listening impression were asked, ranging from “How strong was your listening effort?” (the only item referring directly to perceived LE) over ‘To what extent did you feel disturbed or bothered by background noise?" to "How in need of recovery do you feel right now?", based on [Schiller et al., 2023a]. These were accompanied by four questions asking whether participants felt in company apart from the speaker, and how plausible and real the background sound and the speech of the lecturer were perceived, e.g., "To what extent did the background noises resemble a real environment?". All of these were rated on a 5-point Likert scale from "not at all" (German: "gar nicht") to "extremely" (German: "außerordentlich"). After finishing all three blocks, a final questionnaire was posed, asking for demographics and a ranking of the visualization fidelity levels. In this questionnaire, participants were also asked to recall all background sounds they remembered and on the next page to rank the actual sounds (given) by their disturbance. Furthermore, they had to rate the disturbance of three aspects (i.e., missing or static visual representations, and non-continuous background sounds during answering questions). In total, the experiment lasted for around 90 minutes, of which 50 to 60 minutes were spent immersed. The study was approved by the ethics committee of the Faculty of Arts and Humanities (ref. 2022_016_FB7_RWTH Aachen) and the experimental protocol was carried out in accordance with the Declaration of Helsinki.

6.2.2. Results

The analysis was performed using R (version 4.3.2) [R-Core-Team, 2015].

Participants

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

Behavioral Result

To assess whether participants' behavioral LE was affected by the level of visual fidelity, we analyzed secondary (vibrotactile) task performance and response times, as well as the percentage of correctly answered questions of the primary task. Data was modeled using generalized linear mixed-effects models (GLMMs). Regarding secondary task performance, the final GLMM included the fixed effect Condition (`Single-Task Baseline`, `Dual-Task (None)`, `Dual-Task`

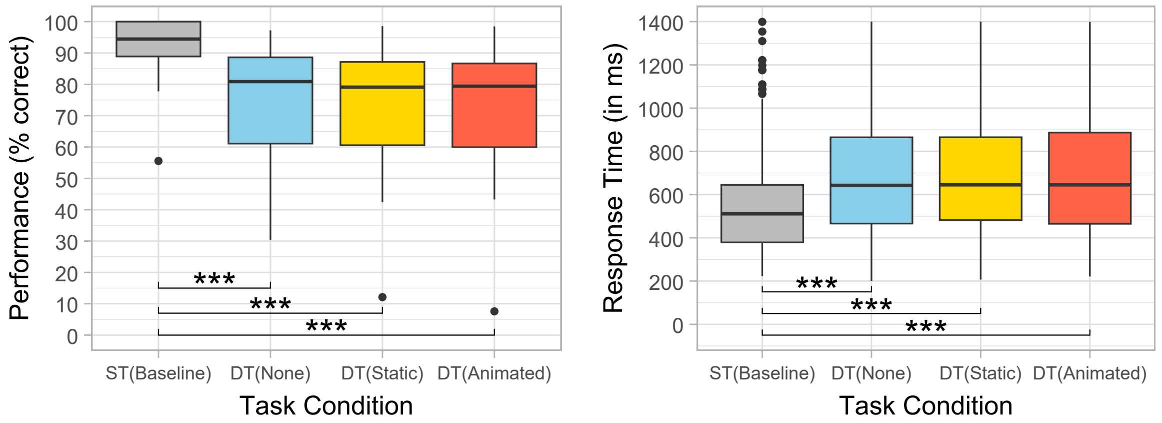


Figure 6.5.: Secondary (vibrotactile) task results for the performance outcomes in % correct (left) and response time in ms (right) as a function of visual fidelity and task condition (`Single-Tasking` (ST) vs. `Dual-Tasking` (DT)). $***$ depicts $p < .001$.

`(Static)`, and `Dual-Task (Animated)`) and random intercepts for Participant, Trial, and Vibration Pattern. This model was specified with a binomial distribution and logit link function, considering that the outcome variable was binary (i.e., either correct or false). Regarding response time, the final GLMM again included the fixed effect Condition and, random intercepts for Participant, Trial, and Vibration Pattern. This model was specified with a Gamma distribution and log link function. Post-hoc comparisons were conducted using the Tukey Method, based on estimated marginal means calculated with the *emmeans* package [Lenth, 2024].

Tab. 6.3 shows the descriptive results for the **primary (HTR) task** of answering the text-related questions. While there were no significant main effects of fidelity or task, there was a significant interaction effect of both ($\chi^2(2) = 10.56$, $p = .005$). However, post-hoc tests did not reveal significant pairwise differences with only single-tasking `Static` vs `Animated` coming close ($p = .075$) and all other p's $> .35$.

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

Table 6.3.: Primary (HTR) task results (percentage of correctly answered questions) as a function of fidelity and single- vs. dual-tasking

Regarding the **secondary (vibrotactile) task**, Fig. 6.5 depicts participants' performance (left) and response times (right) for the `Single-Task Baseline` condition, and the three visual fidelity levels when dual-tasking. Note that, in contrast to the primary task, the secondary task Baseline condition was not performed under each visual fidelity.

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

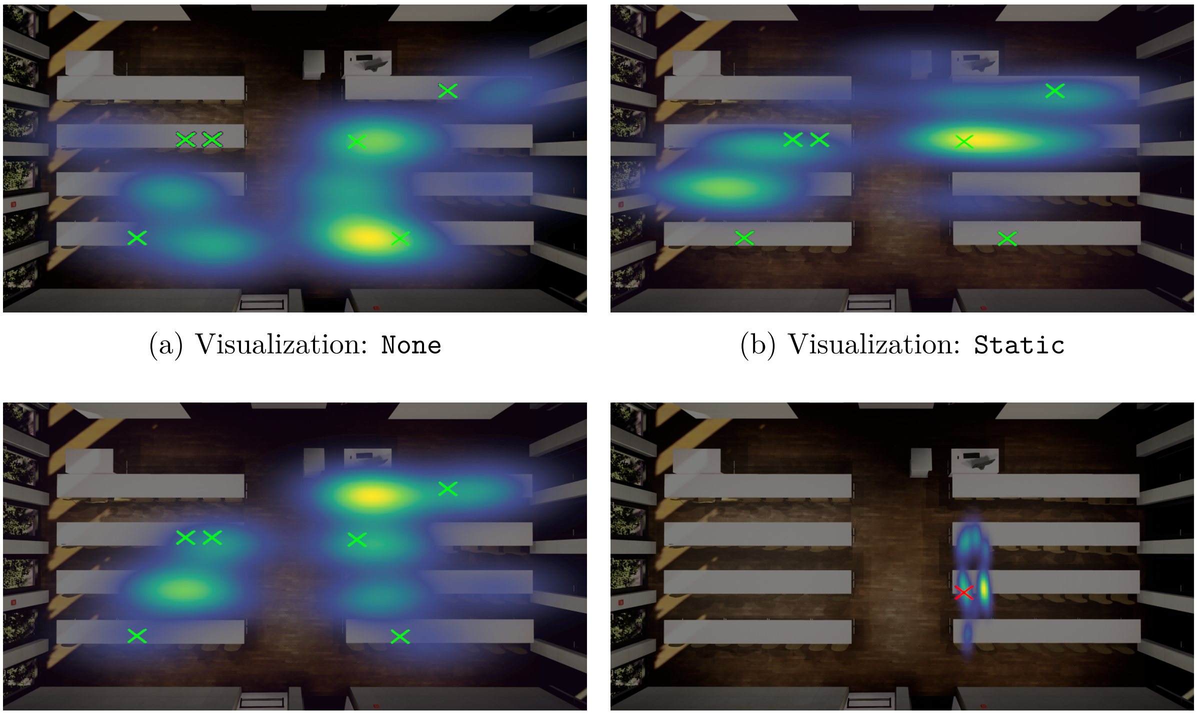
Similar results were obtained for response time measures. Overall, response times also varied significantly across the conditions ($\chi^2(3) = 124.73$, $p < .001$). That is, participants responded fastest in the `Single-Task Baseline` condition but were significantly slower in each of the three `Dual-Task`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

Subjective Evaluation

Following the subjective ratings in the questionnaires between the study blocks and at the end will be analyzed. If not stated differently, we performed 1-way repeated-measures ANOVAs and post-hoc Bonferroni-corrected t-tests for statistic analysis. If the data violated the normality assumption (validated via Shapiro-Wilk's tests), Friedman tests were conducted with potential Bonferroni-corrected Wilcoxon signed-rank tests as post-hoc tests.

Regarding reported presence using the **igroup presence questionnaire** [Schubert et al., 2001], there were no significant differences between the visualization conditions for all subscales (*sense of being there*: $p = .93$, $M = 4.17$, $SD = 1.51$; *Spatial Presence*: $p = .17$, $M = 4.03$, $SD = 1.18$; *Involvement*: $p = .17$, $M = 3.86$, $SD = 1.27$; *Experienced Realism*: $p = .48$, $M = 2.50$, $SD = 1.06$). The same is true for the **Godspeed's** Anthropomorphism scale [Bartneck et al., 2009] ($p = .57$, $M = 2.77$, $SD = 0.86$). Analyzing the answers to the single **social presence** question from the SPS [Bailenson et al., 2001] referring to the speaker only, a Friedman test revealed a significant effect of visualization ($\chi^2(27) = 7.58$, $p = .02$). Post-hoc tests showed a significant effect ($p = .01$) only between `Static` ($M = 2.41$, $SD = 1.12$) and `Animated` ($M = 2.89$, $SD = 1.25$) visualizations, with `None` scoring in between ($M = 2.56$, $SD = 1.16$).

Beyond the originally published results in [Ehret et al., 2024b], we also evaluate the answers of participants to where thy remember their own position and that of their virtual peers. Regarding these positions, there was, however, no significant difference between the visualizations by means of 1-way ANOVAs for the number of peers marked ($M = 2.93$ while 6 were actually present, $p = .51$) nor for the mean distance to their actual seating position ($M = 0.78$ m, $SD = 0.47$, $p = .30$) or the peer positions ($M = 1.07$ m, $SD = 0.38$, $p = .41$). The peer position distance was computed by taking the distance to the closest peer for every reported



(a) Visualization: `None`

(b) Visualization: `Static`

(c) Visualization: `Animated`

(d) Participant’s position

Figure 6.6.: Estimated seating positions of peers in all three visualization conditions and estimated own seating position of the participants (combined from all conditions). The actual positions of the peers are marked with green crosses and the position of the participants in red.

position and then computing the mean distance per participant. Heatmap visualizations of the estimated seating positions can be found in Fig. 6.6. This question was only asked after the first block to measure only the influence of one visualization condition. Although the orders were counterbalanced, due to excluded participants the number of answers per group differed slightly (9 for `None`, 10 for `Static`, and 8 for `Animated`). Looking closer into the data, there was also no difference between the visualizations of peers marked in rows behind the participant ($p = .22$) nor for the amount that participants looked to the back of the classroom ($p = .75$).

After finishing the three study blocks, participants were asked for **preference** with regard to disturbance, realism, and in general, shown in Fig. 6.7. For each condition, a rating of 1 (preferred) to 3 (least preferred) was gathered. A Friedman test of preferences with regard to disturbance showed no significant difference ($p = .15$). However, for realism a significant effect was found ($\chi^2(27) = 18.1$, $p < .001$) and post-hoc tests showed that `Animated` ($M = 1.33$, $SD = 0.62$) was significantly preferred over `Static` ($M = 2.30$, $SD = 0.72$, $p = .001$) and `None` ($M = 2.30$, $SD = 0.72$, $p = .003$), while the latter two were not significantly different. For general preference, a similar trend emerged ($\chi^2(27) = 9.1$, $p = .018$), where `Animated` ($M = 1.56$, $SD = 0.80$) was preferred over `Static` ($M = 2.22$, $SD = 0.70$), approaching

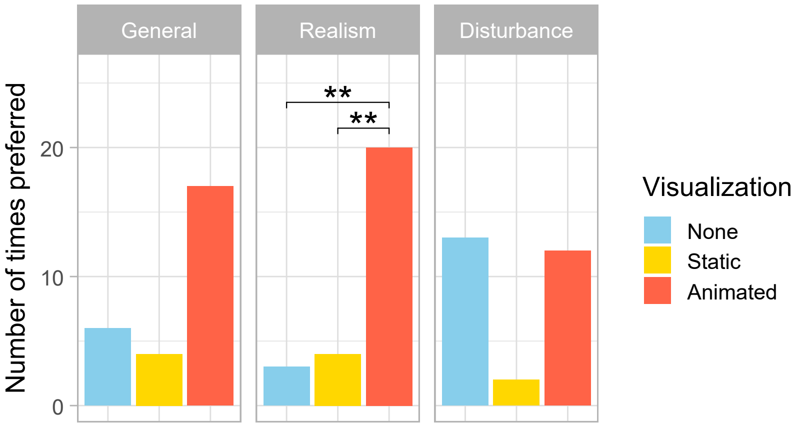


Figure 6.7.: The number of times a visualization was picked as most preferred, in general, and with regard to subjectively perceived realism or disturbance. $**$ indicates $p < .01$.

statistical significance ($p = .052$), and over `None` ($M = 2.22$, $SD = 0.80$, $p = .088$), although these differences did not reach conventional levels of significance.

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

Before revealing in the next question which **background sounds** we included, we asked participants to state which sounds they remember, allowing multiple answers. Twenty-five (93%) remembered conversations between the peers, 22 (81%) mentioned mobile phones, 11 (41%) also stated coughing, and eight (30%) outside noises, or some more specifically cars passing or a dog barking. Furthermore, 7 participants (26%) mentioned a constant background noise or referred more specifically to the fan, while 7 (26%) remembered typing sounds and 2 (7%) specifically referred to yawning. Additionally, sounds that were not part of the simulation were mentioned: moving papers (4 participants, 15%), drinking water (3 participants, 11%), and sounds of chairs (2 participants, 7%). On the next page of the questionnaire, we then asked participants to rank the background sounds that they experienced (explicitly given here) by their annoyance. The results of this ranking can be seen in Fig. 6.8, with mean rankings being: Phone Ringing (2.2), Conversation (2.7), Phone Vibrating (3.3), Coughing (5.2), Laptop Typing (5.2), Throat Clearing (5.9), Yawning (7.0), Dog Barking (7.8), Car Passing (8.4), Fan (9.0), Window Blends (9.7), and Other (11.6).

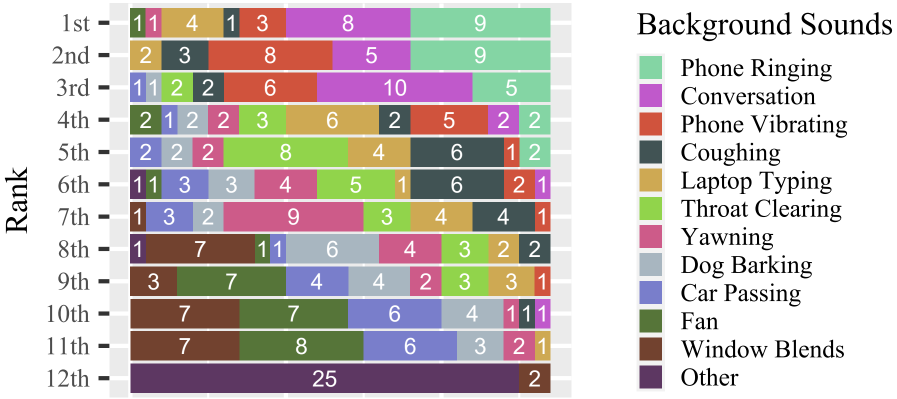


Figure 6.8.: Ranking of the annoyance of the background sounds from most annoying (1st) to least (12th). Numbers indicate the number of occurrences of a raking.

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

6.2.3. Discussion

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

When asked for their general preference, participants clearly preferred background sources being visualized in high fidelity (`Animated`), albeit only significantly with regard to realism. Additionally, the general preference also shows a clear trend towards the `Animated` level. Surprisingly, no clear preference emerged for `Static` or, in our case, partially abstract representations (wooden mannequins) over not visualizing the background noise sources at all (`None`). In fact, participants rated having static representations as more disturbing compared to their absence. These outcomes led us to only partially accept **H1** (expected preference P: $P(\texttt{Animated}) > P(\texttt{Static}) > P(\texttt{None})$) for high-fidelity visualizations. Consequently, when embedding background sound sources, a vivid representation would be the most favorable choice. However, before embedding only placeholders (`Static`), it might be advisable to refrain from introducing any virtual representation (`None`). This is further supported by participants' responses when asked to choose the least disturbing condition: votes were equally distributed between `None` and `Animated`, while `Static` was only preferred by a much smaller fraction (see Fig. 6.7).

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

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

Going beyond the results presented in [Ehret et al., 2024b], we also analyzed participants' gaze behavior for this work. While the statistical evaluation did not show significant differences in the metrics evaluated for where participants remembered their peers, the heatmaps at least show some trend between the visualizations (see Fig. 6.6). Most interestingly, while potentially also not surprising, participants remembered peers more strongly within their visual field of view for the conditions where they actually saw representations, compared to when they could only heard the peers. The peer sitting directly behind participants was remembered more clearly in the audio-only condition, as visible in the heatmaps. This probably indicates that the visual sense is used more dominantly, however, in cases of audio-visual incoherence also acoustical signals are considered more strongly.

Our analysis did not reveal a significant effect of visual fidelity on behavioral indicators of LE and thereby potentially user engagement. Although participants performed significantly weaker and gave slower responses in the secondary task during dual-task conditions, compared to the single-task conditions, this discrepancy was unaffected by the level of visual fidelity. Consequently, our results suggest that participants' LE during a listening task in VR appears to be independent of how accurately the prevailing background sounds are visually represented. This observation is particularly interesting as participants indicated that their subjectively perceived listening impression was negatively influenced by the animated peers (`Animated`), albeit also not significantly for the single perceived LE question. Although not entirely congruent, these ratings partly support our prior assumption that high visual fidelity might divert them from their cognitive task at hand. However, one potential explanation could be occasional glitches or imperfections observed in the animations of the high-fidelity peers (e.g., the back of a peer penetrating the back of the chair shown in Fig. 6.4, 2nd from left). We invested considerable time refining peer movements, yet occasional glitches arose due to inherent limitations in the motion capture method. Importantly, we deliberately chose a diverse array of movements over a limited set of highly refined animations. This decision prioritized a close-to-real-life simulation, emphasizing realistic animation that closely mirrors peers' behaviors. Nevertheless, this poses a **limitation** of the presented study. Another potential shortcoming was revealed by the fact that several participants reported remembering sounds of people drinking or chairs moving, which we did not include in our general soundscapes. Although one participant mentioned in the open-ended comments "The sounds from the workspace of the experimenter were transmitted quite loudly.", suggesting that the talk-back microphone, used for set-up communication with the participant inside the sound-proof booth, was inadvertently left active, these recollections can also be intrusions (false memories which are not uncommon in eyewitness testimony). Repeating the experiment more carefully avoiding inadvertently acoustic noise and examining the impact of audio-visual coherence on potential false memories stands as an intriguing avenue for future research. A third potential limitation of our study is the choice of wooden mannequins as peer representations in `Static`. Despite our intention to mitigate behavioral realism discrepancies, we introduced a visual incongruity between `Static` and `Animated`, particularly as the wooden mannequins' realism contrasted with the overall realistic IVE. This likely resulted in lower social presence ratings towards the ECA in the front in `Static`, suggesting an impact on participants' perceptions. This emphasized the need to carefully consider visual congruity in future studies for an unbiased participant experience. This also has to be considered, when regarding these results in the light of the previous study (Sec. 6.1). In this study the HTR dual-task did not turn out to be as powerful for predicting social presence as the questionnaires, as no significant differences between the presented conditions were found in the behavioral LE. However, in this study no deliberately induced deficiencies in the ECA's performance were present but instead, the focus was on variations in the ECA's surrounding environment. Consequently, attributing causality to any observed correlations between these factors would have been inherently speculative.

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

6.3. Consolidating Discussion

In the two presented studies, we estimated, besides others, the feasibility of the HTR paradigm to evaluate social scenarios in VR with regard to their realism, evaluated for ECAs' performance as well as audio-visual coherence of (potentially human) background sound sources. In general, our results hint that the HTR can be used as a proxy measure to detect severe degradation in ECA performance but fails to differentiate subtle manipulations, such as the degradation of ECA co-verbal gestures or audio-visual coherence of background characters. In the latter cases, questionnaire-based metrics potentially are still capable of producing more insightful results, especially since they can be framed to evaluate particular aspects of the scenario. For example, in the second presented study (Sec. 6.2), we did not specifically measure the social presence of the (potentially only acoustically) present peers, while we could have done so using a well-tailored questionnaire. However, the cognitive spare capacity (more precisely termed behavioral listening effort there) would not be able to differentiate between the presenting ECA in front and the virtual peers, only yielding a single measure per scenario.

Chapter 7

Discussion

We already discussed specific aspects in the individual chapters but would also like to conclude this work with a general discussion merging together the individual findings on a higher level and individually addressing the research objectives presented in Sec. 1

The primary research objective (**RO1**) we addressed in this thesis was to evaluate the influence of multimodal realism of ECAs on perceived social presence. While we were not able to show this influence in all presented studies, it was evident for the ECAs' voice (see Sec. 4.1 and Sec. 6.1) and missing articulation movements (see Sec. 6.1). For the other evaluated realism modalities, like auralization (see Sec. 4.2), co-verbal gestures (see Sec. 5.3.4) and non-verbal turn-taking signals (see Sec. 5.5) we were not able to show significant differences with regard to perceived social presence. However, in most of these evaluations other metrics showed significant differences. For example, during the turn-taking evaluation we found significant differences in the participants' ratings for the conversational flow clarity and post-hoc analysis of the auralization study showed that participants still preferred the directional auralization they rated as more natural. This bring us to believe that social presence is sensitive to more complex perception and subconscious evaluations of the presented ECAs rather than mere naturalness/realism and was therefore not influenced significantly by the latter ones. Another explanation could be that only severe changes have an influence and we therefore failed to find significant differences in all our studies, as the alteration of omitting articulation movement or using synthetic voices stood out more obvious as compared to the other more subtle manipulation. Furthermore, the indifferent results could possibly be caused by the aforementioned shortcomings of questionnaire-based social presence evaluation instruments which potentially are not sensitive enough to capture such subtle differences. Nevertheless, our results can be a valuable starting point for ECA developers to prioritize limited development resources if social presence is the primary goal.

We further hypothesize that one aspect of this can be the relation to the perceptual mismatch hypothesis [Kätsyri et al., 2015]. It states that the behavior of an ECA across all modalities has to reach a certain realism to elicit more positive affinity and thereby potentially also social presence. Furthermore, Kätsyri et al. [2015] state that "the reviewed findings that individuals are increasingly sensitive to atypical features on more human-like characters would suggest that avoiding the uncanny valley will become exponentially more difficult as the characters' overall appearance approaches the level of full human-likeness." This potentially led to the indifferent findings when looking into the directional auralization of ECAs' voices in Sec. 4.2. Besides the participant comments towards the difficulties in filling out the *SPS* questionnaire (see App. B.2), there were also comments rating the overall realism of the simulation too low, which potentially masked the more advanced realism gains we achieved by adding speaker directivity. This becomes especially problematic since very realistic visualizations like *MetaHumans* recently became commonly available and, for example, Seymour et al. [2021] even states that they visually crossed the uncanny valley. However, often other modalities like the behavioral realism lack behind, even increasing this discrepancy.

Another caveat with regard to behavioral realism is that in contrast to visual fidelity there is not an objective more realistic realization due to large interpersonal differences in actual human behavior. For example, all humans have their very individual style of moving. We encountered that especially when observing gazing behavior in different individuals, which can differ strongly with some people nearly always averting gaze while speaking and others focusing the interlocutor throughout their speech act. Therefore, there exists not a single correct realization, which was also shown with regard to back-channels, which can be adapted to portray specific personalities [Sevin et al., 2010], often using the classes from the Big Five personality traits [Norman, 1963]. This also poses a major challenge to statistical models realizing such behavior, which tend towards the mean of the input data. However, this mean performance potentially yields something less natural than each of the individual performances in the training data. Nevertheless, recent implementations for co-verbal gestures found solutions to solve this problem, for example, by adding speaker identifiers as input [Wolfert et al., 2022].

Regarding research objective **RO2**, which was to evaluate the influence of different turn-taking cues on the perception of turn-taking clearness, we presented two studies in Sec. 5.5. While showing that manipulating co-verbal gestures in the gap between sentences (e.g., holding the gesture throughout the gap if the turn is held) significantly improved clearness, we were unable to show the same for inhalation and gazing cues. These, however, showed non-significant tendencies to positively influence the perception of turn-taking. Furthermore, we evaluated whether additional listeners following the turn-taking cues given and thereby producing additional turn-taking cues themselves would benefit perception. Contrary to our expectation these listening agents did not improve perception, at least not using the evaluated social model developed. Taken together, we believe that co-verbal channels can provide additional and helpful turn-taking cues. Further research building on the presented finding is required to evaluate if the presented cues presented in this thesis which did not show significant improvements, could be further enhanced to increase their efficiency.

Another key finding presented in Sec. 5.5.5 and Sec. 6.2, which addresses our third research objective (**RO3**), is that designers of social VR simulations should carefully evaluate the inclusion of additional background characters. The presence of such characters can significantly impact user experience and perceptions, requiring thoughtful consideration in the design process. The results of both studies leaned towards rather not including background characters if they could not be realized appropriately. In Sec. 5.5.5 we found that agents added as additional listeners to a virtual conversation with ECAs were perceived as worse with regard to the clearness of the conversational flow than not having them at all when they did not adhered to a social model but rather performed random listening behavior that was not context-sensitive to the conversation. Similar results were found in Sec. 6.2, where static wooden mannequins were rated equally unrealistic to not having visual representation of the acoustic background sources at all and even more disturbing than not having any.

The last central research objective of this work is whether the HTR paradigm performed in a dual-task (see Sec. 6.1) can be used as objective proxy for measuring social presence (**RO4**). This is obviously also subject to the aforementioned restrictions. While in tendency we were able to find similar differences to those found by traditional social presence questionnaires, there are a few limitations that should be carefully considered. First of all, we were only able to find significances for the strongest performance degraders used in the evaluation study, like replacing the natural, recorded speech material with rather artificial synthetic ones. For example, objectively degraded gestures or omitting co-verbal gestures altogether failed to show significant differences both in the employed social presence questionnaires as well in the HTR performance measuring cognitive spare capacity. A possible explanation for this is that the HTR task—listening to two speakers presenting family stories—is not a very social task with the potential of sparking a lot of social presence as the ECAs content-wise do not react to the users. Furthermore, the stories by themselves feel constructed and are presented in a non-spontaneous way but feeling rather scripted and rigid. This task was specifically design to be more realistic when measuring memory performance and listening effort during controlled psychology studies (cf. [Schlittmeier et al., 2023]) than tasks that were traditionally used in those experiments, like repeating single words or numbers. However, it was not particularly designed to mimic social interactions with all their delicate relational dynamics. Therefore, it still lacks the interactivity and natural feeling potentially required to spark high perceived social presence in general. A potential solution to overcome this is to use large language models (LLMs) to produce dynamic speech content, potentially also reacting to the listeners, while still being able to make sure that all information is shared by the user. This would, however, also require to synthesize the speech and gestures which in turn could again degrade fidelity and controllability. A last general consideration when approaching the correlation between an objective measure and a questionnaire metric, is that when the confidence into the questionnaire instrument is low we cannot be sure whether the objective metric really measures the desired concept if its discriminative power goes beyond that of the questionnaire or whether potentially something different is measures (see, e.g., [Bailenson et al., 2004]). This consideration was also addressed by Slater [2004] who generally challenged the usage of questionnaires and in conclusion also proposed other measurement techniques, like the configuration transition method [Slater et al., 2010] or sentiment analysis [Slater et al., 2022]. Therefore, the usage of the HTR paradigm as proxy for social presence has to be considered with care and other metrics should still be considered, including abandoning the concept of social presence altogether if mere naturalness of a modified audio-visual behavior realization seems to suffice to answer the underlying research question.

Last but not least, a topic that reoccurred during post-study interviews and questionnaires is whether it is actually desirable—and the highest design goal—to represent the reality as closely as possible when simulating ECAs. While this potentially is very desirable for psychological studies where results should be generalizable to the real world, there can be VR scenarios were this does not hold. For example, in the studies regarding speaker directivity (Sec. 4.2) some participants consistently preferred the less natural omnidirectional directivity. While this does not depict speech perception in reality as closely it potentially leads to better intelligibility, especially when the ECA is facing away from the user, where it would otherwise sound muffled and less loud. In the case of using an ECA as guide in a virtual museum where understanding what was said is of high importance, this requirement might outweigh the gained realism and be preferred. Since we are in a virtual environment this kind of "super-natural" behavior can go even further. As described by Xenakis et al. [2023] an ECA could make eye contact with multiple human users at the same time when desirable or non-attentive gazing could be "fixed" during mediated communication [Roth et al., 2019a]. Wieland et al. [2023] go even further and augment the gaze signals of ECAs by means of additional visual, auditory or even tactile cues to make them clearer, for example, for visually impaired persons. This way the full potential of the VR technology can be leveraged rather than "just reconstructing reality". We would therefore argue that, when designing ECAs, it should always be carefully considered whether pure realism is really the highest goal of the application or which (communicative) functions should be fulfilled and whether potentially surpassing reality would be beneficial. We also already made a step into this direction when designing our turn-taking cues (see Sec. 5.5) where we tried to make the utilized natural cues clear and understandable rather than fully covering the fuzziness of human behavior while still utilizing participants' trained knowledge of how to interpret non-verbal turn-taking cues.

7.1. Future Work

The presented work can only be considered a first step towards implementing truly believable ECAs and employing them for manifold applications, ultimately enhancing the effectiveness of ECAs in a wide range of use-cases.

In Sec. 5.4, we provided an overview of back-channels and detailed our implementation of a subset of these elements within an ECA's behavior. However, it would be meaningful to further evaluate how these can facilitate interactions with ECAs that act as advanced (emotional) user interface to steer and manipulate an application, as already often available in the form of voice assistants, like *Alexa* or *Siri* (see [Hoy, 2018]). If these were embedded into a VR/AR application as ECAs their embodiment could enhance their communication efficiency, particularly through the use of back-channeling to convey understanding or confusion already while the user is speaking. This real-time feedback allows users to adjust their communication strategies even before completing their requests, improving overall interaction quality and clarifying the assistant's capabilities. Additionally, back-channel tracking of the users is an interesting endeavour to generate more contextually appropriate ECA behavior. We took an initial step in this direction by utilizing tracking capabilities in a *Vive Pro Eye* with an attached tracker for lower facial movements to assess attitudes toward ECAs in a classroom simulator designed for teacher training. In a pilot study with aspiring and experienced teachers, the system was able to reasonably track the attitudes of the teachers and trigger appropriate responses by the virtual students, while the visualized behavior fell short of representing organic classrooms. The promising results with regard to attitude tracking from this project highlight the potential for further exploration in this area.

Another promising avenue to improve on the shortcomings of the presented social presence measures, is to train machine learning models to judge perceived social presence based on observing and fusing different behavioral aspects of participants. Nevertheless, such models would also be subject to the aforementioned challenge of not having a certain ground truth as questionnaires are also deemed not the perfect instrument. This would, however, help to improve and guide the further conceptualization of social presence as proposed by Sterna and Zibrek [2021] and integrate more delicate user behavior and potentially also temporal developments if suitable models are used.

Future research should further investigate the masking effects of individual behavioral modalities on social presence, specifically whether the low fidelity realization of one modality can obscure the influence of another that may retain a higher level of realism even in its degraded version. This is particularly relevant given our findings from Sec. 4.2, where added acoustic realism from static or dynamic speaker directivity went unnoticed in social presence measures. Examining these effects, such as determining if degraded gestures have minimal impact when articulation movements are absent, could enhance our understanding of how different modalities interact and contribute to social presence, ultimately informing more effective design strategies for social VR simulations. This is also particularly important when designing ECA interactions that closely resemble real-life scenarios, ensuring high ecological validity in psychology studies.

Another important aspect to further evaluate is the integration of more advanced machine learning techniques for gesture improvements instead of the algorithmic ones presented in Sec. 5.3.3. Furthermore, future research should carefully evaluate the appropriateness of using generated gestures, taking into account their potential advantages and limitations as discussed in Sec. 5.3.1. Automatic gesture generation has improved significantly in quality over recent years, and the advantages of quickly generating diverse, potentially artifact-free gestures may outweigh the disadvantages of not having complete control.

Lastly, evaluating the influence of the display technology used is essential, especially given its significant relevance to us as VR researchers. There exists research showing that VR improves social presence (e.g., [Guimarães et al., 2020]). However, there are no conclusive results towards the influence of the immersiveness of the display system. Sun and Botev [2023] compared desktop, AR, and VR presentations but did not find a significant influence on the willingness to delegate critical tasks to the embedded ECAs. We planned to use the experimental setup as presented in Sec. 6.1 to evaluate differences between an HMD and the AixCAVE (see App. A.5.1) with regard to the measured subjective and objective metrics. We hypothesize that the fact that one can see one's own body in a CAVE environment potentially increases social presence compared to an HMD with "only" a body-avatar. This experiment was already conceptualized and implemented but its execution had to be postponed due to time constraints.

By systematically addressing these research directions, we can deepen our understanding of social presence and directly enhance the effectiveness of ECAs across various applications. This progress will lead to more engaging and realistic interactions, improving user experience in various contexts.

Chapter 8

Conclusion

In this dissertation, we have explored the multimodal facets of embodied conversational agents (ECAs) during verbal interactions, focusing on enhancing the naturalness of human-agent interactions to improve their effectiveness and thus user engagement. By addressing critical challenges in creating believable ECAs, such as voice quality, prosody, speech auralization, gesture quality, and turn-taking mechanisms, we have generated valuable insights into how to effectively improve ECA behavior while also advancing our understanding of their effects on perceived social presence, a vital metric for assessing the human-like qualities of ECAs.

Our findings reveal that only significant enhancements, like utilizing recorded voice instead of synthesized speech and implementing appropriate articulation movements, truly elevate social presence. While more subtle changes did not yield substantial improvements in this metric, they provided intriguing insights into participant preferences regarding naturalness of the ECA and the resulting interaction. For example, participants disagreed on rating the naturalness of auralization using speaker directivity, but consistently preferred what they deemed more natural.

The introduction of the *StudyFramework* has equipped both experienced developers and novices with a valuable tool for conducting factorial-design user studies in virtual reality (VR), demonstrating its usability across different expertise levels. Additionally, we proposed the Heard Text Recall (HTR) paradigm as an innovative objective metric for social presence, bridging gaps in existing subjective assessments and laying groundwork for quantifiable evaluations. While our evaluation showed promising results in its capability to differentiate severe changes in ECA behavior, it was unable to discriminate more subtle changes. However, the latter was also partially true for the questionnaire-based instruments employed.

Building on these contributions, future research can delve deeper into enhancing behavioral aspects such as back-channeling capabilities or investigating how low fidelity in one modality may mask positive effects from other improved aspects. The multimodal realizations presented here are readily available as an extensible Unreal Engine plugin, providing researchers with a solid foundation to further explore these exciting avenues without needing to reimplement previous work.

Together, these advancements not only contribute to our understanding of ECAs but also pave the way for creating more engaging and realistic interactions that can transform various applications——from education to entertainment——ultimately enriching user experiences in various domains.

Appendix A

Study Material

This appendix provides additional information about study material used for experiments described in this thesis. Furthermore, several scene used during studies are shown in more detail and some additional information about technical systems used throughout this work is given.

A.1. Heard Text Recall Paradigm

For several studies we utilized texts from the established Heard Text Recall (HTR) paradigm [Schlittmeier et al., 2023], consisting of 34 German texts providing information on three generations of family members with regard to their relationships, professions, hobbies, etc. All stories are self-contained and do not have an overlap with other stories, especially concerning the names of the family members. Each text is accompanied by nine content-questions asking for pieces of this information, which can in some cases only be answered by combining information given in separate statements. Recordings of the texts of a female and a male speaker and respective face trackings (see Sec. 5.1.3) are available in [Ermert et al., 2022]. For presentations by two speakers, the HTR material also provides information which sentence should be spoken by which speaker to provide natural turn shifts while not changing the speaker for every sentence. An example text with content questions can be found in Tab. A.1. Questions regarding specific persons, for example, the second question, have always to be answered by giving the name of the person, e.g., “Edith” rather than “the mother”. The latter answer would be rated as wrong, since participants are always explicitly told to answer by giving names and not other explanations if asked for a person.

Table A.1.: An example text from the HTR paradigm [Schlittmeier et al., 2023], including questions and English translations. The speaker (Sp.) of each sentence is given and questions are shown at the bottom.

Sp.	**German**	**English**
a	Seit ihrer Kindheit hat Emma den Wunsch Sängerin zu werden.	Emma has wanted to be a singer since she was a child.
b	Schon ihre Oma Anneliese war in ihren jungen Jahren eine begnadete Sopranistin.	Her grandmother Anneliese was already a gifted soprano in her younger years.
b	Sie ist letztes Jahr im Alter von 85 Jahren verstorben.	She passed away last year at the age of 85.
a	Emma erinnert sich gerne an die Nachmittage mit ihrer Oma, an denen sie gemeinsam im großen Liederbuch schmökerten.	Emma likes to remember the afternoons with her grandma when they used to browse through the big songbook together.
a	Emmas Geschwister hatten währenddessen im Wohnzimmer Stadt-Land-Fluss gespielt.	Meanwhile, Emma's siblings were playing City-Country-River in the living room.
b	Konrad und Leonie war kein musikalisches Talent in die Wiege gelegt worden.	Konrad and Leonie were not born with musical talent.
b	Sie spielen in ihrer Freizeit am liebsten Gesellschaftsspiele, genau wie ihr Vater Rolf.	They like to play board games in their free time, just like their father Rolf.
b	Er konnte sich besonders gut beim Bingo von seinem stressigen Berufsleben als Bänker erholen.	He was particularly good at bingo to recover from his stressful professional life as a banker.
a	Seine Frau Edith hingegen spielte gerne Geige.	His wife Edith, on the other hand, enjoyed playing the violin.
a	Sie hatte jedes Jahr an Heiligabend ihre Mutter Anneliese und ihre Tochter Emma mit sanften Klängen begleitet, während die beiden Weihnachtslieder sangen.	Every year on Christmas Eve, she accompanied her mother Anneliese and her daughter Emma with soft sounds while they sang Christmas carols.
	Questions:	
	Wer hat den Wunsch Sängerin zu werden?	Who has the desire to become a singer?
	Wer spielte jedes Jahr an Heiligabend Geige?	Who played the violin every year on Christmas Eve?
	Wer erholt sich am besten beim Bingo?	Who recovers best at playing Bingo?
	Wie viele Geschwister hat Emma?	How many siblings does Emma have?
	In welchem Verwandtschaftsverhältnis steht Konrad zu Edith?	What is the relationship between Konrad and Edith?
	In welchem Verwandtschaftsverhältnis steht Leonie zu Anneliese?	What is Leonie's relationship to Anneliese?
	Welches Gesellschaftsspiel spielte Leonie im Wohnzimmer?	Which game did Leonie play in the living room?
	Welchen Beruf übt Rolf aus?	What is Rolf's profession?
	In welchem Alter starb Annelise?	At what age did Annelise die?

This paradigm can either be applied by only performing the listening task. However, it can also be executed in a **dual-task paradigm** [Gagné et al., 2017], where the primary listening task is accompanied by a secondary task. As proposed in [Mohanathasan et al., 2022], we used a vibrotactile task as secondary task if the HTR was to be performed in a dual-task paradigm. While listening to the stories being told, participants had to respond to vibrational pattern being presented via the controllers. The pattern consisted of two vibrations, either short (200 ms) or long (600 ms), with a gap of 300 ms in between. Always one of the resulting four different pattern (short-short, short-long, long-short, or long-long) is randomly presented and participants had to click a dedicated button either on the right controller, if both vibrations had the same length, or on the left controller otherwise during the 2,000 ms following the pattern presentation. It was logged whether the answer given was correct or not and how long after the offset of the last vibration the button was pressed. If neither button was pressed during these 2,000 ms, a missed trial was logged. After 2,000 ms the next random pattern is presented.

A.2. Materials used in the Prosody Study

A.2.1. Texts of Scenarios S2 to S4

Here we present the three additional texts used in the prosody study presented in Sec. 4.1. The first text is already shown in Sec. 4.1.

Table A.2.: Conversation of the second scenario given by a male ECA (A) and a female ECA (B). Accented syllables are written in bold face and the nuclear accent in bold capitals. The *adequate* prosody was used for S_{human} whereas *TTS prosody* was used for $S_{\text{human+TTS}}$ as well as S_{TTS}. For the latter, inadequate nuclear accents are highlighted in red. An English translation of the text is given in the right-hand column.

S2	German (adequate prosody)	German (TTS prosody)	English translation
A	**HAL**lo, habt ihr beide morgen **Zeit** für einen ge**müt**lichen **SPIE**leabend? Ich hätte mal wieder **Lust** auf eine Runde **Sied**ler von Ca**TAN**.	**HAL**lo, **habt** ihr beide morgen **Zeit** für einen ge**müt**lichen **SPIE**leabend? Ich **hät**te mal wieder **Lust** auf eine **Run**de Siedler von Ca**TAN**.	Hello, do you both have time tomorrow for a cozy game evening? I would like to play “Settlers of Catan” again.
B	**Das** ist ja eine **tol**le **IDEE**, aber **mor**gen sind wir leider schon ver**PLANT**.	**Das** ist ja eine **tol**le **IDEE**, aber **mor**gen sind wir **lei**der schon ver**PLANT**.	That’s a great idea but we already have other plans for tomorrow.
A	**SCHA**de. Wie **wä**re es denn alterna**tiv** mit einem Abend gegen **En**de nächster **WO**che?	**SCHA**de. **Wie** wäre es denn alterna**tiv** mit einem **A**bend gegen **En**de **nächs**ter **WO**che?	Too bad. What about an evening towards the end of next week as an alternative?
B	**Ja**, das klingt für **mich** erst einmal **GUT**, aber ich müsste **ER**win sicherheitshalber noch fragen. Ich melde mich **mor**gen zu**RÜCK**.	**Ja**, das **klingt** für mich **erst** einmal **GUT**, aber ich **müss**te Er**win** **SI**cherheitshalber noch fragen. Ich **mel**de mich **mor**gen zu**RÜCK**.	Yes, that sounds good to me, but I have to ask Erwin to be on the safe side. I’ll get back to you tomorrow.

Table A.3.: Conversation of the third and fourth scenario given by a male ECA (A) and a female ECA (B). Accented syllables are written in bold face and the nuclear accent in bold capitals. The *adequate* prosody was used for S_{human} whereas *TTS prosody* was used for $S_{\text{human+TTS}}$ as well as S_{TTS}. For the latter, inadequate nuclear accents are highlighted in red. An English translation of the text is given in the right-hand column.

S3	**German (adequate prosody)**	**German (TTS prosody)**	**English translation**
A	**HAL**lo, ich **woll**te mich nach **Flü**gen für **ei**ne Person von **Ham**burg nach **MEL**bourne erkundigen - am **lieb**sten **BUSI**ness Class.	**HAL**lo, ich **woll**te mich nach **Flü**gen für eine Per**son** von **Ham**burg nach **Mel**bourne er**KUN**digen - am **lieb**sten **BUSI**ness Class.	Hello, I want to book a flight for one person from Hamburg to Melbourne - preferably business class.
B	**GER**ne – an welches **Da**tum hätten Sie dabei ge**DACHT**?	**GER**ne – an **wel**ches **Da**tum hätten Sie dabei ge**DACHT**?	You're welcome - which date do you have in mind?
A	Ich müsste am **neun**ten oder **zehn**ten **Mai** in Australien **AN**kommen. Wenn es **geht** bereits am **VOR**mittag, sodass ich noch im **Hel**len in mein Ho**TEL** komme.	Ich **müss**te am **neun**ten oder **zehn**ten **Mai** in Aus**TRA**lien ankommen. Wenn es **geht** bereits am **VOR**mittag, sodass ich noch im **Hel**len in mein Ho**TEL** komme.	I have to arrive in Australia on the 9th or 10th of May. I would prefer landing in the morning, to be able to arrive in the hotel during daytime.
B	Alles **KLAR**, dann werde ich Ihnen **gleich** mal ein paar **Mög**lichkeiten he**raus**suchen und **ZU**kommen lassen.	Alles **KLAR**, **dann** werde ich Ihnen **gleich** mal ein paar **Mög**lichkeiten heraussuchen und **ZU**kommen lassen.	All right, then I'll select some options and send them to you right away.

S4	**German (adequate prosody)**	**German (TTS prosody)**	**English translation**
A	**HAL**lo, **wann** und **wo** soll ich Dich **mor**gen zum **FUSS**balltraining abholen?	**HAL**lo, **wann** und wo soll ich Dich **mor**gen zum **Fuss**balltraining **AB**holen?	Hello, when and where should I pick you up tomorrow for our football training?
B	Wie **wä**re es mit **sie**ben Uhr **drei**ssig an der **Bus**haltestelle vor der **AN**nakirche? Das **liegt** ja bei **dir** auf dem **WEG**.	**Wie** wäre es mit **sie**ben Uhr **drei**ssig an der **Bus**haltestelle vor der **AN**nakirche? **Das** liegt ja bei dir auf dem **WEG**.	How about 7:30 at the bus stop in front of the Annakirche? That's on your way.
A	**Su**per, das **PASST**. Aber sei **dies**mal bitte **pünkt**licher als die **LETZ**ten beiden Male, sonst sind wir wieder zu **spät** und müssen **fünf** **STRAF**runden laufen.	**Su**per, das **PASST**. **A**ber sei **dies**mal bitte **pünkt**licher als die **letz**ten beiden **MA**le, **sonst** sind wir wieder zu **spät** und müssen **fünf** **STRAF**runden laufen.	Great, that works out. Please be more punctual this time compared to the last two times, otherwise we'll be late again and have to run five penalty laps.
B	Alles **klar**, ich werde mir **MÜ**he geben. Ich **stel**le mir gleich einen **We**cker damit ich **pünkt**lich **LOS**gehe.	Alles **klar**, ich werde mir **MÜ**he geben. Ich **stel**le mir **gleich** einen **We**cker **da**mit ich **pünkt**lich **LOS**gehe.	All right, I'll do my best. I'll set an alarm right away to make sure I leave on time.

A.2.2. Visual Stimuli

These are the visual backgrounds and ECAs used in the study described in Sec. 4.1.

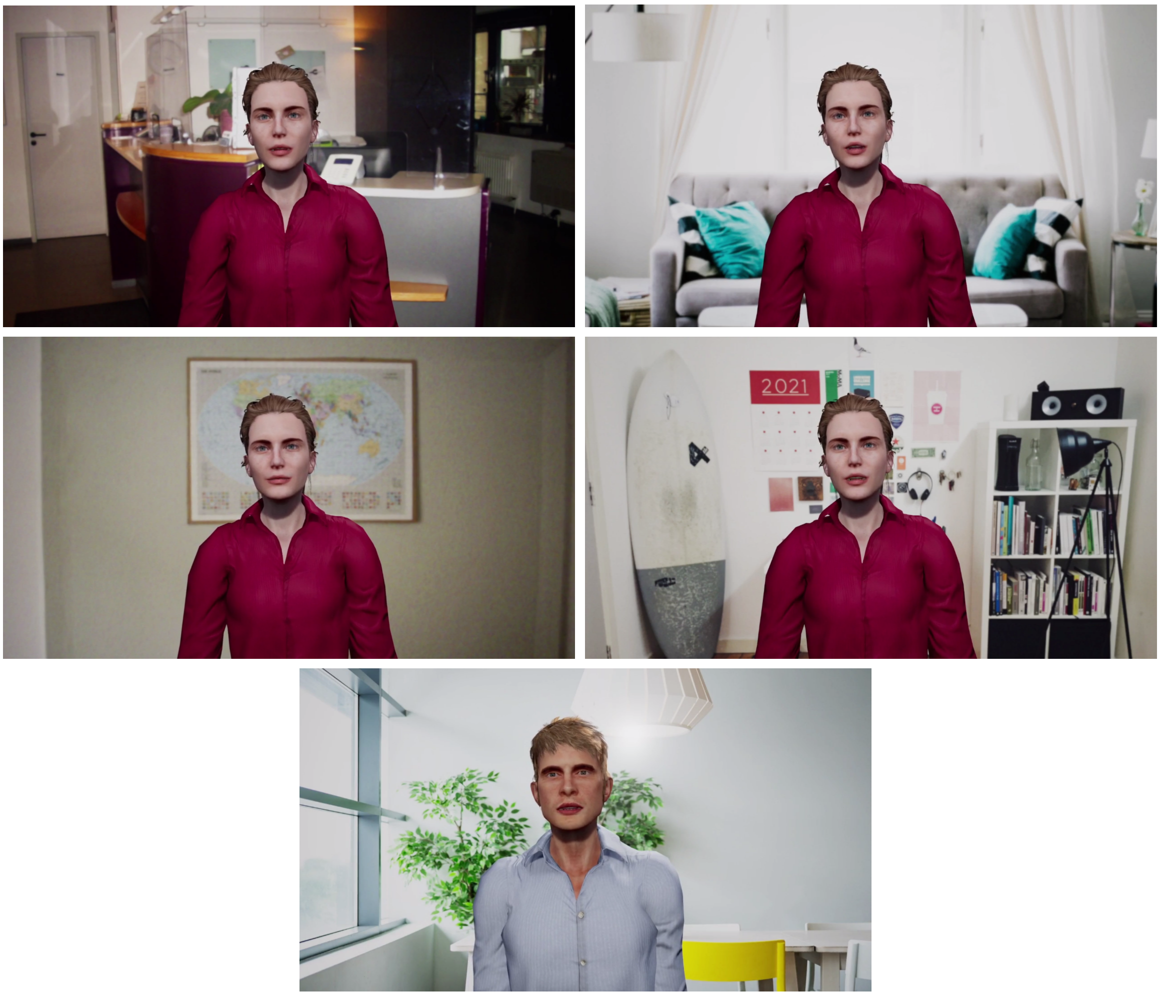


Figure A.1.: Example screenshots from the 4 videos. Top left to bottom right: the backgrounds used for the female ECA in S1 to S4. The male ECA (bottom) was shown in front of the same background for all four scenarios.

A.3. Full Questionnaire Statements used in the StudyFramework Evaluation

Here we present the questionnaire used in the evaluation of the *StudyFramework*, presented in Sec. 3

Nr.	**Full Statement**	**Left Anchor (1)**	**Right Anchor (5)**
Q1	Experience with Unreal Engine (UE) before starting the project:	Completely new to UE	Very familiar with UE
Q2	Experience with C++ programming before starting the project:	Completely new to C++	Expert in C++
Q3	Experience with factorial study design before starting the project:	Complete novice	Expert in study design
Q4	I found the usage of the study setup actor clear and easy.	Strongly Disagree	Strongly Agree
Q5	The different randomization and ordering options were clear to me.	Strongly Disagree	Strongly Agree
Q6	The Wiki helped in finding the information I needed.	Strongly Disagree	Strongly Agree
Q7	The C++/Blueprint Interfaces provided were sufficient.	Strongly Disagree	Strongly Agree
Q8	I had to look into the source code frequently to understand what was going on.	Strongly Disagree	Strongly Agree
Q9	I needed a lot of help to develop the study.	Strongly Disagree	Strongly Agree
Q10	The provided control screen helped conducting the study.	Strongly Disagree	Strongly Agree
Q11	I felt in full control over the study.	Strongly Disagree	Strongly Agree
Q12	I felt confident that I could use the provided recovery options to handle every possible situation.	Strongly Disagree	Strongly Agree
Q13	I used the ”Show Conditions” Option regularly.	Strongly Disagree	Strongly Agree

Table A.4.: Full Statements and their scale’s anchors used in the questionnaire to evaluate the *StudyFramework*. Additionally, the System Usability Scale (SUS), two rankings for most and least helpful features, free feedback fields regarding implementation and execution, as well as demographics, were part of the questionnaire.

A.4. Study Scenes

In the following, further visualizations of used study scenes are shown, for those where the scenes are not already fully shown in the respective chapters, and also going beyond the material published in the respective publications.

A.4.1. Somerset House

This section shows further images of the Somerset House scene used in the dynamic directivity study (Sec. 4.2.3), both the outdoor scene and the indoor museum scene. The outdoor scene was created using a freely available scanned model from Sketchfab: Somerset House site survey scan 2019 by Kimchi and Chips art collective: `https://skfb.ly/6svNI`.

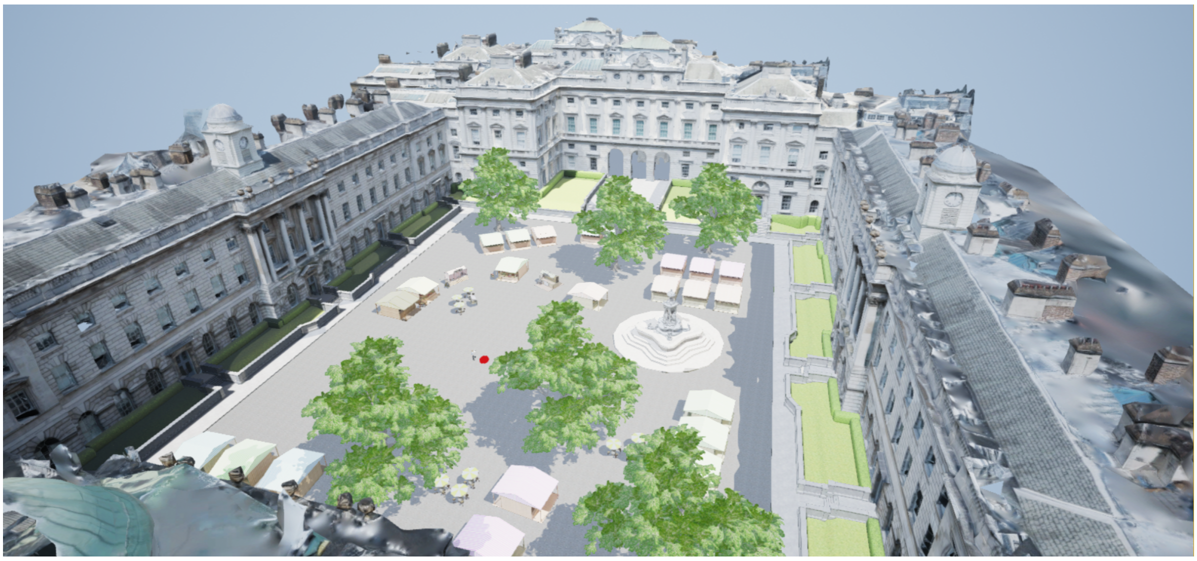


Figure A.2.: Top view of the Somerset House scene. The position of the ECA and the red circle for the participant to stand in are in the middle of the scene.

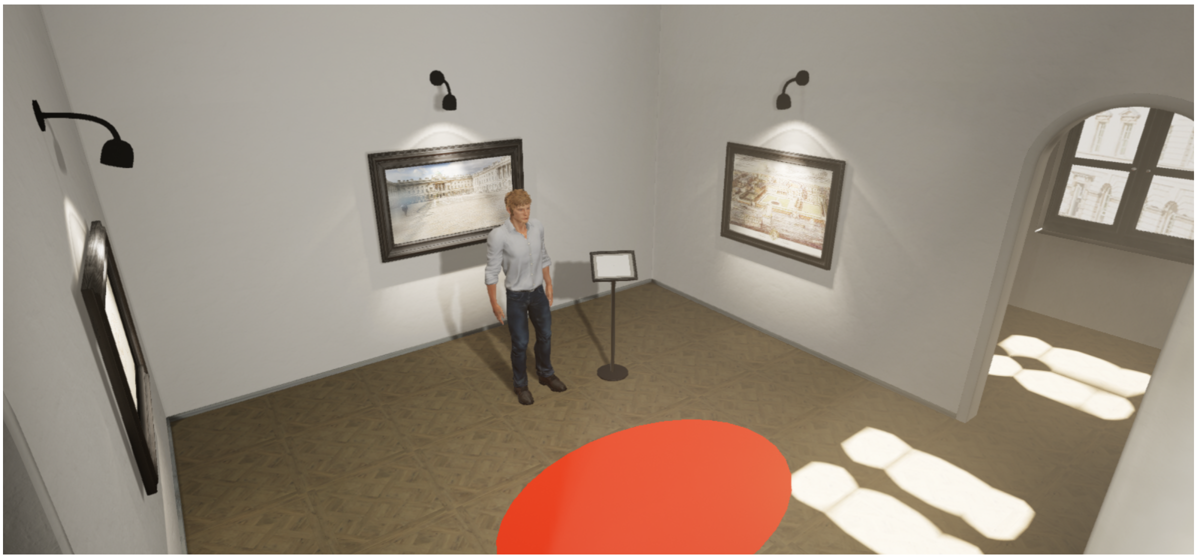


Figure A.3.: The museum scene from the corner of the room. Through the window a view into the courtyard of the Somerset House is visible.

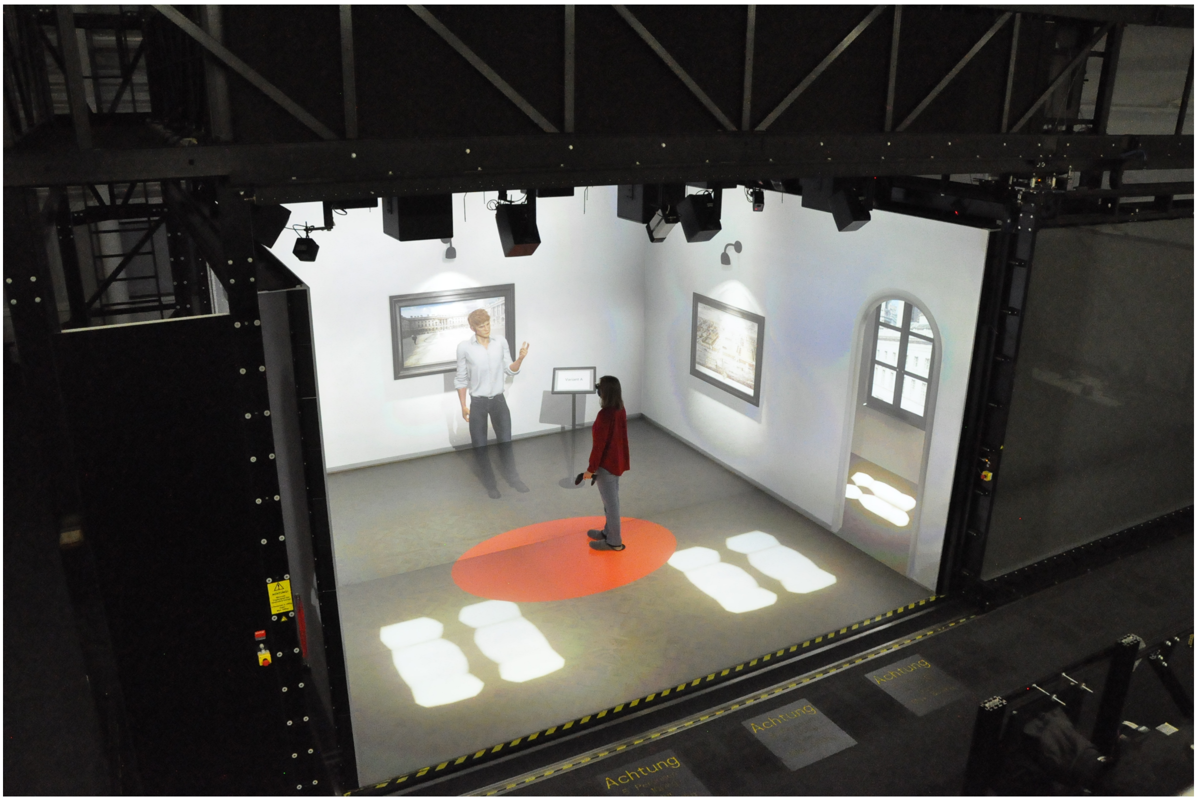


Figure A.4.: The museum scene shown in the AixCAVE with a user. For better visibility the door was kept open for this image, while it was fully closed during the study.

A.4.2. Seminar Room

This scene is based on a replication of the seminar room at the Institute for Hearing Technology and Acoustics (IHTA) at RWTH Aachen University, which is publicly avaliable at [Llorca-Bofí and Vorländer, 2021]. This model provides a very detailed replication of the interior of the scene but the surrounding scene, visible through the windows is of rather low quality. Therefore, we removed the exterior from this model and replaced it with the actual park scene, which is also available at [Llorca-Bofí et al., 2022]. To make it run efficiently in the Unreal Engine, so it can be used in VR, several performance optimizations were performed, which, however, should not have had an influence on the visual quality of the model. Also minor adjustments were made to the model as needed for our study, like adding air vents in the ceiling and a lectern at the front.

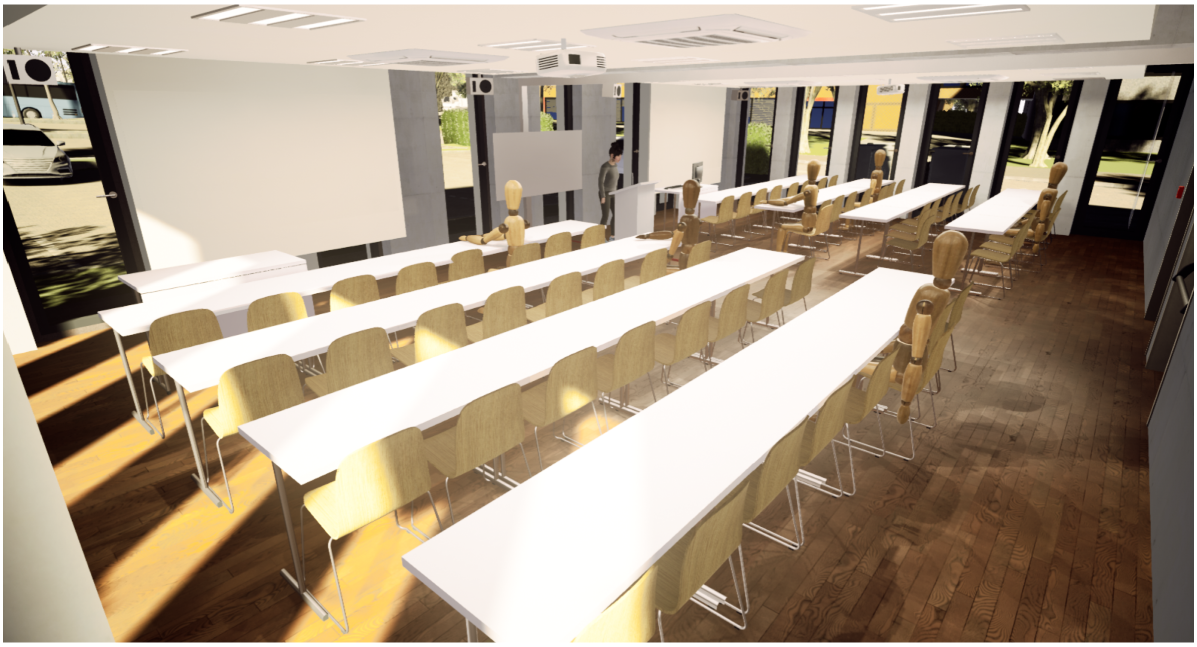


Figure A.5.: A view into the virtual representation of the seminar room at the Institute for Hearing Technology and Acoustics. The room is filled with some wooden mannequins, as used in the study presented in Sec. 6.2.

A.4.3. Living Room Scene

The living room scene used for the studies described in Sec. 5.5, Sec. 6.1, and Sec. 5.3.4 is based on a freely-available scene from the Epic Marketplace: `https://www.unrealengine.com/marketplace/en-US/product/archvis-interior-rendering`. We removed some objects that were catching too much attention and added an night-time outdoor skybox (see Fig.A.6) and adapted the lighting accordingly to improve the rendering of the included MetaHumans.

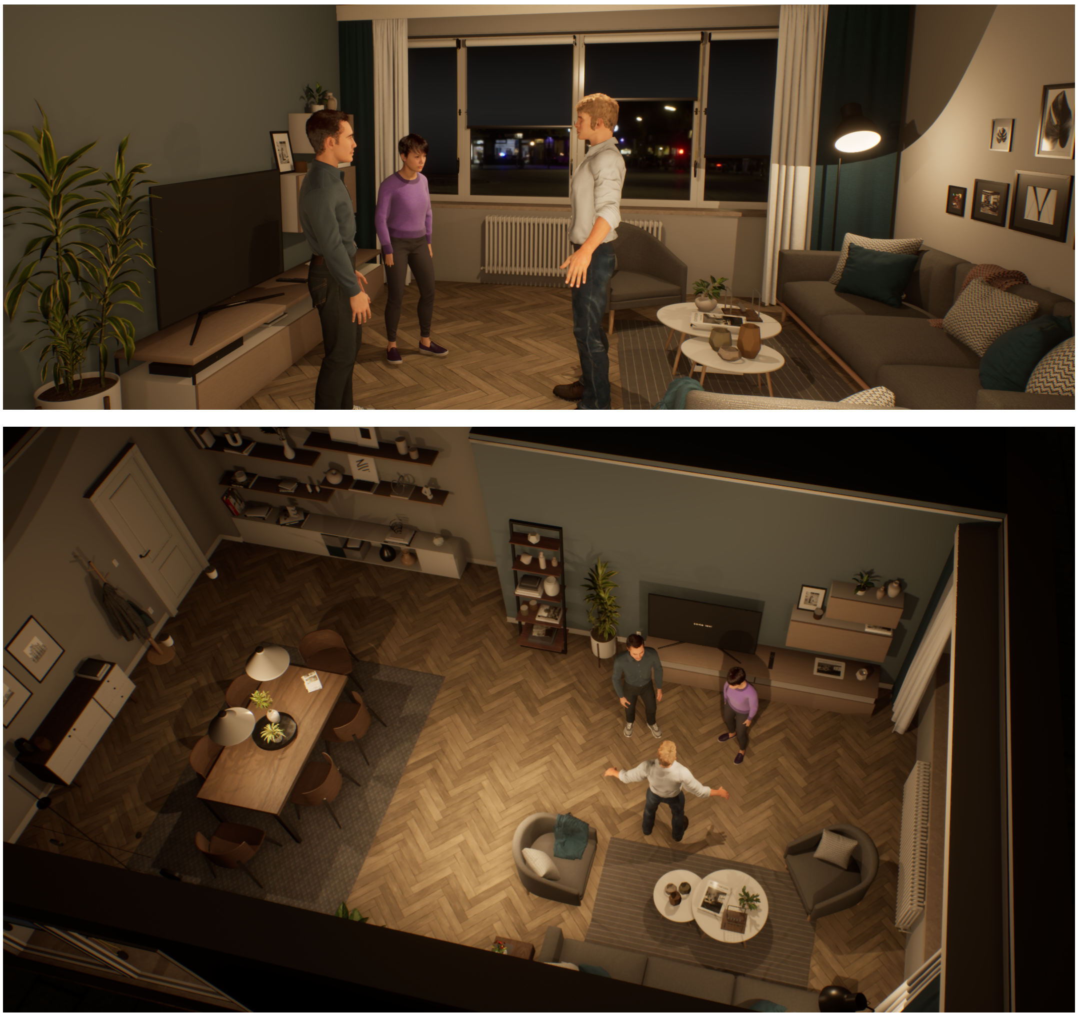


Figure A.6.: Top: Side view of the living room scene, showing two ECAs as in the experiment described in Sec. 6.1 and an exemplary visualization for the user by means of an avatar (bright shirt), illustrated in an A-pose here. The TV behind the ECAs was used to naturally display text during the studies and the outdoor skybox is visible through the windows. Bottom: A top view with removed ceiling and ceiling lamps shows the full extend of the scene.

A.5. Systems

In this appendix, we introduce the AixCAVE which was used for several of the studies presented in this work, as well as computer systems used (if the information was not directly given).

A.5.1. AixCAVE

The AixCAVE is a five-sided CAVE [Cruz-Neira et al., 1992] with a size of $5.25\,\mathrm{m} \times 5.25\,\mathrm{m} \times 3.30\,\mathrm{m}$ ($w \times d \times h$) [Kuhlen and Hentschel, 2014]. In the studies presented in this thesis participant were able to walk naturally though the virtual scenes since the walkable area of the scene matched the dimensions of the AixCAVE (see Fig. A.4). Users wore tracked active stereo glasses and interacted with an ART Flystick 2. The interaction device is tracked with six degrees of freedom (6DOF) and provides six buttons and an analog stick. The ceiling is equipped with an acoustic system consisting of 12 studio loudspeakers and 9 sub-woofers. It can be used to generate two separate virtual sound sources next to the ears of the user using dynamic cross-talk cancellation [Schröder et al., 2010; Masiero and Vorländer, 2014] to generate binaural audio. Furthermore two surveillance cameras are mounted at the ceiling with which the operator can monitor the users even when the door is closed. Additionally, microphones mounted in the ceiling enable the operator to listen to what users say in the AixCAVE and a microphone installed at the experimenter desk allows to talk to participants in case of emergency.

The renderings for the AixCAVE are computed by a 24-node graphics cluster. Each node contains two NVIDIA Quadro P6000, 192GB working memory, and two Intel Xeon Skylake CPUs (10 cores each).

A.5.2. Computer Hardware

The HMD experiments described in Sec. 5.5 and Sec. 5.3.4 were executed using a desktop computer with an Intel Core i9-10900X processor, 32GB working memory, and an NVIDIA RTX 3080 Ti graphics card.

The HMD experiments described in Sec. 6.1 was executed on a desktop computer with an Intel Core i9-13900KF processor, 64GB working memory, and an NVIDIA RTX 4090 graphics card.

Appendix B

Social Presence Questionnaires

This appendix should give an overview about the most-commonly used questionnaires to measure social presence. Questionnaires are sorted by publication year. This is also meant a reference for those questionnaires of these used throughout this work. A systematic review about social presence was written by Oh et al. [2018] and a further analysis was done by Fitrianie et al. [2019], of which the full results can be found at `https://osf.io/b7hyx`. This list represents the questionnaires we found to be most used, however, it does by no means represent a conclusive lists of questionnaires. Questionnaires explicitly not integrated for not providing much added value here or being focused on a very special interaction, like sharing photos together, were, for example, [Greef and Ijsselsteijn, 2004; Guadagno et al., 2007; Qiu and Benbasat, 2008; Poeschl and Doering, 2013, 2015; Li et al., 2019a; Wolfert et al., 2024]. For further reading with regard to social presence for computer-mediated communication between two humans, which is not a key focus of this work, we direct the interested reader to [Kreijns et al., 2022].

B.1. Semantic Differences Survey (SDS)

A very early instrument measuring social presences was developed by [Short et al., 1976]. It had 10 semantic differentials (other sources state it were 4 [Kreijns et al., 2022], 5 or even 24) which had to be rated on 7-point bipolar scale. Apparently not all items are listed in the book, here we present the set as found by Immohr et al. [2022].

impersonal	personal
unsociable	sociable
insensitive	sensitive
cold	warm
colorless	colorful
passive	active
closed	open
ugly	beautiful
small	large

B.2. Social Presence Survey (SPS)

The Social Presence Survey (SPS) was developed by Bailenson et al. [2001]. It is to rated on a 7-point Likert scale between -3 and 3. The actual anchor labels are not given in [Bailenson et al., 2001].

1.	I perceive that I am in the presence of another person in the room with me.	
2.	I feel that the person is watching me and is aware of my presence.	
3.	The thought that the person is not a real person crosses my mind often.	inv
4.	The person appears to be sentient (conscious and alive) to me.	
5.	I perceive the person as being only a computerized image, not as a real person.	inv

We used this, for example, in [Wendt et al., 2018; Bönsch et al., 2018a; Wendt et al., 2019; Ehret et al., 2020].

B.3. Networked Minds Measure of Social Presence

This is the original version of the Networked Minds Measure of Social Presence as conceived in [Biocca, 2001]. The items are rated on 7-point Likert scales. Labels and numbers for the anchors are not explicitly specified in [Biocca, 2001].

	Co-presence - Isolation/Aloneness
1.	I often felt as if I was all alone
2.	I think the other individual often felt alone.

	Co-presence - Mutual Awareness
3.	I hardly noticed another individual.
4.	The other individual didn't notice me in the room.
5.	I was often aware of others in the environment.
6.	Others were often aware of me in the room.
7.	I think the other individual often felt alone.
8.	I often felt as if I was all alone.

	Co-presence - Attentional Allocation
9.	I sometimes pretended to pay attention to the other individual.
10.	The other individual sometimes pretended to pay attention to me.
11.	The other individual paid close attention to me.
12.	I paid close attention to the other individual.
13.	My partner was easily distracted when other things were going on around us.
14.	I was easily distracted when other things were going on around me.
15.	The other individual tended to ignore me.
16.	I tended to ignore the other individual.

	Psychological Involvement - Empathy
17.	When I was happy, the other was happy.
18.	When the other was happy, I was happy.
19.	The other individual was influenced by my moods.
20.	I was influenced by my partner's moods.
21.	The other's mood did NOT affect my mood/emotional-state.
22.	My mood did NOT affect the other's mood/emotional-state.

	Psychological Involvement - Mutual Understanding
23.	My opinions were clear to the other.
24.	The opinions of the other were clear.
25.	My thoughts were clear to my partner.
26.	The other individual's thoughts were clear to me.
27.	The other understood what I meant.
28.	I understood what the other meant.

	Behavioral Engagement - Behavioral Interdependence
29.	My actions were dependent on the other's actions.
30.	The other's actions were dependent on my actions.
31.	My behavior was in direct response to the other's behavior.
32.	The behavior of the other was I direct response to my behavior.
33.	What the other did affected what I did.
34.	What I did affected what the other did.

	Behavioral Engagement - Mutual Assistance
35.	My partner did not help me very much.
36.	I did not help the other very much.
37.	My partner worked with me to complete the task.
38.	I worked with the other individual to complete the task.
	Behavioral Engagement - Dependent Action
39.	The other could not act without me.
40.	I could not act with the other.

B.3.1. Revised Networked Minds Social Presence Measure

The Revised Networked Minds Social Presence Measure was later presented in [Harms and Biocca, 2004]. Again 7-point Likert scales are used without explicitly stating the anchors.

	Co-presence
1.	I noticed (my partner).
2.	(My partner) noticed me.
3.	(My partner's) presence was obvious to me.
4.	My presence was obvious to (my partner).
5.	(My partner) caught my attention.
6.	I caught (my partner's) attention.

	Attentional Allocation
7.	I was easily distracted from (my partner) when other things were going on.
8.	(My partner) was easily distracted from me when other things were going on.
9.	I remained focused on (my partner) throughout our interaction.
10.	(My partner) remained focused on me throughout our interaction.
11.	(My partner) did not receive my full attention.
12.	I did not receive (my partner's) full attention.

	Perceived Message Understanding
13.	My thoughts were clear to (my partner).
14.	(My partner's) thoughts were clear to me.
15.	It was easy to understand (my partner).
16.	(My partner) found it easy to understand me.
17.	Understanding (my partner) was difficult.
18.	(My partner) had difficulty understanding me.

	Perceived Affective Understanding
19.	I could tell how (my partner) felt.
20.	(My partner) could tell how I felt.
21.	(My partner's) emotions were not clear to me.
22.	My emotions were not clear to (my partner).
23.	I could describe (my partner's) feelings accurately.
24.	(My partner) could describe my feelings accurately.

	Perceived Emotional Interdependence
25.	I was sometimes influenced by (my partner's) moods.
26.	(My partner) was sometimes influenced by my moods.
27.	(My partner's) feelings influenced the mood of our interaction.
28.	My feelings influenced the mood of our interaction.
29.	(My partner's) attitudes influenced how I felt.
30.	My attitudes influenced how (my partner) felt.

	Perceived Behavioral Interdependence
31.	My behavior was often in direct response to (my partner's) behavior.
32.	The behavior of (my partner) was often in direct response to my behavior.
33.	I reciprocated (my partner's) actions.
34.	(My partner) reciprocated my actions.
35.	(My partner's) behavior was closely tied to my behavior.
36.	My behavior was closely tied to (my partner's) behavior.

B.4. Questionnaire by Nowak and Biocca

Another often used questionnaire is the not explicitly named one presented in [Nowak and Biocca, 2003]. For *Self-reported copresence* and *Perceived other's copresence* 5-point Likert scales are used between 1 ("strongly agree") and 5 ("strongly disagree"). For the *Telepresence scale* a 7-point Likert scale between 1 ("Not at All") and 7 ("Very Much") is used, while the *Social presence* items are rated on a sliding scale.

	Self-reported copresence:
1.	I did not want a deeper relationship with my interaction partner.
2.	I wanted to maintain a sense of distance between us.
3.	I was unwilling to share personal information with my interaction partner.
4.	I wanted to make the conversation more intimate.
5.	I tried to create a sense of closeness between us.
6.	I was interested in talking to my interaction partner.

	Perceived other's copresence:
8.	My interaction partner was intensely involved in our interaction.
9.	My interaction partner seemed to find our interaction stimulating.
10.	My interaction partner communicated coldness rather than warmth.
11.	My interaction partner created a sense of distance between us.
12.	My interaction partner seemed detached during our interaction.
13.	My interaction partner was unwilling to share personal information with me.
14.	My interaction partner made our conversation seem intimate.
15.	My interaction partner created a sense of distance between us.
16.	My interaction partner created a sense of closeness between us.
17.	My interaction partner acted bored by our conversation.
18.	My interaction partner was interested in talking to me.
19.	My interaction partner showed enthusiasm while talking to me.

	Telepresence scale:
20.	How involving was the experience?
21.	How intense was the experience?
22.	To what extent did you feel like you were inside the environment you saw/heard?
23.	To what extent did you feel immersed in the environment you saw/heard?
24.	To what extent did you feel surrounded by the environment you saw/heard?

	Social presence:
25.	To what extent did you feel able to assess your partner's reactions to what you said?
26.	To what extent was this like a face-to-face meeting?
28.	To what extent was this like you were in the same room with your partner?
29.	To what extent did your partner seem "real"?
30.	How likely is it that you would choose to use this system of interaction for a meeting in which you wanted to persuade others of something?
31.	To what extent did you feel you could get to know someone that you met only through this system?

B.5. Questionnaire by Bailenson, Swinth et al.

This questionnaire was proposed and used in [Bailenson et al., 2005]. It was rated on a 7-point Likert scale from 0 (“strongly disagree”) to 6 (“strongly agre”). “inv” denotes reverse-coded items.

	Copresence:	
1.	Even when the “other” was present, I still felt alone in the virtual room.	inv
2.	I felt like there was someone else in the room with me.	
3.	I felt like the “other” was aware of my presence in the room.	

	Embarrassment:	
1.	I would be willing to change clothes in front of the “other”.	
2.	I would be willing to pick my nose in front of the “other”.	
3.	I would be willing to act out a scene from the movie “Titanic” in front of the “other”.	

	Likability:	
1.	I like the “other”.	
2.	I would like to meet this “other” again.	
3.	The “other” is attractive.	
4.	Spending time with the “other” was NOT satisfying.	inv

B.6. Temple Presence Inventory

The Temple Presence Inventory [Lombard et al., 2009] goes beyond only measuring social presence, but has specific sub-dimensions dedicated to it. For the sake of completeness we reproduce the questionnaire here in its entirety. Items are rated on 7-point Likert scales with different labels, which can be found at `http://matthewlombard.com/research/p2_ab.html`.

	TPI - Spatial presence:
1.	How much did it seem as if the objects and people you saw/heard had come to the place you were?
2.	How much did it seem as if you could reach out and touch the objects or people you saw/heard?
3.	How often when an object seemed to be headed toward you did you want to move to get out of its way?
4.	To what extent did you experience a sense of being there inside the environment you saw/heard?
5.	To what extent did it seem that sounds came from specific different locations?
6.	How often did you want to or try to touch something you saw/heard?
7.	Did the experience seem more like looking at the events/people on a movie screen or more like looking at the events/people through a window?

	TPI - Social presence - Actor within medium (parasocial interaction):
8.	How often did you have the sensation that people you saw/heard could also see/hear you?
9.	To what extent did you feel you could interact with the person or people you saw/heard?
10.	How much did it seem as if you and the people you saw/heard both left the places where you were and went to a new place?
11.	How much did it seem as if you and the people you saw/heard were together in the same place?
12.	How often did it feel as if someone you saw/heard in the environment was talking directly to you?
13.	How often did you want to or did you make eye-contact with someone you saw/heard?
14.	Seeing and hearing a person through a medium constitutes an interaction with him or her. How much control over the interaction with the person or people you saw/heard did you feel you had?

	TPI - Social presence - Passive interpersonal:
15.	During the media experience how well were you able to observe the facial expressions of the people you saw/heard?
16.	During the media experience how well were you able to observe the changes in tone of voice of the people you saw/heard?
17.	During the media experience how well were you able to observe the style of dress of the people you saw/heard?
18.	During the media experience how well were you able to observe the body language of the people you saw/heard?

	TPI - Social presence - Active interpersonal:
19.	How often did you make a sound out loud (e.g. laugh or speak) in response to someone you saw/heard in the media environment?
20.	How often did you smile in response to someone you saw/heard in the media environment?
21.	How often did you want to or did you speak to a person you saw/heard in the media environment?

	TPI - Engagement (mental immersion):
22.	To what extent did you feel mentally immersed in the experience?
23.	How involving was the experience?
24.	How completely were your senses engaged?
25.	To what extent did you experience a sensation of reality?
26.	How relaxing or exciting was the experience?
27.	How engaging was the story?

	TPI - Social richness:
	Please circle the number that best describes your evaluation of the media experience:
28.	Remote - Immediate (7 points)
29.	Unemotional - Emotional (7 points)
30.	Unresponsive - Responsive (7 points)
31.	Dead - Lively (7 points)
32.	Impersonal - Personal (7 points)
33.	Insensitive - Sensitive (7 points)
34.	Unsociable - Sociable (7 points)

	TPI - Social realism:
35.	The events I saw/heard would occur in the real world
36.	The events I saw/heard could occur in the real world
37.	The way in which the events I saw/heard occurred is a lot like the way they occur in the real world

	TPI - Perceptual realism:
38.	Overall how much did touching the things and people in the environment you saw/heard feel like it would if you had experienced them directly?
39.	How much did the heat or coolness (temperature) of the environment you saw/heard feel like it would if you had experienced it directly?
40.	Overall, how much did the things and people in the environment you saw/heard smell like they would had you experienced them directly?
41.	Overall, how much did the things and people in the environment you saw/heard look they would if you had experience them directly?
42.	Overall, how much did the things and people in the environment you saw/heard sound like they would if you had experienced them directly?

B.7. Godspeed Questionnaire

This questionnaire was developed to rate robots in [Bartneck et al., 2009]. Items are rated on a 5-point bipolar scale.

Please rate your impression of [the agent] on these scales:	
Godspeed I: Anthropomorphism	
Fake	Natural
Machinelike	Humanlike
Unconscious	Conscious
Artificial	Lifelike
Moving rigidly	Moving elegantly
Godspeed II: Animacy	
Dead	Alive
Stagnant	Lively
Mechanical	Organic
Artificial	Lifelike
Inert	Interactive
Apathetic	Responsive
Godspeed III: Likeability	
Dislike	Like
Unfriendly	Friendly
Unkind	Kind
Unpleasant	Pleasant
Awful	Nice
Godspeed IV: Perceived Intelligence	
Incompetent	Competent
Ignorant	Knowledgeable
Irresponsible	Responsible
Unintelligent	Intelligent
Foolish	Sensible
Godspeed V: Perceived Safety	
Anxious	Relaxed
Calm	Agitated
Quiescent	Surprised

B.8. Multimodal Presence Scale for Virtual Reality Environments (MPS)

This questionnaire was proposed and validated in [Makransky et al., 2017]. Items are rated on a 5-point Likert scale between “completely disagree” (1) , “disagree” (2), “neither disagree nor agree” (3), “agree” (4), and “strongly agree” (5). It also has a German translation available, published in [Volkmann et al., 2018].

	Physical Presence
PHYS2	The virtual environment seemed real to me.
PHYS3	I had a sense of acting in the virtual environment, rather than operating something from outside.
PHYS4	My experience in the virtual environment seemed consistent with my experiences in the real world.
PHYS5	While I was in the virtual environment, I had a sense of “being there”.
PHYS10	I was completely captivated by the virtual world.

	Social Presence
SOC1	I felt like I was in the presence of another person in the virtual environment.
SOC2	I felt that the people in the virtual environment were aware of my presence.
SOC3	The people in the virtual environment appeared to be sentient (conscious and alive) to me.
SOC5	During the simulation there were times where the computer interface seemed to disappear, and I felt like I was working directly with another person.
SOC7	I had a sense that I was interacting with other people in the virtual environment, rather than a computer simulation.

	Self-presence
SELF2	I felt like my virtual embodiment was an extension of my real body within the virtual environment.
SELF3	When something happened to my virtual embodiment, it felt like it was happening to my real body.
SELF4	I felt like my real arm was projected into the virtual environment through my virtual embodiment.
SELF6	I felt like my real hand was inside of the virtual environment.
SELF7	During the simulation, I felt like my virtual embodiment and my real body became one and the same.

B.9. Robotic Social Attributes Scale (RoSAS)

The questionnaire was developed and evaluated in [Carpinella et al., 2017] as an improvement to the Godspeed questionnaire to rate robots.

It has 3 factors with in total 18 items, which are rated using a 9-point Likert scale from “definitely not associated” (1) to “definitely associated” (9).

	Warmth
1.	Happy
2.	Feeling
3.	Social
4.	Organic
5.	Compassionate
6.	Emotional
	Competence
7.	Capable
8.	Responsive
9.	Interactive
10.	Reliable
11.	Competent
12.	Knowledgable
	Discomfort
13.	Scary
14.	Strange
15.	Awkward
16.	Dangerous
17.	Awful
18.	Aggressive

B.10. Copresence With Virtual Humans in Mixed Reality

This questionnaire was developed in [Pimentel and Vinkers, 2021] specifically to rate virtual humans in mixed reality, here more specifically augemented reality (AR). The 15 items are rated on a 5-point Likert scale from 1 (“completely disagree”) to 5 (“completely agree”).

	TPI - Spatial presence:
1.	I felt that I was in the same space as the other person.
2.	It felt like the other person was with me.
3.	I felt that the other person and I were together in the same space.
4.	I felt that the other person and I were sharing the same physical space.
5.	I felt that I was in the presence of the other person.
6.	I felt that the other person paid attention to me.
7.	I felt that the other person responded to my nonverbal expressions (e.g., gestures, facial expressions).
8.	I felt that the other person responded to shifts in my movements (e.g., posture, position).
9.	The other person responded to my actions.
10.	I felt that the other person was attentive to what I was doing.
11.	I think that the other person noticed what I was paying attention to.
12.	The other person did not acknowledge my presence.
13.	The other person did not react to my behavior.
14.	I felt that the other person was distracted.
15.	I felt that the other person did not give their attention to me.

B.11. ASAQ

Excerpt from the Artificial Social Agent (ASA) Questionnaire [Fitrianie et al., 2022], rated on a 7-point Likert scale. The full questionnaire can be found at `https://doi.org/10.4121/19650846`. We also developed and published a German version of the full questionnaire [Albers et al., 2024]. Here we only show the most relevant constructs to this work. “inv” marks questions of which the score is inverted before computing the mean score of the construct.

	1.1 Human-Like Appearance	
HLA1	[The agent]’s appearance is human	
HLA2	[The agent] has the appearance of a human	
HLA3	[The agent] has a human-like outside	
HLA4	[The agent]’s appearance makes me think of a human	

	1.2 Human-Like Behavior	
HLB1	A human would behave like [the agent]	
HLB2	[The agent]’s manners is consistent with that of people	
HLB3	[The agent] behavior makes me think of human behavior	
HLB4	[The agent] behaves like a real person	
HLB5	[The agent] has a human-like manner	

	1.3 Natural Appearance	
NA1	[The agent] appears like something that could exist in nature	
NA2	[The agent] has a natural physique	
NA3	[The agent]’s resemblance has an organic origin	
NA4	[The agent] seems natural from its outward appearance	
NA5	How [the agent] is represented is realistic	

	1.4 Natural Behavior	
NB1	[The agent] is alive	
NB2	[The agent] acts naturally	
NB3	[The agent] reacts like a living organism	

	4 Agent’s Likeability	
AL1	[The agent]’s appearance is pleasing	
AL2	I like [the agent]	
AL3	I dislike [the agent]	inv
AL4	[The agent] is cooperative	
AL5	I want to hang out with [the agent]	

	5 Agent’s Sociability	
AS1	[The agent] can easily mix socially	
AS2	It is easy to mingle with [the agent]	
AS3	[The agent] interacts socially with [me / the user]	

	13 Agent’s Coherence	
AC1	[The agent]’s behavior does not make sense	inv
AC2	[The agent]’s behavior is irrational	inv
AC3	[The agent] is inconsistent	inv
AC4	[The agent] appears confused	inv

	16 Social Presence.	
SP1	[The agent] has a social presence	
SP2	[The agent] is a social entity	
SP3	[I have / The user has] the same social presence as [the agent]	

	19 User-Agent Interplay	
UAI1	[My / The user’s] emotions influence the mood of the interaction	
UAI2	[The agent] reciprocates [my / the user’s] actions	
UAI3	[The agent]’s and [my / the user’s] behaviors are in direct response to each other’s behavior	
UAI4	[The agent]’s and [my / the user’s] emotions change to what [we / they] do to each other	

B.12. Virtual Human Plausibility Questionnaire (VHPQ)

This questionnaire was developed in [Mal et al., 2022] to rate virtual humans and to operationalize the virtual human plausibility concept presented in [Latoschik and Wienrich, 2022]. Items are assessed on a 7-point Likert scale from 1 “does not apply at all”) to 7 (“completely applies”).

	Appearance and Behavior Plausibility (ABP)
ABP1	The behavior of the virtual character seemed to be plausible to me.
ABP2	The appearance of the virtual character seemed to be plausible to me.
ABP3	The virtual character’s behavior matched its appearance.
ABP4	The behavior and appearance of the virtual character were coherent.
ABP5	The virtual character behaved as I would expect it to behave.
ABP6	I could predict by the virtual character’s appearance how it would behave.
ABP7	The virtual character behaved in the virtual environment as I would expect it to.

	Match to the Virtual Environment (MVE)
MVE1	The virtual character fit into the virtual environment.
MVE2	The virtual character was a plausible part of the virtual environment.
MVE3	The appearance of the virtual character and the virtual environment matched.
MVE4	The behavior of the virtual character matched with the virtual environment.

B.13. Virtual Agent Believability Scale

This scale was developed and evaluated in [Guo et al., 2023]. The item are rated on 7-point Likert scale with the labels: “strongly disagree” (1), “disagree” (2), “somewhat disagree” (3), “neither agree nor disagree” (4), “somewhat agree” (5), “agree” (6), and “strongly agree” (7).

	Visual properties (or appearance):
1.	The visual design of the virtual agent caught my attention.
2.	I think the virtual agent’s appearance is aesthetically pleasing.
3.	I think the virtual agent’s visual design is realistic.
	Behavior:
4.	The virtual agent’s behavior drew my attention.
5.	I felt the virtual agent’s behavior was coherent and natural.
6.	I think the virtual agent’s behavior was easy to understand.
7.	I felt the virtual agent’s behavior was appropriate to the context.
8.	I felt sometimes the virtual agent behaved inappropriately.
	Awareness:
9.	I felt that the virtual agent perceived the environment around him/her/them.
10.	I felt that the virtual agent reacted to the change in the environment.
11.	I felt that the virtual agent was aware of my presence.
12.	I felt that the virtual agent was aware of the presence of other virtual agents.
13.	The virtual agent was unaware of its surroundings.
	Social relationships:
14.	The virtual agent interacted socially with me.
15.	I felt that the virtual agent was able to coordinate with me.
16.	The virtual agent interacted socially with the other virtual agent(s).
17.	I felt that the virtual agent was able to coordinate with the other virtual agent(s).
	Intelligence:
18.	I felt that the virtual agent was able to make plans.
19.	I felt that the virtual agent learned from past experiences.
20.	I felt that the virtual agent seemed to have memory.
	Emotion:
21.	I felt that the virtual agent was capable of having feelings.
22.	I felt that the virtual agent expressed emotions.
23.	I felt that the virtual agent’s expressed emotions were easy to understand.
24.	I felt that the virtual agent’s expressed emotions were appropriate to the context.

	Personality:
25.	I felt that the virtual agent had a personality.
26.	I felt that the virtual agent was extraverted and enthusiastic.
27.	I felt that the virtual agent was sympathetic and warm.
28.	I felt that the virtual agent was dependable and self-disciplined.
29.	I felt that the virtual agent was emotionally stable.
30.	I felt that the virtual agent was open to new experiences.

	Personality:
31.	I felt that the virtual agent seemed to have self-awareness.
32.	The virtual agent took actions without inputs from others.
33.	The virtual agent seemed to have its own goals.

	Overall believability:
34.	I felt that the virtual agent was believable.
35.	I felt that the virtual agent behaved like a real person.
36.	I enjoy the interaction with the virtual agent.

B.14. igroup presence questionnaire (IPQ)

The igroup presence questionnaire (IPQ) [Schubert et al., 2001] measures presence and not social presence. It is presented here as well, as it was used in one of our studies, presented in Sec. 6.2. It is also available in German at: `http://www.igroup.org/pq/ipq/index.php`.

The questionnaire contains one general items and three subscales with individual items which are rated on different 7-point Likert scales (see `http://www.igroup.org/pq/ipq/index.php`).

	General:
G1	In the computer generated world I had a sense of ”being there”
	Spatial Presence:
SP1	Somehow I felt that the virtual world surrounded me.
SP2	I felt like I was just perceiving pictures.
SP3	I did not feel present in the virtual space.
SP4	I had a sense of acting in the virtual space, rather than operating something from outside.
SP5	I felt present in the virtual space.
	Involvement:
INV1	How aware were you of the real world surrounding while navigating in the virtual world? (i.e. sounds, room temperature, other people, etc.)?
INV2	I was not aware of my real environment.
INV3	I still paid attention to the real environment.
INV4	I was completely captivated by the virtual world.
	Experiences Realism:
REAL1	How real did the virtual world seem to you?
REAL2	How much did your experience in the virtual environment seem consistent with your real world experience?
REAL3	How real did the virtual world seem to you?
REAL4	The virtual world seemed more realistic than the real world.

Appendix C

Publications: Content and Contributions

In the following chapter, we provide an overview of the publications the author of this thesis (Jonathan Ehret, né Wendt) was involved with. The contribution of each of the authors is stated for all publications that were reproduced in this thesis. All of the presented publications underwent a peer-review process, non-peer-reviewed submission, like conference talks, are omitted. This chapter is split in three sections, first all scientific papers are presented that are reproduced in this thesis (App. C.1), followed by other peer-reviewed publications (App. C.2), concluded by a list of published datasets and software in App. C.3. All publications within a section are presented in the order of publication. Content and contribution descriptions are, where applicable re-used or adapted from [Bönsch, 2024] for consistency.

C.1. Publications Reproduced in this Thesis

[Wendt et al., 2018]

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

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

Type Workshop Paper

Venue IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE) @ IEEE Virtual Reality Conference

Content This study investigates how sound directionality from an embodied conversational agent (ECA) affects a user's perception of the social presence perceived for an ECA in a VR environment. A pre-study with 8 participants found that while some participants noticed a difference in sound between the two auralization conditions (omnidirectional and directional), there were issues with the study design that need to be addressed before a larger scale study can be conducted delving into the impact of an ECA's speech auralization.

Contribution The study was designed, implemented, performed, and analyzed by Jonathan Ehret, né Wendt. Jonas Stienen assisted in the implementation of the acoustic rendering while Benjamin Weyers, Tom Vierjahn, and Andrea Bönsch provided valuable feedback and ideas on the study design. Benjamin Weyers and Jonathan Ehret performed the data analysis of the study. The written paper originates from Jonathan Ehret, while Michael Vorländer, Torsten W. Kuhlen, and all other co-authors provided guidance and feedback in writing the publication.

[Wendt et al., 2019]

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

Authors **Jonathan Wendt**, Benjamin Weyers, Jonas Stienen, Andrea Bönsch, Michael Vorländer, and Torsten W. Kuhlen

Type Conference Paper

Venue ACM International Conference on Intelligent Virtual Agents

Content The paper investigates the impact of adding directivity to the speech sound source of ECAs in VR environments and finds no significant effects on perceived social presence and realism of the ECA's voice, suggesting that other factors, such as overall realism and social context, may play a more prominent role in user perception.

Contribution The study was designed, implemented, performed, and analyzed by Jonathan Ehret, né Wendt. Jonas Stienen assisted in the implementation of the acoustic rendering. Benjamin Weyers, Andrea Bönsch, and Jonas Stienen provided valuable feedback on the study design, and Benjamin Weyers assisted with valuable input during the study analysis. The written paper originates from Jonathan Ehret, while Michael Vorländer, Torsten W. Kuhlen, and all other co-authors provided guidance in writing the publication.

[Ehret et al., 2020]

Title Evaluating the Influence of Phoneme-Dependent Dynamic Speaker Directivity of Embodied Conversational Agents' Speech

Authors **Jonathan Ehret**, Jonas Stienen, Chris Brozdowski, Andrea Bönsch, Irene Mittelberg, Michael Vorländer, and Torsten W. Kuhlen

Type Conference Paper

Venue ACM International Conference on Intelligent Virtual Agents

Content The paper explores the use of directional filters to enhance the realism of ECAs by simulating the directionality of speech sounds based on the currently uttered speech (coined dynamic directivity). In a VR-based study comparing different speech auralization methods, participants could not distinguish between static and dynamic directionality, and preferences aligned with naturalness ratings, but there was no consensus on the most natural method.

Contribution The study was implemented, performed, and analyzed by the lead author Jonathan Ehret. Jonas Stienen provided simulations of the directivity filters needed. The study design was developed by Jonathan Ehret, Chris Brozdowsk, Andrea Bönsch, and Jonas Stienen. Chris Brozdowski and Irene Mittelberg provided valuable insights into phonetics and developed the speech material, while also doing the motion capturing. Jonas Stienen was in charge of recording and post-processing the audio material. Chris Brozdowski provided valuable insights for the analysis of the gathered data. The written paper originates from Jonathan Ehret, while Andrea Bönsch, Irene Mittelberg, Michael Vorländer and Torsten W. Kuhlen provided guidance in writing the publication.

[Ehret et al., 2021]

Title Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents' Speech?

Authors **Jonathan Ehret**, Andrea Bönsch, Lukas Aspöck, Christine T. Röhr, Stefan Baumann, Martine Grice, Janina Fels, Torsten W. Kuhlen

Type Journal Paper

Journal ACM Transactions on Applied Perception

Content This paper investigates the impact of prosody, specifically accent placement, on the perceived naturalness and aliveness of ECAs' speech. The study compares inadequate prosody generated by text-to-speech engines, inadequate prosody imitated by trained human speakers, and adequate prosody produced by those speakers, with results showing that adequate prosody is an important factor for perceiving speech as natural. Moreover, the presence of ECA embodiments does not significantly affect the perception of naturalness in the ECAs' speech.

Contribution The study design was developed by Jonathan Ehret, Lukas Aspöck, Stefan Baumann, and Andrea Bönsch. The video material and acoustic renderings for the study were generated by Jonathan Ehret. Christine T. Röhr implemented the online study and analyzed the data together with Jonathan Ehret. Stefan Baumann and Martine Grice provided guidance on the linguistic aspects of this project. The written paper originates from Jonathan Ehret, while Andrea Bönsch, Christine T. Röhr, Stefan Baumann, Janina Fels, and Torsten W. Kuhlen provided guidance in writing the publication.

[Ehret et al., 2023]

Title Who's next? Integrating Non-Verbal Turn-Taking Cues for Embodied Conversational Agents

Authors **Jonathan Ehret**, Andrea Bönsch, Patrick Nossol, Cosima A. Ermert, Chinthusa Mohanathasan, Sabine J. Schlittmeier, Janina Fels, Torsten W. Kuhlen

Type Conference Paper

Venue ACM International Conference on Intelligent Virtual Agents

Content This paper studies how ECAs can better simulate turn-taking in conversations, which is an important aspect of natural human communication. The authors propose a system that generates non-verbal cues through gestures, gaze, and breathing to indicate when the ECA wants to take or yield the turn. The system was evaluated in a VR study where participants interacted with ECAs. The results show that the gesture manipulation had an effect on the interaction but there was no significant impact on the perceived social presence of the ECAs.

Contribution The study design was conceived and implemented by Jonathan Ehret, with support in improving the study by Andrea Bönsch, Cosima Ermert, and Chinthusa Mohanathasan. Patrick Nossol supported in recording and post-processing the used full-body animations. The study was conducted by Jonathan Ehret and Patrick Nossol. The data analysis was done by Jonathan Ehret with support of Chinthusa Mohanathasan. The written paper originates from Jonathan Ehret, while Sabine J. Schlittmeier, Janina Fels, Torsten W. Kuhlen and all other co-authors provided guidance in writing the publication.

[Ehret et al., 2024b]

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

Authors **Jonathan Ehret***, Andrea Bönsch*, Isabel S. Schiller, Carolin Breuer, Lukas Aspöck, Janina Fels, Sabine J. Schlittmeier, and Torsten W. Kuhlen (* These authors contributed equally to this work.)

Type Workshop Paper

Venue IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE) @ IEEE Virtual Reality Conference

Content This paper investigates the audiovisual coherence in VR environments by exploring the necessity of visually representing background sound sources. The study examines how different visual representations (animated vs static vs none) of background noise sources affect user experience. Findings suggest that animated visualization is subjectively preferred for optimal coherence, especially for sounds from virtual humans, but it doesn't influence objective performance in a listening task.

Contribution The study design was developed by Jonathan Ehret, Andrea Bönsch, and Isabel S. Schiller. Jonathan Ehret implemented the study and Carolin Breuer and Lukas Aspöck supported in the acoustical setup and by recording the acoustic stimuli. The data analysis was performed by Jonathan Ehret and Isabel S. Schiller. The written paper originates from Jonathan Ehret (primarily chapters 2 and 3) and Andrea Bönsch (primarily chapter 1 and 4) who contributed equally to the manuscript, while Sabine J. Schlittmeier, Janina Fels, Torsten W. Kuhlen and all other co-authors provided guidance in writing the publication.

[Ehret et al., 2024a]

Title StudyFramework: Comfortably Setting up and Conducting Factorial-Design Studies Using the Unreal Engine

Authors **Jonathan Ehret**, Andrea Bönsch, Janina Fels, Sabine J. Schlittmeier, Torsten W. Kuhlen

Type Workshop Paper

Venue Open Access Tools (OAT) and Libraries for Virtual Reality @ IEEE Conference on Virtual Reality

Content This paper introduces the plugin *StudyFramework*, which streamlines setting up and conducting factorial-design VR-based user studies within the Unreal Engine, saving time and reducing errors for researchers.

Contribution The StudyFramework was developed and continuously improved by Jonathan Ehret with support in the conception by Andrea Bönsch. The presented evaluation was conceived and conducted by Jonathan Ehret. The manuscript originated from Jonathan Ehret, while Sabine J. Schlittmeier, Janina Fels, Torsten W. Kuhlen, and Andrea Bönsch provided guidance in writing the publication.

[Ehret et al., 2025a]

Title Exploring Gaze Dynamics: Initial Findings on the Role of Listening Bystanders in Conversational Interactions

Authors **Jonathan Ehret**, Valentin Dasbach, Jan-Nikjas Hartmann, Janina Fels, Torsten W. Kuhlen, and Andrea Bönsch

Type Workshop Paper

Venue IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE) @ IEEE Virtual Reality Conference

Content This work-in-progress paper investigates how virtual listening bystanders influence participants' gaze behavior and their perception of turn-taking during scripted conversations with ECAs. Participants had to listen to two ECAs presenting family stories, while three additional agents were acting as listneing bystanders. Thereby they either adhered to a social model, performed random gazing behavior, or were not present at all. It was found that adding additional bystanders increased the time participants took to gaze at the next speaker after a turn change. Furthermore, random gazing bystanders were particularly noted for obscuring conversational flow.

Contribution The study design was developed by Jonathan Ehret and Valentin Dasbach. Jonathan Ehret implemented the study prototype which was extended by Valentin Dasbach to address the research objective. The data analysis was performed by Valentin Dasbach and Jonathan Ehret. The written paper originates from Andrea Bönsch and Jonathan Ehret. Jan-Nikjas Hartmann supported in generating the figures presented in the paper, while Janina Fels, Torsten W. Kuhlen and all other co-authors provided guidance in writing the publication.

[Ehret et al., 2025b]

Title Objectifying Social Presence: Evaluating Degraded Speech Performance in ECAs Using the Heard Text Recall Paradigm

Authors **Jonathan Ehret**, Jonas Schüppen, Chinthusa Mohanathasan, Cosima A. Ermert, Janina Fels, Sabine J. Schlittmeier, Torsten W. Kuhlen, Andrea Bönsch

Type Journal Paper

Journal IEEE Transactions on Visualization and Computer Graphics

Content This paper's contribution is threefold. First, different gesture manipulations are discussed and evaluated with their influence on perceived naturalness. Second, a study is conducted in which participants have to listen to family stories being told by two ECAs. The performance of the ECAs in deliberately degraded, e.g., by removing lip sync or using distorted co-verbal gestures. The influence of these degraders on perceived social presence is evaluated, finding significant deductions only for omitting lip sync and using artificial voices. Lastly, during listening to the stories cognitive spare capacity is measured by means of a dual-task paradigm. This metric is evaluated as promising objective proxy for perceived social presence.

Contribution The initial conceptualization of the two presented studies originated from Jonathan Ehret. Jonas Schüppen implemented gesture manipulations and conducted the gesture manipulation study as part of his master thesis, under close guidance by Jonathan Ehret throughout the entire development and study process. Andrea Bönsch, Cosima Ermert and Chinthusa Mohanathasan advised on methodology, and Torsten W. Kuhlen offered valuable conceptual feedback. The written paper originated from Jonathan Ehret, while Janina Fels, Sabine J. Schlittmeier, and all co-authors reviewed and edited the manuscript.

C.2. Other Publications

[Bönsch et al., 2016]

Title Collision Avoidance in the Presence of a Virtual Agent in Small-Scale Virtual Environments

Authors Andrea Bönsch, Benjamin Weyers, **Jonathan Wendt**, Sebastian Freitag, and Torsten W. Kuhlen

Type Technote

Venue IEEE Symposium on 3D User Interfaces

Content The paper discusses the need for collision avoidance strategies for virtual agents in small-scale IVEs in a CAVE-like environment. It presents the results of a user study conducted in a small office setting where participants preferred collaborative collision avoidance with VAs, expecting the VAs to step aside to create space for them while being willing to adjust their own paths.

Contribution Honorable Mention for Best Technote

[Bönsch et al., 2017c]

Title Peers At Work: Economic Real-Effort Experiments In The Presence of Virtual Co-Workers

Authors Andrea Bönsch, **Jonathan Wendt**, Heiko Overath, Özgür Gürerk, Christine Harbring, Christian Grund, Thomas Kittsteiner, and Torsten W. Kuhlen

Type Poster

Venue IEEE Virtual Reality Conference

Content This poster uses VR to conduct different economic experiments where people sort objects and investigates how people's performance is affected by seeing a virtual coworker, represented by a VA, sorting objects at the same time. The study finds that people work harder when their virtual coworker is also productive and that competition with a virtual coworker motivates people more than simply being paid for the objects they sort correctly.

[Bönsch et al., 2017a]

Title Score-Based Recommendation for Efficiently Selecting Individual Virtual Agents in Multi-Agent Systems

Authors Andrea Bönsch, Robert Trisnadi, **Jonathan Wendt**, Tom Vierjahn, and Torsten W. Kuhlen

Type Poster

Venue 23rd ACM Symposium on Virtual Reality Software and Technology (VRST)

Content The poster proposes a recommendation system to aid human operators who control the VAs' behavior, e.g., displaying certain gestures, in a populated VR environment. This system recommends VAs to the operator based on the user's distance and gaze direction toward the VAs, making the selection process faster and less error-prone. This can improve the overall user experience in VR environments with multiple virtual agents.

[Bönsch et al., 2018a]

Title Social VR: How Personal Space is Affected by Virtual Agents' Emotions

Authors Andrea Bönsch, Sina Radke, Heiko Overath, Laura M. Asche, **Jonathan Wendt**, Tom Vierjahn, Ute Habel, and Torsten W. Kuhlen

Type Conference Paper

Venue IEEE Conference on Virtual Reality and 3D User Interfaces

Content The paper presents findings from a controlled experiment involving German males aged 18 to 30 years, exploring personal space preferences in social interactions. It indicates that the size of PS varies based on the emotional facial expression of virtual agents (VAs) (angry vs. happy) and the number of VAs (single vs. group), with larger distances preferred when approached by angry VAs and when approached by a group of VAs.

[Bönsch et al., 2018b]

Title Towards Understanding the Influence of a Virtual Agent's Emotional Expression on Personal Space

Authors Andrea Bönsch, Sina Radke, **Jonathan Wendt**, Tom Vierjahn, Ute Habel, and Torsten W. Kuhlen

Type Workshop Paper

Venue IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE) @ IEEE Virtual Reality Conference

Content Building on the promising results from [Bönsch et al., 2018a], this paper proposes a new study design for further investigation on the impact of emotional expressions displayed by VAs on individuals' PS preferences.

[Bönsch et al., 2019]

Title Evaluation of Omnipresent Virtual Agents Embedded as Temporarily Required Assistants in Immersive Environments

Authors Andrea Bönsch, Jan Hoffmann, **Jonathen Wendt**, and Torsten W. Kuhlen

Type Workshop Paper

Venue IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE) @ IEEE Virtual Reality Conference

Content The paper evaluates the idle behavior of a virtual assistant. The virtual agent (VA), embedded as a temporary required assistant, is always present, while it either shows a following behavior, where it follows the user closely, or a busy behavior, where it stays nearby, keeping itself busy and approaches only when asked. A user study indicates that participants prefer the following behavior, which also leads to faster response times compared to the busy behavior.

[Bönsch et al., 2020d]

Title Towards a Graphical User Interface for Exploring and Fine-Tuning Crowd Simulations

Authors Andrea Bönsch, Marcel Jonda, **Jonathan Ehret**, and Torsten W. Kuhlen

Type Workshop Paper

Venue IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE) @ IEEE Virtual Reality Conference

Content This paper proposes a graphical user interface that allows users to evaluate different crowd simulation algorithms and adjust their parameters to achieve a more natural and believable crowd behavior.

[Bönsch et al., 2020a]

Title Immersive Sketching to Author Crowd Movements in Real-time

Authors Andrea Bönsch, Sebastian J. Barton, **Jonathan Ehret**, and Torsten W. Kuhlen

Type Conference Paper

Venue ACM International Conference on Intelligent Virtual Agents

Content The paper introduces a sketch-based interface for efficiently controlling the movement of virtual crowds in 3D virtual reality environments, presenting initial promising results of a proof-of-concept and discussing potential improvements and future directions.

Contribution Andrea Bönsch designed and analyzed the user study. Sebastian J. Barton implemented and conducted the study, as part of his bachelor thesis, under close guidance and advice throughout the entire development and study process by Andrea Bönsch. Jonathan Ehret provided valuable feedback on the study design and supported the analysis. Torsten W. Kuhlen provided guidance in writing the publication.

[Bönsch et al., 2020e]

Title The Impact of a Virtual Agent's Non-Verbal Emotional Expression on a User's Personal Space Preferences

Authors Andrea Bönsch, Sina Radke, **Jonathan Ehret**, Ute Habel, and Torsten W. Kuhlen

Type Conference Paper

Venue ACM International Conference on Intelligent Virtual Agents

Content Extending [Bönsch et al., 2018a] and [Bönsch et al., 2018b], this paper investigates PS preferences in VR interactions with VAs, finding that subjects preferred larger PS when approached by an angry VA (conveyed via mimics ,full-body animations, and footstep sound), replicating previous findings, and further observing differences in PS preferences across different immersive settings, namely an HMD and the aixCAVE.

[Bönsch et al., 2020b]

Title Inferring a User's Intent on Joining or Passing by Social Groups

Authors Andrea Bönsch, Alexander R. Bluhm, **Jonathan Ehret**, and Torsten W. Kuhlen

Type Conference Paper

Venue ACM International Conference on Intelligent Virtual Agents

Content he paper addresses the need for modeling interactions between users and social groups of VAs in VR applications, specifically focusing on differentiating between joining and passing by a group. To improve the interactive capabilities of VAs in such situations, the authors propose a classification scheme that infers user intent from social cues and triggers realistic non-verbal actions by the VAs. The results from the pilot study are promising.

[Bönsch et al., 2021a]

Title Indirect User Guidance by Pedestrians in Virtual Environments

Authors Andrea Bönsch, Katharina Güths, **Jonathan Ehret**, Torsten W. Kuhlen

Type Poster

Venue International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments

Content The poster explores unaided wayfinding strategies using virtual pedestrians as subtle social cues to guide users through unfamiliar IVEs, with a brief overview of required pedestrian behavior and the results of an initial feasibility study suggesting its potential effectiveness.

[Bönsch et al., 2021b]

Title Being Guided or Having Exploratory Freedom: User Preferences of a Virtual Agent's Behavior in a Museum

Authors Andrea Bönsch, David Hashem, **Jonathan Ehret**, Torsten W. Kuhlen

Type Conference Paper

Venue ACM International Conference on Intelligent Virtual Agents

Content The paper investigates whether users prefer a virtual guide or a free exploration accompanied by a VA in a virtual museum. The conducted user study indicates that combining both approaches may result in higher user acceptance.

Award GALA Audience Award for the submitted gala video showcasing the application.

[Bönsch et al., 2022]

Title An Embodied Conversational Agent Supporting Scene Exploration by Switching between Guiding and Accompanying

Authors Andrea Bönsch, Daniel Rupp, **Jonathan Ehret**, Torsten W. Kuhlen

Type Late-Breaking Report

Venue IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE) @ IEEE Virtual Reality Conference

Content This late-breaking report introduced the concept of a VA designed to serve as a virtual guide and a knowledgeable companion for interactive and structured exploration of unfamiliar IVEs, with a brief overview of its behavioral design and a mention of an upcoming user study. The idea is based on the results of [Bönsch et al., 2021b], while its implementation and study results are discussed in [Bönsch et al., 2024]

[Ehret et al., 2022]

Title Natural Turn-Taking with Embodied Conversational Agents

Authors **Jonathan Ehret**, Andrea Bönsch, Torsten W. Kuhlen

Type Late-Breaking Report

Venue IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE) @ IEEE Virtual Reality Conference

Content This late-breaking details plans on exploring whether turn-taking gestures of a VA improve how users take turns in a user–agent conversation. The idea resulted in [Ehret et al., 2023]

[Ehret, 2022]

Title Verbal Interactions with Embodied Conversational Agents

Authors **Jonathan Ehret**

Type Doctoral Consortium Abstract

Venue ACM International Conference on Intelligent Virtual Agents

Content An abstract describing Jonathan Ehret's research plan, specifically lining out the ideas that led to [Ehret et al., 2023] (evaluating different non-verbal turn-taking signals) and [Ehret et al., 2025b] (evaluating cognitive spare capacity as proxy for measuring social presence)

[Bönsch et al., 2023b]

Title	Whom Do You Follow? Pedestrian Flows Constraining the User's Navigation during Scene Exploration
Authors	Andrea Bönsch, Lukas B. Zimmermann, **Jonathan Ehret**, and Torsten W. Kuhlen
Type	Poster
Venue	ACM International Conference on Intelligent Virtual Agents
Content	The poster presents a work-in-progress combining two wayfinding techniques, the River Analogy and virtual pedestrian flows, to guide users through scenes by having them follow a chosen pedestrian as if it were a boat on a river, with initial study results on different visualization methods for this approach.

[Bönsch et al., 2023a]

Title	Where Do They Go? Overhearing Conversing Pedestrian Groups during Scene Exploration
Authors	Andrea Bönsch, Till Sittart, **Jonathan Ehret**, Torsten W. Kuhlen
Type	Poster
Venue	ACM International Conference on Intelligent Virtual Agents
Content	The poster presents a novel approach to scene exploration in IVEs, using conversing pedestrian groups as cues to indirectly guide users to unseen points of interest, with insights from a feasibility study comparing this approach to non-talkative groups and groups discussing random topics.

[Bönsch et al., 2024]

Title	Wayfinding in Immersive Virtual Environments as Social Activity Supported by Virtual Agents
Authors	Andrea Bönsch, **Jonathan Ehret**, Daniel Rupp, Torsten W. Kuhlen
Type	Journal Paper
Journal	Frontiers in Virtual Reality, Section Virtual Reality and Human Behaviour
Content	This paper examines the impact of VAs on user experience, comfort, and scene knowledge acquisition during wayfinding in IVEs, comparing unsupported wayfinding to strong and weak social wayfinding conditions through a between-subject study involving 60 participants. The findings highlight the efficiency of social wayfinding support, particularly the strong type, while also suggesting potential for further exploration of weak social wayfinding techniques.

[Albers et al., 2024]

Title German and Dutch Translations of the Artificial-Social-Agent Questionnaire Instrument for Evaluating Human-Agent Interactions

Authors Nele Albers*, Andrea Bönsch*, **Jonathan Ehret**, Boleslav A. Khodakov, and Willem-Paul Brinkman (* These authors contributed equally to this work.)

Type Extended Abstract

Venue ACM International Conference on Intelligent Virtual Agents

Content This extended abstract details the process of translating the standardized Artificial-Social-Agent Questionnaire (ASAQ) to Dutch and German, highlighting the translation challenges and validation results. Summative assessments with bilingual participants showed strong correlations with the original English version, ensuring the translations' reliability for evaluating human-agent interactions in Dutch and German-speaking populations.

[Schiller et al., 2024]

Title A lecturer's voice quality and its effect on memory, listening effort, and perception in a VR environment

Authors Isabel S. Schiller, Carolin Breuer, Lukas Aspöck, **Jonathan Ehret**, Andrea Bönsch, Torsten W. Kuhlen, Janina Fels, Sabine J. Schlittmeier

Type Journal Paper

Journal Scientific Reports

Content This paper evaluates whether the voice quality of a virtual speaker (here in particular hoarseness) has an influence on the (perceived) listening effort for participants taking part in a lecture in virtual reality. The evaluation was conducted with 50 participants to evaluate whether VR poses a viable option for conducting such research. However, no effect on the objectively measured listening effort could be found, while subjective metrics showed differences.

C.3. Published Datasets/Software

[Ermert et al., 2022]

Title AuViST - An Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm

Authors Cosima A. Ermert, Chinthusa Mohanathasan, **Jonathan Ehret**, Sabine J. Schlittmeier, Torsten W. Kuhlen, Janina Fels

Type Dataset

Content The Audio-Visual Speech and Text (AuViST) database provides additional material to the heardtext-recall (HTR) paradigm by Schlittmeier et al. [2023]. German audio recordings in male and female voice as well as matching face tracking data are provided for all texts.

Availability doi:10.18154/RWTH-2023-05543

[Gilbert et al., 2024]

Title RWTH VR Group Unreal Engine Toolkit

Authors David Gilbert, **Jonathan Ehret**, Marcel Krüger, Sebastian Pape, Daniel Rupp, Kris Tabea Helwig, Timon Römer, Simon Oehrl, Ali Can Demiralp, Faysal Qurabi, Kamil Karwacki, Torsten W. Kuhlen

Type Unreal Engine Plugin

Content The RWTH VR Toolkit is an Unreal Engine Plugin that can be used to facilitate the development of VR applications. Especially the possibility to run those applications on HMDs but just as well deploy them into a CAVE while debugging a desktop setting is a key feature.

Availability doi:10.5281/ZENODO.10817754

Title *StudyFramework*

Authors **Jonathan Ehret**, Andrea Bönsch, Marius Schmeling, Malte Kögel, Patrick Nossol, Paul Weiser, Konstantin Kühlem

Type Unreal Engine Plugin

Content The StudyFramework facilitates setting up and conducting factorial-design studies using the Unreal Engine. It is described in Sec. 3.

Availability `https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/unreal-study-framework`

Title *Character Plugin*

Authors **Jonathan Ehret**, Andrea Bönsch, Daniel Rupp, Denis Kuznietsov, Patrick Nossol, Jan-Nikjas Hartmann, Faysal Qurabi, Malte Kögel

Type Unreal Engine Plugin

Content This Unreal Engine plugin facilitates adding virtual characters, especially virtual humans, to an application. It provides beyond others basic functionality with regard to gazing (see Se. 5.2), gestures (see Sec. 5.3.2), facial animation (see Sec. 5.1, locomotion and crowd algorithms.

Availability `https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/character-plugin`

Title *Avatar Plugin*

Authors Patrick Nossol, **Jonathan Ehret**, Andrea Bönsch, Jonas Schüppen

Type Unreal Engine Plugin

Content This Unreal Engine plugin facilitates adding body-avatars to a VR application using an HMD. Its basic functionality is described in Sec. 3.4.

Availability `https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/avatar-plugin`

Title *MoCap Plugin*

Authors Patrick Nossol, **Jonathan Ehret**, Andrea Bönsch, Jonas Schüppen

Type Unreal Engine Plugin

Content This Unreal Engine plugin provides similar functionality to the Avatar Plugin shown above, however, has a focus on capturing full-body motion for later replay. It uses the full 6-tracker setup described in Sec. 3.4.

Availability `https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/MoCapPlugin`

Bibliography

Nadine Aburumman, Marco Gillies, Jamie A. Ward, and Antonia F.de C. Hamilton. Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions. *International Journal of Human-Computer Studies*, 164:102819, 8 2022. doi:10.1016/j.ijhcs.2022.102819.

David Ackermann, Christoph Böhm, Fabian Brinkmann, and Stefan Weinzierl. The Acoustical Effect of Musicians' Movements During Musical Performances. *Acta Acust united Ac*, 105: 356–367, 2019. doi:10.3813/AAA.919319.

Alex Adkins. *The Importance of Hand Motions for Communication and Interaction in Virtual Reality*. PhD thesis, 2022. URL `https://tigerprints.clemson.edu/all_dissertations`.

Henny Admoni and Brian Scassellati. Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction*, 6:25–63, 2017. doi:10.5898/jhri.6.1.admoni.

Leonel Aguilar, Michal Gath-Morad, Jascha Grübel, Jasper Ermatinger, Hantao Zhao, Stefan Wehrli, Robert W. Sumner, Ce Zhang, Dirk Helbing, Christoph Hölscher, Dirk Helbing, and Christoph Hölscher. Experiments as Code and its application to VR studies in human-building interaction. *Scientific Reports*, 14, 2024. doi:10.1038/s41598-024-60791-3.

Nele Albers, Andrea Bönsch, Jonathan Ehret, Boleslav A. Khodakov, and Willem-Paul Brinkman. German and Dutch Translations of the Artificial-Social-Agent Questionnaire Instrument for Evaluating Human-Agent Interaction. In *IVA '24: 24th ACM International Conference on Intelligent Virtual Agents*, 2024. doi:10.1145/3652988.3673928.

Amy L. Alexander, Tad Brunyé, Jason Sidman, and Shawn A. Weil. From gaming to training: A review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in pc-based simulations and games. In *DARWARS Training Impact Group*, 2005. URL `http://cs.engr.uky.edu/~sgware/reading/papers/alexander2005gaming.pdf`.

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum*, 39:487–496, 5 2020. doi:10.1111/cgf.13946.

Maryam Alimardani, Robyn de Roode, Julija Vaitonyte, and Max M. Louwerse. Effect of a Virtual Agent's Appearance and Voice on Uncanny Valley and Trust in Human-Agent Collaboration. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, 2024. doi:10.1145/3652988.3673970.

Jont B. Allen and David A. Berkley. Image Method for Efficiently Simulating Small-room Acoustics. *J. Acoust. Soc. Am.*, 65:943–950, 1979. doi:10.1121/1.382599.

Katrin Allmendinger. Social presence in synchronous virtual learning situations: The role of nonverbal signals displayed by avatars. *Educational Psychology Review*, 22:41–56, 3 2010. doi:10.1007/s10648-010-9117-8.

Keith Anderson, Elisabeth André, T. Baur, Sara Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, Kaśka Porayska-Pomsta, P. Rizzo, and Nicolas Sabouret. The TARDIS Framework: Intelligent Virtual Agents for Social Coaching in Job Interviews. In *Advances in Computer Entertainment*, pages 476–491, 2013. doi:10.1007/978-3-319-03161-3_35.

Razvan Andrei, Oliver Benjamin Engermann, Christian Glerup Sørensen, and Markus Löchtefeld. Examining the Effects of Eye-tracking on Dyadic Conversations in Virtual Reality. In *International Conference on Mobile and Ubiquitous Multimedia (MUM '23)*, 2023. doi:10.1145/3626705.3627794.

Sean Andrist, Bilge Mutlu, and Michael Gleicher. Conversational Gaze Aversion for Virtual Agents. In *International Workshop on Intelligent Virtual Agents*, pages 249–262, 2013. doi:10.1007/978-3-642-40415-3_22.

Deepali Aneja, Daniel McDuff, and Shital Shah. A High-Fidelity Open Embodied Avatar with Lip Syncing and Expression Capabilities. In *2019 International Conference on Multimodal Interaction*, pages 69–73. ACM, 2019. doi:10.1145/3340555.3353744.

Rui Filipe Antunes and Luís Correia. Virtual simulations of ancient sites inhabited by autonomous characters: Lessons from the development of Easy-population. *Digital Applications in Archaeology and Cultural Heritage*, 26, 2022. doi:10.1016/j.daach.2022.e00237.

Takayuki Arai. The Replication of Chiba and Kajiyama's Mechanical Models of the Human Vocal Cavity. *J. Phonetic Soc. Jpn.*, 5:31–38, 2001. doi:10.24467/onseikenkyu.5.2_31.

Stephanie Arévalo Arboleda, Christian Kunert, Jakob Hartbrich, Christian Schneiderwind, Chenyao Diao, Christoph Gerhardt, Tatiana Surdu, Florian Weidner, Wolfgang Broll, Stephan Werner, and Alexander Raake. Beyond Looks: A Study on Agent Movement and Audiovisual Spatial Coherence in Augmented Reality. In *IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 502–512, 2024. doi:10.1109/VR58804.2024.00071.

Matthew P Aylett and Christopher J Pidcock. The CereVoice Characterful Speech Synthesiser SDK. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, pages 413–414, 2007. doi:10.1007/978-3-540-74997-4_65.

R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412, 2008. doi:10.1016/j.jml.2007.12.005.

Jeremy N Bailenson and Nick Yee. Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments. *Society*, 16:814–819, 2005. doi:10.1111/j.1467-9280.2005.01619.x.

Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 10:583–598, 2001. doi:10.1162/105474601753272844.

Jeremy N Bailenson, Eyal Aharoni, Andrew C Beall, Rosanna E Guadagno, Aleksandar Dimov, and Jim Blascovich. Comparing behavioral and self-report measures of embodied agents' social presence in immersive virtual environments. In *Proceedings of the 7th Annual International Workshop on PRESENCE*, 2004.

Jeremy N. Bailenson, Kim Swinth, Crystal Hoyt, Susan Persky, Alex Dimov, and Jim Blascovich. The Independent and Interactive Effects of Embodied-Agent Appearance and Behavior on Self-Report, Cognitive, and Behavioral Markers of Copresence in Immersive Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 14:379–393, 8 2005. doi:10.1162/105474605774785235.

Grayson Bailey, Olaf Kammler, Rene Weiser, Sven Schneider, and Ekaterina Fuchkina. Integrating Immersive Virtual Environment User Studies into Architectural Design Practice: A Pre-Occupancy User Study of Train Station Waiting Preferences With VREVAL. *Proceedings of the 2022 Annual Modeling and Simulation Conference, ANNSIM 2022*, pages 644–655, 2022. doi:10.23919/ANNSIM55834.2022.9859371.

Gérard Bailly, Stephan Raidt, and Frédéric Elisei. Gaze, conversational agents and face-to-face communication. *Speech Communication*, 52:598–612, 6 2010. doi:10.1016/J.SPECOM.2010.02.015.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 59–66, 2018. doi:10.1109/FG.2018.00019.

Christoph Bartneck, Dana Kulic, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int J Soc Robot*, 1:71–81, 2009. doi:10.1007/s12369-008-0001-3.

Douglas Bates, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 2015. doi:10.18637/jss.v067.i01.

Helen L Bear and Richard Harvey. Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, 95:40–67, 2017. doi:10.1016/j.specom.2017.07.001.

Adam O. Bebko and Nikolaus F. Troje. bmlTUX: Design and Control of Experiments in Virtual Reality and Beyond. *i-Perception*, 11, 7 2020. doi:10.1177/2041669520938400.

Durand R Begault. *3-D Sound for Virtual Reality and Multimedia*. PhD thesis, 2000. URL https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20010044352.pdf.

Gottfried Behler, Martin Pollow, and Michael Vorländer. Measurements of Musical Instruments with Surrounding Spherical Arrays. In *Proc. Acoustics Nantes Conf.*, pages 761–765, 2012. URL https://hal.archives-ouvertes.fr/hal-00811213/.

Gary Bente, Sabine Rüggenberg, Nicole C. Krämer, and Felix Eschenburg. Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human communication research*, 34:287–318, 2008. doi:10.1111/j.1468-2958.2008.00322.x.

Anna Rita Bentivoglio, Susan B. Bressman, Emanuele Cassetta, Donatella Carretta, Pietro Tonali, and Alberto Albanese. Analysis of blink rate patterns in normal subjects. *Movement Disorders*, 12:1028–1034, 1997. doi:10.1002/MDS.870120629.

Alexander Berger and Markus Kiefer. Comparison of Different Response Time Outlier Exclusion Methods: A Simulation Study. *Frontiers in Psychology*, 12:675558, 2021. doi:10.3389/FPSYG.2021.675558.

Kirsten Bergmann and Manuela Macedonia. A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance. In *International Conference on Intelligent Virtual Agents*, pages 139–148, 2013. doi:10.1007/978-3-642-40415-3_12.

Kirsten Bergmann, Holly P. Branigan, and Stefan Kopp. Exploring the alignment space - lexical and gestural alignment with real and virtual humans. *Frontiers in ICT*, 2, 2015. doi:10.3389/FICT.2015.00007.

Ulysses Bernardet, Sin-Hwa Kanq, Andrew Feng, Steve Dipaola, and Ari Shapiro. Speech Breathing in Virtual Humans: An Interactive Model and Empirical Study. In *IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE)*, pages 1–9, 2019. doi:10.1109/VHCIE.2019.8714737.

Elisabetta Bevacqua, Dirk Heylen, Catherine Pelachaud, and Marion Tellier. Facial Feedback Signals for ECAs. In *AISB*, pages 328–334, 2007. URL https://hal.science/hal-00433312.

Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. A listening agent exhibiting variable behaviour. In *International Workshop on Intelligent Virtual Agents*, pages 262–269, 2008. doi:10.1007/978-3-540-85483-8_27.

Elisabetta Bevacqua, Sathish Pammi, Sylwia Julia Hyniewska, Marc Schröder, and Catherine Pelachaud. Multimodal Backchannels for Embodied Conversational Agents. In *International Conference on Intelligent Virtual Agents*, pages 194–200, 2010. doi:10.1007/978-3-642-15892-6_21.

Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, pages 2027–2036, 2021. doi:10.1145/3474085.3475223.

F Biocca, C Harms, and J Burgoon. Criteria and scope conditions for a theory and measure of social presence. *Presence: Teleoperators & Virtual Environments*, 12:456–480, 2001. URL https://www.researchgate.net/publication/239665882_Criteria_And_Scope_Conditions_For_A_Theory_And_Measure_Of_Social_Presence.

Frank Biocca. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *International Workshop on Presence*, 2001. URL http://matthewlombard.com/ISPR/Proceedings/2001/Biocca2.pdf.

Frank Biocca and Chad Harms. Defining and measuring social presence: Contribution to the networked minds theory and measure. In *Proceedings of PRESENCE*, 2002. URL http://matthewlombard.com/ISPR/Proceedings/2002/Biocca%20and%20Harms.pdf.

Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT press, 1997.

Loën Boban, Lucas Strauss, Hugo Decroix, Bruno Herbelin, and Ronan Boulic. Unintentional synchronization with self-avatar for upper- and lower-body movements. *Frontiers in Virtual Reality*, 4, 2 2023. doi:10.3389/frvir.2023.1073549.

Dan Bohus and Eric Horvitz. Facilitating Multiparty Dialog with Gaze, Gesture, and Speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 2010. doi:10.1145/1891903.1891910.

Michael Bonfert, Thomas Muender, Ryan P Mcmahan, Frank Steinicke, Doug Bowman, Virginia Tech, Rainer Malaka, and Tanja Döring. The Interaction Fidelity Model: A Taxonomy to Communicate the Different Aspects of Fidelity in Virtual Reality. *International Journal of Human–Computer Interaction*, 1, 2024. doi:10.1080/10447318.2024.2400377.

Miguel Borges, Andrew Symington, Brian Coltin, Trey Smith, and Rodrigo Ventura. HTC Vive: Analysis and Accuracy Improvement. *IEEE International Conference on Intelligent Robots and Systems*, pages 2610–2615, 12 2018. doi:10.1109/IROS.2018.8593707.

Elif Bozkurt, Yücel Yemez, and Engin Erzin. Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication*, 85:29–42, 2016. doi:10.1016/j.specom.2016.10.004.

David Bridges, Alain Pitiot, Michael R. MacAskill, and Jonathan W. Peirce. The timing megastudy: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, 8, 7 2020. doi:10.7717/peerj.9414.

John Brooke. SUS-A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 1995.

Jack Brookes, Matthew Warburton, Mshari Alghadier, Mark Mon-Williams, and Faisal Mushtaq. Studying human behavior with virtual reality: The Unity Experiment Framework. *Behavior Research Methods*, 52:455–463, 4 2020. doi:10.3758/s13428-019-01242-0.

Judee K Burgoon, Joseph A Bonito, Paul Benjamin Lowry, Sean L Humpherys, Gregory D Moody, James E Gaskin, and Justin Scott Giboney. Application of Expectancy Violations Theory to communication with and judgments about embodied agents during a decision-making task. *Int. J. Human-Computer Studies*, 91:24–36, 2016. doi:10.1016/j.ijhcs.2016.02.002.

Kenneth P Burnham and David R Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33, 2004. doi:10.1177/0049124104268644.

Hendrik Buschmeier and Stefan Kopp. Communicative Listener Feedback in Human-Agent Interaction: Artificial Speakers Need to Be Attentive and Adaptive Socially Interactive Agents Track. In *Proc. ofthe 17th International Conference on Autonom- ous Agents and Multiagent Systems (AAMAS 2018)*, pages 1213–1221, 2018. URL `https://dl.acm.org/doi/abs/10.5555/3237383.3237880`.

Andrea Bönsch. *Social Wayfinding Strategies to Explore Immersive Virtual Environments*. RWTH Aachen University, 2024. doi:10.18154/RWTH-2024-07063.

Andrea Bönsch, Benjamin Weyers, Jonathan Wendt, Sebastian Freitag, and Torsten W. Kuhlen. Collision avoidance in the presence of a virtual agent in small-scale virtual environments. In *IEEE Symposium on 3D User Interfaces (3DUI)*, pages 145–148, 2016. doi:10.1109/3DUI.2016.7460045.

Andrea Bönsch, Robert Trisnadi, Jonathan Wendt, Tom Vierjahn, and Torsten W Kuhlen. Score-Based Recommendation for Efficiently Selecting In-dividual Virtual Agents in Multi-Agent Systems. In *Proceedings of 23rd ACM Symposium on Virtual Reality Software and Technology*, page 74, 2017a. doi:10.1145/3139131.3141215.

Andrea Bönsch, Tom Vierjahn, and Torsten W. Kuhlen. Evaluation of Approaching-Strategies of Temporarily Required Virtual Assistants in Immersive Environments. In *IEEE Symposium on 3D User Interfaces*, pages 69–72, 2017b. doi:10.1109/3DUI.2017.7893319.

Andrea Bönsch, Jonathan Wendt, Heiko Overath, Özgür Gürerk, Christine Harbring, Christian Grund, Thomas Kittsteiner, and Torsten W. Kuhlen. Peers at work: Economic real-effort experiments in the presence of virtual co-workers. *Proceedings - IEEE Virtual Reality Conference*, pages 301–302, 2017c. doi:10.1109/VR.2017.7892296.

Andrea Bönsch, Sina Radke, Heiko Overath, Laura M Asché, Jonathan Wendt, Tom Vierjahn, Ute Habel, and Torsten W. Kuhlen. Social VR: How Personal Space is Affected by Virtual Agents' Emotions. In *Proceedings of IEEE Virtual Reality Conference 2018*, 2018a. doi:10.1109/VR.2018.8446480.

Andrea Bönsch, Sina Radke, Jonathan Wendt, Tom Vierjahn, Ute Habel, and Torsten W Kuhlen. Towards Understanding the Influence of a Virtual Agent's Emotional Expression on Personal Space. In *IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE)*, 2018b. URL `https://vr.rwth-aachen.de/media/papers/VHCIE18_Boensch_EmotionEffectsPS_opt.pdf`.

Andrea Bönsch, Jan Hoffman, Jonathan Wendt, and Torsten W. Kuhlen. Evaluation of Omnipresent Virtual Agents embedded as Temporarily Required Assistants in Immersive Environments. In *Virtual Humans and Crowds for Immersive Environments (VHCIE) at IEEE VR. 2019*, 2019. doi:10.1109/VHCIE.2019.8714726.

Andrea Bönsch, Sebastian J. Barton, Jonathan Ehret, and Torsten W. Kuhlen. Immersive Sketching to Author Crowd Movements in Real-time. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA 2020*, page 3, 2020a. doi:10.1145/3383652.3423883.

Andrea Bönsch, Alexander R. Bluhm, Jonathan Ehret, and Torsten W. Kuhlen. Inferring a User's Intent on Joining or Passing by Social Groups. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA 2020*, 2020b. doi:10.1145/3383652.3423862.

Andrea Bönsch, Ute Habel, Jonathan Ehret, Sina Radke, and Torsten Kuhlen. *The Impact of a Virtual Agent's Non-Verbal Emotional Expression on a User's Personal Space Preferences: Supplemental Material.* 2020c. doi:10.18154/RWTH-2020-09106.

Andrea Bönsch, Marcel Jonda, Jonathan Ehret, and Torsten W. Kuhlen. Towards a Graphical User Interface for Exploring and Fine-Tuning Crowd Simulations. *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, pages 160–164, 2020d. doi:10.1109/VRW50115.2020.00033.

Andrea Bönsch, Sina Radke, Jonathan Ehret, Ute Habel, and Torsten W. Kuhlen. The Impact of a Virtual Agent's Non-Verbal Emotional Expression on a User's Personal Space Preferences. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA 2020*, 2020e. doi:10.1145/3383652.3423888.

Andrea Bönsch, Katharina Güths, Jonathan Ehret, and Torsten W. Kuhlen. Indirect User Guidance by Pedestrians in Virtual Environments. In *International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, 2021a. doi:10.2312/egve.20211336.

Andrea Bönsch, David Hashem, Jonathan Ehret, and Torsten W. Kuhlen. Being Guided or Having Exploratory Freedom: User Preferences of a Virtual Agent's Behavior in a Museum. *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, IVA 2021*, pages 33–40, 2021b. doi:10.1145/3472306.3478339.

Andrea Bönsch, Daniel Rupp, Jonathan Ehret, and Torsten W. Kuhlen. An Embodied Conversational Agent Supporting Scene Exploration by Switching between Guiding and Accompanying. In *IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE)*, 2022. URL `https://vr.rwth-aachen.de/media/papers/212/LBR_VHCIE_Boensch_red.pdf`.

Andrea Bönsch, Till Sittart, Jonathan Ehret, and Torsten W. Kuhlen. Where Do They Go? Overhearing Conversing Pedestrian Groups during Scene Exploration. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, 2023a. doi:10.1145/3570945.3607351.

Andrea Bönsch, Lukas B. Zimmermann, Jonathan Ehret, and Torsten W. Kuhlen. Whom Do You Follow? Pedestrian Flows Constraining the User's Navigation during Scene Exploration. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, 2023b. doi:10.1145/3570945.3607350.

Andrea Bönsch, Jonathan Ehret, Daniel Rupp, and Torsten W. Kuhlen. Wayfinding in immersive virtual environments as social activity supported by virtual agents. *Frontiers in Virtual Reality*, 4:1334795, 2024. doi:10.3389/FRVIR.2023.1334795.

João Paulo Cabral, Benjamin R Cowan, Katja Zibrek, and Rachel McDonnell. The Influence of Synthetic Voice on the Evaluation of a Virtual Character. In *Proc. Interspeech*, pages 229–233, 2017. doi:10.21437/Interspeech.2017-325.

Angelo Cafaro, Nadine Glas, and Catherine Pelachaud. The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, pages 911–920. International Foundation for Autonomous Agents and Multiagent Systems, 2016. URL `https://dl.acm.org/citation.cfm?id=2937059`.

Julia Cambre and Chinmay Kulkarni. One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proc. ACM Hum.-Comput. Interact*, 3: 19, 2019. doi:10.1145/3359325.

Qiongdan Cao, Hui Yu, Paul Charisse, Si Qiao, and Brett Stevens. Is High-Fidelity Important for Human-like Virtual Avatars in Human Computer Interactions? *International Journal of Network Dynamics and Intelligence*, pages 15–23, 3 2023. doi:10.53941/IJNDI0201008.

Colleen M Carpinella, Michael A Perez, Alisa B Wyman, and Steven J Stroessner. The Robotic Social Attributes Scale (RoSAS): Development and Validation. In *ACM/IEEE International Conference on human-robot interaction*, pages 254 – 262, 2017. doi:10.1145/2909824.3020208.

J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjailmsson, and H. Yan. Embodiment in conversational interfaces: Rea. *Conference on Human Factors in Computing Systems - Proceedings*, pages 520–527, 1999. doi:10.1145/302979.303150.

Justine Cassell. Embodied conversational interface agents. *Communications of the ACM*, pages 70–78, 2000. doi:10.1145/332051.332075.

Justine Cassell, Yukiko I Nakano, Timothy W Bickmore, Candace L Sidner, and Charles Rich. Non-Verbal Cues for Discourse Structure. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 114–123, 2001a. doi:10.3115/1073012.1073028.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. BEAT: the Behavior Expression Animation Toolkit. In *Proceedings of SIGGRAPH'01*, pages 477–486. ACM Press, 2001b. doi:10.1145/383259.383315.

Justine Cassell, Paul A Tepper, Kim Ferriman, and Kristina Striegnitz. *Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions*, pages 133–160. 2007. doi:10.1002/9780470512470.ch8.

Roser Cañigueral and Antonia F.de C. Hamilton. Being watched: Effects of an audience on eye gaze and prosocial behaviour. *Acta Psychologica*, 195:50–63, 4 2019. doi:10.1016/j.actpsy.2019.02.002.

Roser Cañigueral, Jamie A. Ward, and Antonia F.de C. Hamilton. Effects of being watched on eye gaze and facial displays of typical and autistic individuals during conversation. *Autism*, 25:210–226, 1 2021. doi:10.1177/1362361320951691.

Young Woon Cha, Husam Shaik, Qian Zhang, Fan Feng, Andrei State, Adrian Ilie, and Henry Fuchs. Mobile, Egocentric Human Body Motion Reconstruction Using Only Eyeglasses-mounted Cameras and a Few Body-worn Inertial Sensors. In *IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 607–616, 2021. doi:10.1109/VR50410.2021.00087.

J C P Chan, H Leung, J K T Tang, and T Komura. A Virtual Reality Dance Training System Using Motion Capture Technology. *IEEE Transactions on Learning Technologies*, 4:187–195, 2011. doi:10.1109/TLT.2010.27.

Noël Chateau, Valérie Maffiolo, Nathalie Pican, and Marc Mersiol. The Effect of Embodied Conversational Agents' Speech Quality on Users' Attention and Emotion. In *International Conference on Affective Computing and Intelligent Interaction*, pages 652–659, 2005. doi:10.1007/11573548_84.

Jiali Chen, Yong Liu, Zhimeng Zhang, Changjie Fan, and Yu Ding. Text-driven Visual Prosody Generation for Embodied Conversational Agents. In *ACM International Conference on Intelligent Virtual Agents (IVA '19)*, pages 108–110. ACM, 2019. doi:10.1145/3308532.3329445.

Lei Chen and Mary P. Harper. Multimodal Floor Control Shift Detection. In *Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interfaces*, pages 15–22, 2009. doi:10.1145/1647314.1647320.

Yanbo Cheng and Yingying Wang. Evaluating the Effect of Outfit on Personality Perception in Virtual Characters. *Virtual Worlds*, 3:21–39, 2024. doi:10.3390/virtualworlds3010002.

CC Chiu and S Marsella. Gesture generation with low-dimensional embeddings. *Proceedings of the 2014 international conference*, 2014. URL `http://dl.acm.org/citation.cfm?id=2615857`.

Chung-Cheng Chiu and Stacy Marsella. How to train your avatar: a data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*, pages 127–140, 2011. doi:10.1007/978-3-642-23974-8_14.

Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. Predicting Co-verbal Gestures: A Deep and Temporal Modeling Approach. In *International Conference on Intelligent Virtual Agents*, pages 152–166. Springer, Cham, 2015. doi:10.1007/978-3-319-21996-7_17.

Minsoo Choi, Alexandros Koilias, Matias Volonte, Dominic Kao, and Christos Mousas. Exploring the Appearance and Voice Mismatch of Virtual Characters. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2023. doi:10.1109/ISMAR-Adjunct60411.2023.00118.

Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, Ari Shapiro, and Stefan Scherer. Exploring feedback strategies to improve public speaking: An interactive Virtual Audience Framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1143–1154, 2015. doi:10.1145/2750858.2806060.

Mingyuan Chu and Peter Hagoort. Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General*, 143:1726–1741, 2014. doi:10.1037/A0036281.

Emna Chérif and Jean-François Lemoine. Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice:. *Recherche et Applications en Marketing (English Edition)*, 34:28–47, 2019. doi:10.1177/2051570719829432.

Andrew P. Clark, Kate L. Howard, Andy T. Woods, Ian S. Penton-Voak, and Christof Neumann. Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PLOS ONE*, 13:e0190393, 1 2018. doi:10.1371/JOURNAL.PONE.0190393.

Michelle Cohn, Patrik Jonell, Taylor Kim, Jonas Beskow, and Georgia Zellou. Embodiment and gender interact in alignment to TTS voices. In *Proceedings of the Cognitive Science Society*, pages 220–226, 2020. URL `https://cognitivesciencesociety.org/cogsci20/papers/0044/0044.pdf`.

Sean Commins, Joseph Duffin, Keylor Chaves, Diarmuid Leahy, Kevin Corcoran, Michelle Caffrey, Lisa Keenan, Deirdre Finan, and Conor Thornberry. NavWell: A simplified virtual-reality platform for spatial navigation and memory experiments. *Behavior Research Methods*, 52:1189–1207, 6 2020. doi:10.3758/S13428-019-01310-5.

Susan Wagner Cook, Howard S. Friedman, Katherine A. Duggan, Jian Cui, and Voicu Popescu. Hand Gesture and Mathematics Learning: Lessons From an Avatar. *Cognitive Science*, 41: 518–535, 2017. doi:10.1111/cogs.12344.

Carolina Cruz-Neira, Daniel Sandin, Robert V Kenyon, and John C Hart. The cave - audio visual experience automatic virtual environment. *Communications of The ACM*, 1992. doi:10.1145/129888.129892.

Ronald Cumbal, Daniel Alexander Kazzi, Vincent Winberg, and Olov Engwall. Shaping unbalanced multi-party interactions through adaptive robot backchannels. In *ACM International Conference on Intelligent Virtual Agents (IVA '22)*. Association for Computing Machinery, Inc, 9 2022. doi:10.1145/3514197.3549680.

Ronald Cumbal, Reshma Kantharaju, Maike Paetzel-Prüsmann, and James Kennedy. Let Me Finish First - The Effect of Interruption-Handling Strategy on the Perceived Personality of a Social Agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '24)*, 2024. doi:10.1145/3652988.3673916.

Douglas W. Cunningham and Christian Wallraven. *Experimental Design: From User Studies to Psychphysics*. A K Peters/CRC Press, 2012. doi:10.1201/b11308.

Anne Cutler. *Errors of stress and intonation*, pages 67–80. New York, Academic Press, 1980.

Erwan David, Jesús Gutiérrez, Melissa Lè Hoa Võ, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. The Salient360! toolbox: Handling gaze data in 3D made easy. *Computers & Graphics*, 119:103890, 2024. doi:10.1016/J.CAG.2024.103890.

Robert O. Davis, Joseph Vincent, and Taejung Park. Reconsidering the Voice Prinicple with Non-native Language Speakers. *Computers and Education*, 140, 2019. doi:10.1016/j.compedu.2019.103605.

Aline W. de Borst and Beatrice de Gelder. Is it the real deal? Perception of virtual characters versus humans: an affective cognitive neuroscience perspective. *Front. Psychol.*, 6, 2015. doi:10.3389/fpsyg.2015.00576.

Ferdinand de Coninck, Zerrin Yumak, Guntur Sandino, and Remco Veltkamp. Nonverbal Behavior Generation for Virtual Characters in Group Conversations. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 2019. doi:10.1109/AIVR46125.2019.00016.

Henrique Galvan Debarba, Sylvain Chague, and Caecilia Charbonnier. On the Plausibility of Virtual Body Animation Features in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, 28:1880–1893, 4 2022. doi:10.1109/TVCG.2020.3025175.

Zhigang Deng, J. P. Lewis, and Ulrich Neumann. Automated eye motion using texture synthesis. *IEEE Computer Graphics and Applications*, 25:24–30, 3 2005. doi:10.1109/MCG.2005.35.

Zhigang Deng, Pei-Ying Chiang, Pamela Fox, and Ulrich Neumann. Animating Blendshape Faces by Cross-Mapping Motion Capture Data. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 43–48, 2006. doi:10.1145/1111411.1111419.

David Devault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014. URL `https://dl.acm.org/doi/abs/10.5555/2615731.2617415`.

David Devault, Johnathan Mell, and Jonathan Gratch. Toward Natural Turn-Taking in a Virtual Human Negotiation Agent. In *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015. URL `https://cdn.aaai.org/ocs/10335/10335-45289-1-PB.pdf`.

Christina Dicke, Viljakaisa Aaltonen, Anssi Rämö, and Miikka Vilermo. Talk to me: The Influence of Audio Quality on the Perception of Social Presence. In *Proc BCS HCI*, pages 309–318, 2010. URL `https://dl.acm.org/doi/abs/10.5555/2146303.2146349`.

Yu Ding, Yuting Zhang, Meihua Xiao, and Zhigang Deng. A Multifaceted Study on Eye Contact based Speaker Identification in Three-party Conversations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3011–3021. Association for Computing Machinery, 5 2017. doi:10.1145/3025453.3025644.

Tiffany D. Do, Ryan P. McMahan, and Pamela J. Wisniewski. A New Uncanny Valley? The Effects of Speech Fidelity and Human Listener Gender on Social Perceptions of a Virtual-Human Speaker. In *CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2022. doi:10.1145/3491102.3517564.

Georgiana Cristina Dobre, Marco Gillies, Patrick Falk, Jamie A Ward, Antonia F de C Hamilton, and Xueni Pan. Direct Gaze Triggers Higher Frequency of Gaze Change: An Automatic Analysis of Dyads in Unstructured Conversation. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21)*, pages 735–739. ACM, 2021. doi:10.1145/3462244.3479962.

G. Doherty-Sneddon and F. G. Phelps. Gaze aversion: A response to cognitive or social difficulty? *Memory & Cognition*, 33:727–733, 6 2005. doi:10.3758/BF03195338.

Madeline Easley, Jung Hyup Kim, Siddarth Mohanty, Ching-Yun Yu, Varun Pulipati, Sara Mostowfi, Fang Wang, Kangwon Seo, Danielle Oprean, and Danielle Oprean. The Effects of a Virtual Instructor with Realistic Lip Sync in an Augmented Reality Environment. In *16th International Conference on Virtual, Augmented and Mixed Reality*, 2024. doi:10.1007/978-3-031-61041-7_1.

Jens Edlund and Jonas Beskow. Mushypeek: A framework for online investigation of audiovisual dialogue phenomena. *Language and Speech*, 52:351–367, 2009. doi:10.1177/0023830909103179.

Allen L. Edwards. Balanced Latin-Square Designs in Psychological Research. *The American Journal of Psychology*, 64:598–603, 1951. doi:10.2307/1418200.

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization. *ACM Transactions on Graphics*, 35:1–11, 2016. doi:10.1145/2897824.2925984.

Jonathan Ehret. Doctoral Consortium : Verbal Interactions with Embodied Conversational Agents. In *Doctoral Consortium at the 22nd ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, 2022. URL `https://vr.rwth-aachen.de/media/papers/216/IVA_DC_2022_Ehret.pdf`.

Jonathan Ehret, Jonas Stienen, Chris Brozdowski, Andrea Bönsch, Irene Mittelberg, Michael Vorländer, and Torsten W. Kuhlen. Evaluating the Influence of Phoneme-Dependent Dynamic Speaker Directivity of Embodied Conversational Agents' Speech. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA 2020*. Association for Computing Machinery, Inc, 10 2020. doi:10.1145/3383652.3423863.

Jonathan Ehret, Andrea Bönsch, Lukas Aspöck, Christine T. Röhr, Stefan Baumann, Martine Grice, Janina Fels, and Torsten W. Kuhlen. Do Prosody and Embodiment Influence the

Perceived Naturalness of Conversational Agents' Speech? *ACM Transactions on Applied Perception*, 18:21:1–15, 2021. doi:10.1145/3486580.

Jonathan Ehret, Andrea Bönsch, and Torsten W. Kuhlen. Natural Turn-Taking with Embodied Conversational Agents. In *IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE)*, 2022. URL `http://vr.rwth-aachen.de/media/papers/213/LBR_VHCIE_Ehret_red.pdf`.

Jonathan Ehret, Andrea Bönsch, Patrick Nossol, Cosima A Ermert, Chinthusa Mohanathasan, Sabine J Schlittmeier, Janina Fels, and Torsten W Kuhlen. Who's next? Integrating Non-Verbal Turn-Taking Cues for Embodied Conversational Agents. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023. doi:10.1145/3570945.3607312.

Jonathan Ehret, Andrea Bönsch, Janina Fels, Sabine J. Schlittmeier, and Torsten W. Kuhlen. StudyFramework: Comfortably Setting up and Conducting Factorial-Design Studies Using the Unreal Engine. *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 442–449, 3 2024a. doi:10.1109/VRW62533.2024.00087.

Jonathan Ehret, Andrea Bönsch, Isabel S. Schiller, Carolin Breuer, Lukas Aspöck, Janina Fels, Sabine J. Schlittmeier, and Torsten W. Kuhlen. Audiovisual Coherence: Is Embodiment of Background Noise Sources a Necessity? *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 61–67, 3 2024b. doi:10.1109/VRW62533.2024.00017.

Jonathan Ehret, Valentin Dasbach, Jan-Nikjas Hartmann, Janina Fels, Torsten W. Kuhlen, and Andrea Bönsch. Exploring Gaze Dynamics: Initial Findings on the Role of Listening Bystanders in Conversational Interactions. In *IEEE VR Workshop on Virtual Humans and Crowds for Immersive Environments (VHCIE)*, 2025a. doi:10.1109/VRW66409.2025.00151.

Jonathan Ehret, Jonas Schüppen, Chinthusa Mohanathasan, Cosima A. Ermert, Janina Fels, Sabine J. Schlittmeier, Torsten W. Kuhlen, and Andrea Bönsch. Objectifying Social Presence: Evaluating Multimodal Degraders in ECAs Using the Heard Text Recall Paradigm. *IEEE Transactions on Visualization and Computer Graphics [in press]*, 2025b. doi:10.1109/TVCG.2025.3636079.

Paul Ekman and Wallace V. Friesen. Facial Action Coding System. *Environmental Psychology & Nonverbal Behavior*, 1978. doi:10.1037/t27734-000.

Erik Ekstedt and Gabriel Skantze. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Interspeech 2022*, pages 5190–5194. ISCA, 9 2022a. doi:10.21437/Interspeech.2022-10955.

Erik Ekstedt and Gabriel Skantze. How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551. Association for Computational Linguistics, 2022b. doi:10.18653/v1/2022.sigdial-1.51.

Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*, pages 754–768, 2021. doi:10.1145/3472749.3474784.

Cathy Ennis and Carol O'Sullivan. Perceptually plausible formations for virtual conversers. *Computer Animation and Virtual Worlds*, 23:321–329, 2012. doi:10.1002/cav.1453.

Cosima A. Ermert, Jonathan Ehret, Torsten W. Kuhlen, Chinthusa Mohanathasan, Sabine J. Schlittmeier, and Janina Fels. Audio-Visual Content Mismatches in the Serial Recall Paradigm. In *49. Jahrestagung für Akustik, Hamburg, Germany, DAGA 2023*, pages 1429–1430, 2023. URL `https://pub.dega-akustik.de/DAGA_2023/data/articles/000029.pdf`.

Cosima Antonia Ermert, Chinthusa Mohanathasan, Jonathan Ehret, Sabine Janina Schlittmeier, Torsten W. Kuhlen, and Janina Fels. AuViST - An Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm. 2022. doi:10.18154/RWTH-2023-05543.

Lyke Esselink, Marloes Oomen, and Floris Roelofsen. Truedepth Measurements of Facial Expressions: Sensitivity to the Angle Between Camera and Face. In *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2023. doi:10.1109/ICASSPW59220.2023.10193107.

Elodie Etienne, Anne Lise Leclercq, Angélique Remacle, Laurence Dessart, and Michaël Schyns. Perception of avatars nonverbal behaviors in virtual reality. *Psychology & Marketing*, 40: 2464–2481, 2023. doi:10.1002/MAR.21871.

Elodie Etienne, Marion Ristorcelli, Sarah Saufnay, Aurélien Quilez, Rémy Casanova, Michaël Schyns, and Magalie Ochs. A Systematic Review on the Socio-affective Perception of IVAs' Multi-modal behaviour A Systematic Review on the Socio-affective Perception of IVAs' Multi-modal behaviour. In ACM. In *ACM International Conference on Intelligent Virtual Agents (IVA '24)*, 2024. doi:10.1145/3652988.3673943.

Mireille Fares, Catherine Pelachaud, and Nicolas Obin. Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. *European Signal Processing Conference*, 2022-August:593–597, 2022. doi:10.23919/EUSIPCO55093.2022.9909519.

Martin Feick, Niko Kleer, Anthony Tang, and Antonio Krüger. The Virtual Reality Questionnaire Toolkit. *UIST 2020 - Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 68–69, 10 2020. doi:10.1145/3379350.3416188.

Ylva Ferstl. Generating Emotionally Expressive Look-At Animation. In *ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2023. doi:10.1145/3623264.3624438.

Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. In *21th ACM International Conference on Intelligent Virtual Agents (IVA '21)*, 2021. doi:10.1145/3472306.3478338.

Edina Fintor, Lukas Aspöck, Janina Fels, and Sabine J. Schlittmeier. The role of spatial separation of two talkers' auditory stimuli in the listener's memory of running speech: listening effort in a non-noisy conversational setting. *International Journal of Audiology*, pages 1–9, 2021. doi:10.1080/14992027.2021.1922765.

Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem Paul Brinkman. What are we measuring anyway? - A literature survey of questionnaires used in studies reported in the intelligent virtual agent conferences. In *IVA 2019 - Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 159–161. Association for Computing Machinery, Inc, 7 2019. doi:10.1145/3308532.3329421.

Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial-social-agent questionnaire. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8. ACM, 9 2022. doi:10.1145/3514197.3549612.

George Fletcher, Donal Egan, Rachel McDonnell, and Darren Cosker. Improving motion matching for VR avatars by fusing inside-out tracking with outside-in 3D pose estimation. In *The 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2023. URL `https://researchportal.bath.ac.uk/en/publications/improving-motion-matching-for-vr-avatars-by-fusing-inside-out-tra`.

Jean-Pierre Gagné, Jana Besser, and Ulrike Lemke. Behavioral Assessment of Listening Effort Using a Dual-Task Paradigm. *Trends in Hearing*, 21, 2017. doi:10.1177/2331216516687287.

Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M. Angela Sasse. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 529–536. Association for Computing Machinery (ACM), 2003. doi:10.1145/642611.642703.

Maria Garau, Mel Slater, David-Paul Pertaub, and Sharif Razzaque. The responses of people to virtual humans in an immersive virtual environment. *Teleoperators & Virtual Environments*, 14:104–116, 2005. doi:10.1162/1054746053890242.

Aaron S. Geller, Ian K. Schleifer, Per B. Sederberg, Joshua Jacobs, and Michael J. Kahana. PyEPL: A cross-platform experiment-programming library. *Behavior Research Methods*, 39: 950–958, 2007. doi:10.3758/BF03192990.

Kallirroi Georgila, Alan W Black, Kenji Sagae, and David Traum. Practical Evaluation of Human and Synthesized Speech for Virtual Human Dialogue Systems. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, page 3519–3526, 2012. URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/562_Paper.pdf`.

Michele Geronazzo, Simone Spagnol, and Federico Avanzini. Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26: 1247–1260, 7 2018. doi:10.1109/TASLP.2018.2821846.

Michael A. Gerzon. Ambisonics in Multichannel Broadcasting and Video. *Journal of the Audio Engineering Society*, 33:859–871, 11 1985. URL `https://aes2.org/publications/elibrary-page/?id=4419`.

David Gilbert, Jonathan Ehret, Marcel Krüger, Sebastian Paper, Daniel Rupp, Kris Tabea Helwig, Timon Römer, Simon Oehrl, Torsten W. Kuhlen, Ali Can Demiralp, Faysal Qurabi, and Kamil Karwacki. RWTH VR Group Unreal Engine Toolkit. 2024. doi:10.5281/ZENODO.10817754.

Sarah Gillet, Ronald Cumbal, André Pereira, José Lopes, Olov Engwall, and Iolanda Leite. Robot Gaze Can Mediate Participation Imbalance in Groups with Diferent Skill Levels. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*, pages 303–311, 2021. doi:10.1145/3434073.3444670.

Guilherme Goncalves, Pedro Monteiro, Hugo Coelho, Miguel Melo, and Maximino Bessa. Systematic Review on Realism Research Methodologies on Immersive Virtual, Augmented and Mixed Realities. *IEEE Access*, 9:89150–89161, 2021. doi:10.1109/ACCESS.2021.3089946.

Li Gong and Clifford Nass. When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Human Communication Research*, 33:163–193, 2007. doi:10.1111/j.1468-2958.2007.00295.x.

Mar Gonzalez-Franco, Antonella Maselli, Dinei Florencio, Nikolai Smolyanskiy, and Zhengyou Zhang. Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific reports*, 7:1–11, 2017. doi:10.1038/s41598-017-04201-x.

Ific Goude, Alexandre Bruckert, Anne-Helene Olivier, Julien Pettre, Remi Cozot, Kadi Bouatouch, Marc Christie, and Ludovic Hoyet. Real-time Multi-map Saliency-driven Gaze Behavior for Non-conversational Characters. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–13, 2023. doi:10.1109/TVCG.2023.3244679.

Sarah Graf and Valentin Schwind. Inconsistencies of Presence Questionnaires in Virtual Reality. In *26th ACM Symposium on Virtual Reality Software and Technology*, pages 1–3, 2020. doi:10.1145/3385956.3422105.

Simone Grassini and Karin Laumann. Questionnaire Measures and Physiological Correlates of Presence: A Systematic Review. *Frontiers in Psychology*, 0:349, 3 2020. doi:10.3389/FPSYG.2020.00349.

Jonathan Gratch, Jeff Rickel, Elisabeth André, Justine Cassell, Eric Petajan, and Norman Badler. Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intell. Sys.*, 17:54–63, 2002. doi:10.1109/MIS.2002.1024753.

Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, R J van der Werf, and Louis-Philippe Morency. Virtual Rapport. In *Intelligent Virtual Agents: 6th International Conference*, 2006. doi:10.1007/11821830_2.

Jonathan Gratch, David DeVault, and Gale Lucas. The Benefits of Virtual Humans for Teaching Negotiation. In *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, pages 283–294, 2016. doi:10.1007/978-3-319-47665-0_25.

P. De Greef and W. A. Ijsselsteijn. Social Presence in a Home Tele-Application. *CyberPsychol. Behav.*, 4:307–315, 7 2004. doi:10.1089/109493101300117974.

Peter Green and Catriona J. Macleod. SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7:493–498, 4 2016. doi:10.1111/2041-210X.12504.

Helena Grillon and Daniel Thalmann. Eye contact as trigger for modification of virtual character behavior. In *Virtual Rehabilitation*, pages 205–211, 2008. doi:10.1109/ICVR.2008.4625161.

Foteini Grivokostopoulou, Konstantinos Kovas, and Isidoros Perikos. The Effectiveness of Embodied Pedagogical Agents and Their Impact on Students Learning in Virtual Worlds. *Applied Sciences*, 10:1739, 2020. doi:10.3390/app10051739.

Jascha Grübel. The design, experiment, analyse, and reproduce principle for experimentation in virtual reality. *Frontiers in Virtual Reality*, 4:1069423, 4 2023. doi:10.3389/frvir.2023.1069423.

Jascha Grübel, Raphael Weibel, Mike Hao Jiang, Christoph Hölscher, Daniel A. Hackman, and Victor R. Schinazi. EVE: A framework for experiments in virtual environments. In *Spatial Cognition X - 13th Biennial Conference, KogWis 2016*, pages 159–176. Springer Verlag, 2017. doi:10.1007/978-3-319-68189-4_10.

Rosanna E. Guadagno, Jim Blascovich, Jeremy N. Bailenson, and Cade Mccall. Virtual Humans and Persuasion: The Effects of Agency and Behavioral Realism. *Media Psychology*, 10: 1–22, 2007. URL `https://www.tandfonline.com/doi/abs/10.1080/15213260701300865`.

Manuel Guimarães, Rui Prada, Pedro A Santos, João Dias, Arnav Jhala, and Samuel Mascarenhas. The impact of virtual reality in the social presence of a virtual agent. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual*, 10 2020. doi:10.1145/3383652.3423879.

Siqi Guo, Nicoletta Adamo, and Christos Mousas. Developing a Scale for Measuring the Believability of Virtual Agents. In *ICAT-EGVE 2023 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, pages 45–52, 2023. doi:10.2312/egve.20231312.

Tom Gurion, Patrick G T Healey, and Julian Hough. Comparing models of speakers' and listeners' head nods. In *SEMDIAL 2020 (WatchDial), The 24nd workshop on the Semantics and Pragmatics of Dialogue*, 2020. URL `https://www.semdial.org/anthology/Z20-Gurion_semdial_0013.pdf`.

Ramiro H Gálvez, Agustín Gravano, Stefan Beňuš, Rivka Levitan, Marian Trnka, and Julia Hirschberg. An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars. *Speech Communication*, 124:46–67, 2020. doi:10.1016/j.specom.2020.07.007.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation In The Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. doi:10.1109/CVPR.2018.00762.

Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning Speech-driven 3D Conversational Gestures from Video. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2021. doi:10.1145/3472306.3478335.

Uri Hadar, Timothy J. Steiner, E. C. Grant, and F. Clifford Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46, 1983. doi:10.1016/0167-9457(83)90004-0.

Jihun Hamm, Christian G. Kohler, Ruben C. Gur, and Ragini Verma. Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods*, 200:237–256, 9 2011. doi:10.1016/J.JNEUMETH.2011.06.023.

Eugy Han, Cyan DeVeaux, Mark Roman Miller, Gabriella M. Harari, Jeffrey T. Hancock, Nilam Ram, and Jeremy N. Bailenson. Alone Together, Together Alone: The Effects of Social Context on Nonverbal Behavior in Virtual Reality. *PRESENCE: Virtual and Augmented Reality*, 33:425–451, 2024. doi:10.1162/pres_a_00432.

Chad Harms and Frank Biocca. Internal consistency and reliability of the networked minds social presence measure. In *Proceedings of the Seventh Annual International Workshop on Presence*, 2004. URL `http://matthewlombard.com/ISPR/Proceedings/2004/Harms%20and%20Biocca.pdf`.

Arne Hartz, Björn Guth, Mathis Jording, Kai Vogeley, and Martin Schulte-Rüther. Temporal Behavioral Parameters of On-Going Gaze Encounters in a Virtual Environment. *Frontiers in Psychology*, 12:673982, 8 2021. doi:10.3389/FPSYG.2021.673982.

Béatrice S. Hasler, Bernhard Spanlang, and Mel Slater. Virtual race transformation reverses racial in-group bias. *PloS one*, 12:e0174965, 4 2017. doi:10.1371/journal.pone.0174965.

Aleshia Taylor Hayes, Charles E. Hughes, and Jeremy Bailenson. Identifying and Coding Behavioral Indicators of Social Presence With a Social Presence Behavioral Coding System. *Frontiers in Virtual Reality*, 3:62, 6 2022. doi:10.3389/FRVIR.2022.773448.

Yuan He, André Pereira, and Taras Kucherenko. Evaluating data-driven co-speech gestures of embodied conversational agents through real-time interaction. In *ACM International Conference on Intelligent Virtual Agents (IVA '22)*. Association for Computing Machinery, Inc, 9 2022. doi:10.1145/3514197.3549697.

Ilona Heldal, Ralph Schroeder, Anthony Steed, Ann-Sofie Axelsson, Maria Spante, and Josef Wideström. Immersiveness and symmetry in copresent scenarios. In *IEEE Virtual Reality*, pages 171–178, 2005. doi:10.1109/VR.2005.1492771.

Maartje M.E. Hendrikse, Gerard Llorach, Giso Grimm, and Volker Hohmann. Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. *Speech Communication*, 101:70–84, 2018. doi:10.1016/J.SPECOM.2018.05.008.

Claudia Hendrix and Woodrow Barfield. Presence in virtual environments as a function of visual and auditory cues. In *Proceedings Virtual Reality Annual International Symposium'95*, pages 74–82, 1995. doi:10.1109/VRAIS.1995.512482.

Daniel Hepperle, Christian Felix Purps, Jonas Deuchler, and Matthias Wölfel. Aspects of visual avatar appearance: self-representation, display type, and uncanny valley. *Visual Computer*, 38:1227–1244, 4 2022. doi:10.1007/S00371-021-02151-0.

Dirk Heylen. Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3:241–267, 2006. doi:10.1142/S0219843606000746.

Dirk Heylen. *Listening Heads*, pages 241–259. Springer Berlin Heidelberg, 2008. doi:10.1007/978-3-540-79037-2_13.

Dirk Heylen, Ivo van Es, Anton Nijholt, and Betsy van Dijk. Controlling the gaze of conversational agents. *Advances in natural multimodal dialogue systems*, pages 245–262, 2005. doi:10.1007/1-4020-3933-6_11.

Dirk Heylen, Elisabetta Bevacqua, Marion Tellier, and Catherine Pelachaud. Searching for prototypical facial feedback signals. In *International Workshop on Intelligent Virtual Agents*, pages 147–153. Springer Verlag, 2007. doi:10.1007/978-3-540-74997-4_14.

Darragh Higgins, Katja Zibrek, Joao Cabral, Donal Egan, and Rachel McDonnell. Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans. *Computers & Graphics*, 104:116–128, 2022. doi:10.1016/j.cag.2022.03.009.

Laurie Hiyakumoto, Scott Prevost, and Justine Cassell. Semantic and Discourse Information for Text-to-Speech Intonation. In *Concept to Speech Generation Systems*, pages 47–56, 1997. URL `https://aclanthology.org/W97-1207.pdf`.

Anna Hjalmarsson and Catharine Oertel. Gaze direction as a back-channel inviting cue in dialogue. In *IVA 2012 workshop on Realtime Conversational Virtual Agents*, 2012. URL `https://core.ac.uk/reader/572089790#page=52`.

Laura Hoffmann, Nicole C. Krämer, Anh Lam-Chi, and Stefan Kopp. Media Equation Revisited: Do Users Show Polite Reactions towards an Embodied Agent? In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*, pages 159–165, 2009. doi:10.1007/978-3-642-04380-2_19.

Gijs A. Holleman, Roy S. Hessels, Chantal Kemner, and Ignace T.C. Hooge. Implying social interaction and its influence on gaze behavior to the eyes. *PLOS ONE*, 15:e0229203, 2 2020. doi:10.1371/JOURNAL.PONE.0229203.

Judith Holler and Kobin H Kendrick. Unaddressed participants' gaze in multi-person interaction: optimizing recipiency. *Frontiers in Psychology*, 6:76–89, 2015. doi:10.3389/fpsyg.2015.00098.

Gernot Horstmann and Sebastian Loth. The Mona Lisa Illusion-Scientists See Her Looking at Them Though She Isn't. *i-Perception*, 10, 2019. doi:10.1177/2041669518821702.

Matthew B Hoy. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical reference services quarterly*, 37:81–88, 2018. doi:10.1080/02763869.2018.1404391.

Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, 6:1049, 2015. doi:10.3389/fpsyg.2015.01049.

Lixing Huang, Louis Philippe Morency, and Jonathan Gratch. Learning backchannel prediction model from parasocial consensus sampling: A subjective evaluation. In *International Conference on Intelligent Virtual Agents*, pages 159–172, 2010. doi:10.1007/978-3-642-15892-6_17.

Lixing Huang, Louis Philippe Morency, and Jonathan Gratch. Virtual rapport 2.0. In *Intern. Workshop on Intell. Virtual Agents*, pages 68–79, 2011. doi:10.1007/978-3-642-23974-8_8.

Wei-Chia Huang, Sai-Keung Wong, Matias Volonte, and Sabarish V. Babu. Impact of Socio-Demographic Attributes and Mutual Gaze of Virtual Humans on Users' Visual Attention and Collision Avoidance in VR. *IEEE Transactions on Visualization and Computer Graphics*, 30:6146–6163, 2024. doi:10.1109/TVCG.2023.3329515.

Jonatan Hvass, Oliver Larsen, Kasper Vendelbo, Niels Nilsson, Rolf Nordahl, and Stefania Serafin. Visual realism and presence in a virtual reality game. In *The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. IEEE, 2017. doi:10.1109/3DTV.2017.8280421.

Felix Immohr, Gareth Rendle, Erik Hübner, Annika Neidhardt, Luljeta Sinani, Bernd Fröhlich, and Alexander Raake. Exploring Factors Influencing Audiovisual Plausibility and Co-Presence in Multi-Modal VR Communication. In *DAGA 2022 Stuttgart*, 2022.

Felix Immohr, Gareth Rendle, Anton Lammert, Annika Neidhardt, Victoria Meyer Zur Heyde, Bernd Froehlich, and Alexander Raake. Evaluating the Effect of Binaural Auralization on Audiovisual Plausibility and Communication Behavior in Virtual Reality. In *IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 849–858, 2024. doi:10.1109/VR58804.2024.00104.

Benjamin Inden, Zofia Malisz, Petra Wagner, and Ipke Wachsmuth. Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. *ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction*, pages 181–188, 2013. doi:10.1145/2522848.2522890.

Carlos T. Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. A Speech-Driven Hand Gesture Generation Method and Evaluation in Android Robots. In *IEEE Robotics and Automation Letters*, pages 3757–3764, 2018. doi:10.1109/LRA.2018.2856281.

Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. Multimodal fusion using respiration and gaze for predicting next speaker in multi-party meetings. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 99–106, 2015. doi:10.1145/2818346.2820755.

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Trans. Interact. Intell. Syst.*, 6:4:1–33, 2016. doi:10.1145/2757284.

Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel Modeling? In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*, pages 131–138, 2021. doi:10.1145/3472306.3478360.

Sárándi István, Alexander Hermans, and Bastian Leibe. Learning 3D Human Pose Estimation From Dozens of Datasets using a Geometry-Aware Autoencoder to Bridge Between Skeleton Formats. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2956–2966, 2023. doi:10.1109/WACV56688.2023.00297.

Ryoya Ito, Celso M de Melo, Jonathan Gratch, and Kazunori Terada. Emotional Expression Help Regulate the Appropriate Level of Cooperation with Agents. In *12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 27–36, 2024.

Laurent Itti, Nitin Dhavale, and Fréderic Pighin. Photorealistic attention-based gaze animation. In *IEEE International Conference on Multimedia and Expo*, pages 521–524, 2006. doi:10.1109/ICME.2006.262440.

Peter Jax and Peter Vary. Bandwidth Extension of Speech Signals: A Catalyst for the Introduction of Wideband Speech Coding? *IEEE Commun. Mag.*, 44:106–111, 5 2006. doi:10.1109/MCOM.2006.1637954.

Leif Johannsen. *Numerical simulation of voice directivity patterns for different phonemes.* PhD thesis, 2021.

Swati Johar. *Emotion, Affect and Personality in Speech: The Bias of Language and Paralanguage.* Springer International Publishing, 2016. doi:10.1007/978-3-319-28047-9.

Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. Gaze and Turn-Taking Behavior in Casual Conversational Interactions. *ACM Trans. Interact. Intell. Syst*, 3:12:1–30, 2013. doi:http://dx.doi.org/10.1145/2499474.2499481.

Starkey Duncan Jr. On the structure of speaker-auditor interaction during speaking turns. *Language in Socienty*, 2:161–180, 1974. URL `https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S0047404500004322`.

Yvonne Jung, Arjan Kuijper, Dieter Fellner, Michael Kipp, Jan Miksatko, Jonathan Gratch, and Daniel Thalmann. Believable Virtual Characters in Human-Computer Dialogs. In *Eurographics 2011 – State of the Art Reports*, 2011. URL `http://www.michaelkipp.de/publication/Jungetal11.pdf`.

Mathieu Jégou, Liv Lefebvre, and Pierre Chevaillier. A Continuous Model for the Management of Turn-Taking in User-Agent Spoken Interactions Based on the Variations of Prosodic Signals. In *International Conference on Intelligent Virtual Agents*, pages 389–398. Springer, Cham, 2015. doi:10.1007/978-3-319-21996-7_42.

Ni Kang, Willem Paul Brinkman, M. Birna Van Riemsdijk, and Mark Neerincx. The design of virtual audiences: Noticeable and recognizable behavioral styles. *Computers in Human Behavior*, 55:680–694, 2016. doi:10.1016/j.chb.2015.10.008.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion. *ACM Trans. Graph.*, 36: 94, 2017. doi:10.1145/3072959.3073658.

Brian F. G. Katz, Fabien Prezat, and Christophe D'Alessandro. Human Voice Phoneme Directivity Pattern Measurements. *J. Acoust. Soc. Am.*, 120:3359–3359, 11 2006. doi:10.1121/1.4781486.

Judith K Keller, Agon Kusari, Sophie Czok, Birgit Simgen, Frank Steinicke, and Esther K Diekhof. ACHOO-Bless you! Sense of Presence can provoke Proactive Mucosal Immune Responses in Immersive Human-Agent Interactions. In *IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 557–567, 2024. doi:10.1109/VR58804.2024.00076.

Adam Kendon. Some uses of the head shake. *Gesture*, 2:147–182, 2002. doi:10.1075/gest.2.2.03ken.

Adam Kendon and Mark Cook. The consistency of gaze patterns in social interaction. *British Journal of Psychology*, 60:481–494, 1969. doi:10.1111/J.2044-8295.1969.TB01222.X.

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

Bavo Van Kerrebroeck, Giusy Caruso, and Pieter-Jan Maes. A Methodological Framework for Assessing Social Presence in Music Interactions in Virtual Reality. *Frontiers in Psychology*, 12, 2021. doi:10.3389/fpsyg.2021.663725.

Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy Bailenson, Gerd Bruder, and Greg F. Welch. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 105–114. IEEE, 10 2018. doi:10.1109/ISMAR.2018.00039.

Simon Kimmel, Frederike Jung, Andrii Matviienko, Wilko Heuten, and Susanne Boll. Let's Face It: Influence of Facial Expressions on Social Presence in Collaborative Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16. ACM, 2023. doi:10.1145/3544548.3580707.

Scott A. King and Richard E. Parent. Creating speech-synchronized animation. *IEEE Transactions on Visualization and Computer Graphics*, 11:341–352, 5 2005. doi:10.1109/TVCG.2005.43.

Mendel Kleiner, Bengt-Inge Dalenbäck, and Peter Svensson. Auralization - An Overview. *J. Audio Engin Soc*, 41:861–875, 1993. URL `https://aes2.org/publications/elibrary-page/?id=5597`.

Malte Kob. *Physical Modeling of the Singing Voice*. PhD thesis, RWTH Aachen University, 2002.

Ned Kock. Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward e-communication tools. *IEEE Transactions on Professional Communication*, 48:117–130, 6 2005. doi:10.1109/TPC.2005.849649.

Nikolina Koleva, Martín Villalba, Maria Staudte, and Alexander Koller. The Impact of Listener Gaze on Predicting Reference Resolution. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2:812–817, 2015. doi:10.3115/v1/P15-2133.

Kyveli Kompatsiari, Francesca Ciardo, and Agnieszka Wykowska. To follow or not to follow your gaze: The interplay between strategic control and the eye contact effect on gaze-induced attention orienting. *Journal of Experimental Psychology: General*, 151:121–136, 2022. doi:10.1037/xge0001074.

Andrew Kope, Caroline Rose, and Michael Katchabaw. Modeling Autobiographical Memory for Believable Agents. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 23–29, 2013. doi:10.1609/AIIDE.V9I1.12686.

Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Intern. Workshop on Intell. Virtual Agents*, pages 205–217. Springer Verlag, 2006. doi:10.1007/11821830_17.

Karel Kreijns, Kate Xu, and Joshua Weidlich. Social Presence: Conceptualization and Measurement. *Educational Psychology Review*, 34:139–170, 3 2022. doi:10.1007/S10648-021-09623-8.

Brigitte Krenn, Stephanie Schreitter, and Friedrich Neubarth. Speak to me and I tell you who you are! A language-attitude study in a cultural-heritage application. *AI & Society*, 32: 65–77, 2017. doi:10.1007/s00146-014-0569-0.

Niklas Krome and Stefan Kopp. Towards Real-time Co-speech Gesture Generation in Online Interaction in Social XR. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023. doi:10.1145/3570945.3607315.

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. In *IVA 2019 - Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 97–104. Association for Computing Machinery, Inc, 7 2019. doi:10.1145/3308532.3329472.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 242–250, 2020. doi:10.1145/3382507.3418815.

Torsten Wolfgang Kuhlen and Bernd Hentschel. Quo vadis CAVE: Does immersive visualization still matter? *IEEE Computer Graphics and Applications*, 34:14–21, 9 2014. doi:10.1109/MCG.2014.97.

Junyeong Kum, Sunghun Jung, and Myungho Lee. The Effect of Eye Contact in Multi-Party Conversations with Virtual Humans and Mitigating the Mona Lisa Effect. *Electronics*, 13: 430, 2024. doi:10.3390/electronics13020430.

Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brebisson, and Yoshua Bengio. ObamaNet: Photo-realistic lip-sync from text. In *arXiv preprint arXiv:1801.01442*, 12 2017. URL `http://arxiv.org/abs/1801.01442`.

Christos Kyrlitsias and Despina Michael-Grigoriou. Asch conformity experiment using immersive virtual reality. *Computer Animation and Virtual Worlds*, 29:e1804, 9 2018. doi:10.1002/CAV.1804.

Christos Kyrlitsias and Despina Michael-Grigoriou. Social Interaction With Agents and Avatars in Immersive Virtual Environments: A Survey. *Frontiers in Virtual Reality*, 0:168, 1 2022. doi:10.3389/FRVIR.2021.786665.

Christos Kyrlitsias, Despina Michael-Grigoriou, Domna Banakou, and Maria Christofi. Social Conformity in Immersive Virtual Environments: The Impact of Agents' Gaze Behavior. *Frontiers in Psychology*, 11:530913, 9 2020. doi:10.3389/FPSYG.2020.02254.

Jari Kätsyri, Klaus Förger, Meeri Mäkäräinen, and Tapio Takala. A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, 6, 2015. doi:10.3389/fpsyg.2015.00390.

Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in Neurorobotics*, 14:593732, 2020. doi:10.3389/fnbot.2020.593732.

Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings ofthe SIGDIAL 2017 Conference*, pages 127–136, 2017. doi:10.18653/v1/W17-5516.

Divesh Lala, Koji Inoue, Tatsuya Kawahara, and Kei Sawada. Backchannel Generation Model for a Third Party Listener Agent. *HAI 2022 - Proceedings of the 10th Conference on Human-Agent Interaction*, pages 114–122, 12 2022. doi:10.1145/3527188.3561926.

Luchcha Lam, Minsoo Choi, Magzhan Mukanova, Klay Hauser, Fangzheng Zhao, Richard Mayer, Christos Mousas, and Nicoletta Adamo-Villani. Effects of Body Type and Voice Pitch on Perceived Audio-Visual Correspondence and Believability of Virtual Characters. In *ACM Symposium on Applied Perception*, 2023. doi:10.1145/3605495.3605791.

Maurice Lamb, Malin Brundin, Estela Perez Luque, and Erik Billing. Eye-Tracking Beyond Peripersonal Space in Virtual Reality: Validation and Best Practices. *Frontiers in Virtual Reality*, 3, 2022. doi:10.3389/frvir.2022.864653.

Christian Lang, Sven Wachsmuth, Marc Hanheide, and Heiko Wersing. Facial Communicative Signals. *International Journal of Social Robotics*, 4:249–262, 2012. doi:10.1007/s12369-012-0145-z.

Stephen RH Langton, Roger J. Watt, and Vicki Bruce. Do the eyes have it? Cues to the direction of social attention. *Trends in cognitive sciences*, 4:50–59, 2000. doi:10.1016/S1364-6613(99)01436-9.

Marc Erich Latoschik and Carolin Wienrich. Congruence and Plausibility, not Presence?! Pivotal Conditions for XR Experiences and Effects, a Novel Approach. *Frontiers in Virtual Reality*, 3, 4 2022. doi:10.3389/frvir.2022.694433.

PaulJ. Laurienti, RobertA. Kraft, JosephA. Maldjian, JonathanH. Burdette, and MarkT. Wallace. Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, 158:405–414, 1 2004. doi:10.1007/s00221-004-1913-2.

B. H. Le, Xiaohan Ma, and Zhigang Deng. Live Speech Driven Head-and-Eye Motion Generators. *IEEE Transactions on Visualization and Computer Graphics*, 18:1902–1914, 2012. doi:10.1109/TVCG.2012.74.

Geonsun Lee, Yeol Lee, Guan-Ming Su, and Dinesh Manocha. ”May I Speak?”: Multi-modal Attention Guidance in Social VR Group Conversations. *TVCG Special Issue on the 2024 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2024. doi:10.1109/TVCG.2024.3372119.

Kwan Min Lee and Clifford Nass. Designing social presence of social actors in human computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 289–296. ACM, 2003. doi:10.1145/642611.642662.

Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction. *Journal of communication*, 56:754–772, 2006. doi:10.1111/j.1460-2466.2006.00318.x.

Sooha Park Lee, Jeremy B. Badler, and Norman I. Badler. Eyes alive. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques - SIGGRAPH ’02*, page 637, 2002. doi:10.1145/566570.566629.

D. J. Leiner. SoSci Survey (Version 3.2.28) [Computer software]. Available at https://www.soscisurvey.de, 2021.

Russell V. Lenth. R package emmeans: Estimated marginal means, 2024. URL `https://rvlenth.github.io/emmeans/`.

Tobias Lentz. *Binaural technology for virtual reality.* PhD thesis, RWTH Aahcen University, 2008.

Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. Virtual Reality System with Integrated Sound Field Simulation and Reproduction. *EURASIP J. Adv. Sig. Pr.*, 1:1–19, 12 2007. doi:10.1155/2007/70540.

Thomas Leonard and Fred Cummins. The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26:1457–1471, 2011. doi:10.1080/01690965.2010.500218.

Giovanna Leone. Nodding without Understanding: An Explorative Study of How Adolescents Listen to Their Teachers. In *International Conference on Social Informatics*, pages 137–144, 2012. doi:10.1109/SocialInformatics.2012.63.

Stephen C. Levinson. Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, 20:6–14, 1 2016. doi:10.1016/j.tics.2015.10.010.

J P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. Practice and Theory of Blendshape Facial Model. *Eurographics (State of the Art Reports)*, 1, 2014. URL `http://graphics.cs.uh.edu/wp-content/papers/2014/2014-EG-blendshape_STAR.pdf`.

Margaux Lhommet and Stacy Marsella. Gesture with meaning. In *International Workshop on Intelligent Virtual*, pages 303–312, 2013. doi:10.1007/978-3-642-40415-3_27.

Jie Li, Yiping Kong, Thomas Röggla, Francesca De Simone, Swamy Ananthanarayan, Huib de Ridder, Abdallah El Ali, and Pablo Cesar. Measuring and Understanding Photo Sharing Experiences in Social Virtual Reality. In *CHI Conference on Human Factors in Computing Systems (CHI 2019)*. Glasgow, 2019a. doi:10.1145/3290605.3300897.

Yang Li, Jin Huang, Feng Tian, Hong An Wang, and Guo Zhong Dai. Gesture interaction in virtual reality. *Virtual Reality & Intelligent Hardware*, 1:84–112, 2 2019b. doi:10.3724/SP.J.2096-5796.2018.0006.

Chang Liu, Qunfen Lin, Tencent Games, Zijiao Zeng, and Ye Pan. EmoFace: Audio-driven Emotional 3D Face Animation. In *IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 2024a. doi:10.1109/VR58804.2024.00060.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024b. URL `https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EMAGE_Towards_Unified_Holistic_Co-Speech_Gesture_Generation_via_Expressive_Masked_CVPR_2024_paper.html`.

Kang Liu and Joern Ostermann. Realistic head motion synthesis for an image-based talking head. In *Face and Gesture 2011*, pages 221–226. IEEE, 3 2011. doi:10.1109/FG.2011.5771401.

Joan Llobera and Caecilia Charbonnier. Physics-based character animation for Virtual Reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 56–57, 3 2022. doi:10.1109/VRW55335.2022.00021.

Gerard Llorach, Alun Evans, Josep Blat, Giso Grimm, and Volker Hohmann. Web-Based Live Speech-Driven Lip-Sync. In *8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, pages 1–4. IEEE, 9 2016. doi:10.1109/VS-GAMES.2016.7590381.

Josep Llorca-Bofí and Michael Vorländer. IHTAclassroom. Multi-detailed 3D architecture model for sound perception research in Virtual Reality. Zenodo, 2021. doi:10.5281/zenodo.4629716.

Josep Llorca-Bofí, Christian Dreier, Jonas Heck, Jonas Kempin, and Michael Vorländer. IHTApark. Multi-detailed 3D architectural model for sound perception research in Virtual Reality. 2022. doi:10.5281/ZENODO.5905338.

Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. Measuring Presence: The Temple Presence Inventory. In *Proceedings of the 12th International Workshop on Presence*, pages 1–15, 2009. URL `http://matthewlombard.com/ISPR/Proceedings/2009/Lombard_et_al.pdf`.

Sebastian Loth, Gernot Horstmann, Corinna Osterbrink, and Stefan Kopp. Accuracy of perceiving precisely gazing virtual agents. In *International Conference on Intelligent Virtual Agents*, pages 263–268. ACM, 11 2018. doi:10.1145/3267851.3267852.

Jean-Luc Lugrin, Maximilian Ertl, Philipp Krop, Richard Klupfel, Sebastian Stierstorfer, Bianka Weisz, Maximilian Ruck, Johann Schmitt, Nina Schmidt, and Marc Erich Latoschik. Any "Body" There? Avatar Visibility Effects in a Virtual Reality Game. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 17–24, 2018. doi:10.1109/VR.2018.8446229.

Le Luo, Dongdong Weng, Ni Ding, Jie Hao, and Ziqi Ti. The effect of avatar facial expressions on trust building in social virtual reality. *The Visual Computer*, 39:5869–5882, 2023. doi:10.1007/s00371-022-02700-1.

Beatriz López, Álvaro Hernández-Trapote, David Pardo, Raul Santos, and María del Carmen Rodríguez Gancedo. ECA gesture strategies for robust SLDSs. In *Proceedings of the AISB Symposium on Multimodal Output Generation*, 2008.

Mark Ter Maat, Khiet P. Truong, and Dirk Heylen. How turn-taking strategies influence users' impressions of an agent. In *International Conference on Intelligent Virtual Agents*, pages 441–453. Springer, Berlin, Heidelberg, 2010. doi:10.1007/978-3-642-15892-6_48.

R. M. Maatman, Jonathan Gratch, and Stacy Marsella. Natural behavior of a listening agent. In *Intelligent Virtual Agents: 5th International Working Conference*, 2005. doi:10.1007/11550617_3.

Robert Mahony, Tarek Hamel, and Jean-Michel Pflimlin. Nonlinear Complementary Filters on the Special Orthogonal Group. *IEEE Transactions on Automatic Control*, 53:1203–1218, 2008. doi:10.1109/TAC.2008.923738.

Guido Makransky, Lau Lilleholt, and Anders Aaby. 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, 7 2017. doi:10.1016/J.CHB.2017.02.066.

David Mal, Erik Wolf, Nina Dollinger, Mario Botsch, Carolin Wienrich, and Marc Erich Latoschik. Virtual Human Coherence and Plausibility - Towards a Validated Scale. *Proceedings - 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, pages 788–789, 2022. doi:10.1109/VRW55335.2022.00245.

Zofia Malisz, Harald Berthelsen, Jonas Beskow, and Joakim Gustafson. PROMIS: a statistical-parametric speech synthesis system with prominence control via a prominence network. In *10th ISCA Speech Synthesis Workshop*, pages 257–262, 9 2019. doi:10.21437/SSW.2019-46.

Fabrizia Mantovani, Gianluca Castelnuovo, Andrea Gaggioli, and Giuseppe Riva. Virtual reality training for health-care professionals. *Cyberpsychology and Behavior*, 6:389–395, 2003. doi:10.1089/109493103322278772.

Vladislav Maraev, Chiara Mazzocconi, Christine Howes, and Catherine Pelachaud. Towards Investigating Gaze and Laughter Coordination in Socially Interactive Agents. *International Conference on Human-Agent Interaction*, pages 473–475, 2023. doi:10.1145/3623809.3623968.

Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 25, 2013. doi:10.1145/2485895.2485900.

Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 10:325–347, 2019. doi:10.1109/TAFFC.2017.2731763.

Susana Martinez-Conde, Stephen L. Macknik, and David H. Hubel. The role of fixational eye movements in visual perception. *Nature Reviews — Neuroscience*, pages 229–240, 2004. doi:10.1038/nrn1348.

Bruno Masiero and Michael Vorländer. A Framework for the Calculation of Dynamic Crosstalk Cancellation Filters. *IEEE T. Audio Speech*, 22:1345–1354, 2014. doi:10.1109/TASLP.2014.2329184.

David McNeill. *Hand and mind: What gestures reveal about thought.* The University of Chicago Press, 1992.

Ravish Mehra, Lakulish Antani, Sujeong Kim, and Dinesh Manocha. Source and Listener Directivity for Interactive Wave-based Sound Propagation. *IEEE T. Vis. Comput. Gr.*, 20: 495–503, 2014. doi:10.1109/TVCG.2014.38.

Dario Alfonso Cuello Mejía, Hidenobu Sumioka, Hiroshi Ishiguro, and Masahiro Shiomi. Evaluating gaze behaviors as pre-touch reactions for virtual agents. *Frontiers in Psychology*, 14, 2023. doi:10.3389/fpsyg.2023.1129677.

Timo Menzel, Erik Wolf, Stephan Wenninger, Niklas Spinczyk, Lena Holderrieth, Ulrich Schwanecke, Marc Erich Latoschik, and Mario Botsch. WILDAVATARS: Smartphone-Based Reconstruction of Full-Body Avatars in the Wild. *IEEE Transactions on Visualization and Computer Graphics (under review)*, 2024. doi:10.36227/techrxiv.172503940.07538627/v1.

Gregory Mills and Remko Boschker. Using Virtual Reality to Investigate the Emergence of Gaze Conventions in Interpersonal Coordination. In *4th International Conference on Human-Computer Interaction (HCII)*, pages 564–571, 2022. doi:10.1007/978-3-031-19679-9_71.

Jūra Miniota, Siyang Wang, Jonas Beskow, Joakim Gustafson, Éva Székely, and André Pereiral. Hi robot, it's not what you say, it's how you say it. In *32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 307–314, 2023. doi:10.1109/RO-MAN57019.2023.10309427.

R Miyawaki, M Perusquia-Hernandez, N Isoyama, H Uchiyama, and K Kiyokawa. A Data Collection Protocol, Tool and Analysis for the Mapping of Speech Volume to Avatar Facial Animation. In *ICAT-EGVE*, pages 27–34, 2022. URL `https://monicaperusquia.com/publications/2022-Miyawaki-ADataCollectionProtocol,ToolandAnalysisfortheMappingofSpeechVolumetoAvatarFacialAnimation.pdf`.

Chinthusa Mohanathasan, Cosima A. Ermert, Jonathan Ehret, Janina Fels, Torsten W. Kuhlen, and Sabine J. Schlittmeier. Measuring listening effort in adverse listening conditions: Testing two dual task paradigms for upcoming audiovisual virtual reality experiments. In *22. Conference of the European Society for Cognitive Psychology, Lille, France, ESCoP*, 2022. URL `https://publications.rwth-aachen.de/record/852818/files/852818.pdf`.

Chinthusa Mohanathasan, Janina Fels, and Sabine J. Schlittmeier. Listening to two-talker conversations in quiet settings: the role of listeners' cognitive processing capabilities for memory and listening effort. *Scientific Reports*, 14:22764, 2024. doi:10.1038/s41598-024-74085-1.

Chinthusa Mohanathasan, Cosima A. Ermert, Janina Fels, Torsten W. Kuhlen, and Sabine J. Schlittmeier. Exploring short-term memory and listening effort in two-talker conversations: The influence of soft and moderate background noise. *PLoS ONE*, 20:e0318821, 2025. doi:10.1371/JOURNAL.PONE.0318821.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70–84, 1 2010. doi:10.1007/s10458-009-9092-y.

Masahiro Mori. The Uncanny Valley. *Energy*, 7:33–35, 1970.

Fariba Mostajeran, Frank Steinicke, Oscar Javier Ariza Nunez, Dimitrios Gatsios, and Dimitrios Fotiadis. Augmented Reality for Older Adults: Exploring Acceptability of Virtual Coaches for Home-based Balance Training in an Aging Population. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 4 2020. doi:10.1145/3313831.3376565.

Motionbuilder. Smooth Filter, 2024. URL `https://help.autodesk.com/view/MOBPRO/2017/ENU/?guid=GUID-19F2B63F-70F0-4B6B-AE35-3CA8C9232811`.

Samer Al Moubayed, Jonas Beskow, and Björn Granström. Auditory visual prominence From intelligibility to behavior. *Journal on Multimodal User Interfaces*, 3:299–309, 2009. doi:10.1007/s12193-010-0054-0.

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. *Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction*, pages 114–130. Springer Berlin Heidelberg, 2012. doi:10.1007/978-3-642-34584-5_9.

Pat Mulvaney, Brendan Rooney, Maximilian A. Friehs, and John Francis Leader. Social VR design features and experiential outcomes: narrative review and relationship map for dyadic agent conversations. *Virtual Reality*, 28:1–20, 2024. doi:10.1007/S10055-024-00941-0.

Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction - HRI '09*, pages 61–68, 2009. doi:10.1145/1514095.1514109.

Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems*, 1:12:1–33, 2012. doi:10.1145/2070719.2070725.

S. C. Mölbert, A. Thaler, B. J. Mohler, S. Streuber, J. Romero, M. J. Black, S. Zipfel, H.-O. Karnath, K. E. Giel, and K. E. Giel. Assessing body image in anorexia nervosa using biometric self-avatars in virtual reality: Attitudinal components rather than visual body size estimation are distorted. *Psychological Medicine*, 48:642–653, 3 2018. doi:10.1017/S0033291717002008.

Sahil Narang, Andrew Best, and DInesh Manocha. Inferring User Intent using Bayesian Theory of Mind in Shared Avatar-Agent Virtual Environments. *IEEE Transactions on Visualization and Computer Graphics*, 25:2113–2122, 5 2019. doi:10.1109/TVCG.2019.2898800.

Michael Neff. *Hand Gesture Synthesis for Conversational Characters*, pages 1–12. Springer, 2016. doi:10.1007/978-3-319-30808-1_5-1.

Michael Neff. Tunable tension for gesture animation. In *ACM International Conference on Intelligent Virtual Agents (IVA '22)*, 2022. doi:10.1145/3514197.3549631.

Michael Neff, Yingying Wang, Rob Abbott, and Marilyn Walker. Evaluating the Effect of Gesture and Language on Personality Perception in Conversational Agents. In *Proceedings of the 10th international conference on Intelligent virtual agents*, pages 222–235, 2010. doi:10.1007/978-3-642-15892-6_24.

Rolf Nordahl and Niels Christian Nilsson. *The Sound of Being There: Presence and Interactive Audio in Immersive Virtual Reality Virtual*, pages 213–233. Oxford University Press, 2014. doi:10.1093/oxfordhb/9780199797226.013.013.

Warren T. Norman. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology*, 66, 1963. doi:10.1037/h0040291.

Nahal Norouzi, Kangsoo Kim, Gerd Bruder, Austin Erickson, Zubin Choudhary, Yifan Li, and Greg Welch. A Systematic Literature Review of Embodied Augmented Reality Agents in Head-Mounted Display Environments. In *Proceedings of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments*, 2020. doi:10.2312/egve.20201264.

Kristine L Nowak and Frank Biocca. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence*, 12:481–494, 2003. doi:10.1162/105474603322761289.

Magalie Ochs, Nathan Libermann, Axel Boidin, and Thierry Chaminade. Do You Speak to a Human or a Virtual Agent? Automatic Analysis of User's Social Cues during Mediated Communication. In *In Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 197–205, 2017. doi:10.1145/3136755.3136807.

Magalie Ochs, Jérémie Bousquet, Jean Marie Pergandi, and Philippe Blache. Multimodal Behavioral Cues Analysis of the Sense of Presence and Social Presence During a Social Interaction With a Virtual Patient. *Frontiers in Computer Science*, 4:32, 4 2022. doi:10.3389/FCOMP.2022.746804.

Catharine Oertel, Marcin Włodarczak, Jens Edlund, Petra Wagner, and Joakim Gustafson. Gaze Patterns in Turn-Taking. In *INTERSPEECH 2012*, pages 2246–2246, 2012. URL `https://pub.uni-bielefeld.de/record/2502477`.

Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI*, 7, 2020. doi:10.3389/frobt.2020.00092.

Catharine Oertel, Patrik Jonell, Dimosthenis Kontogiorgos, Kenneth Funes Mora, Jean-Marc Odobez, and Joakim Gustafson. Towards an Engagement-Aware Attentive Artificial Listener for Multi-Party Interactions. *Frontiers in Robotics and AI*, 8:1–19, 7 2021. doi:10.3389/FROBT.2021.555913.

Catherine S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Frontiers in Robotics and AI*, 5, 2018. doi:10.3389/frobt.2018.00114.

Ali Oker, Florina Pecune, and Christelle Declercq. Virtual tutor and pupil interaction: A study of empathic feedback as extrinsic motivation for learning. *Education and Information Technologies*, 25:3643–3658, 2020. doi:10.1007/s10639-020-10123-5.

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

Xueni Pan and Antonia F. de C. Hamilton. Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, 109:395–417, 8 2018. doi:10.1111/bjop.12290.

Ye Pan, Shuai Tan, Shengran Cheng, Qunfen Lin, Tencent Games, Zijiao Zeng, and Kenny Mitchell. Expressive Talking Avatars. *IEEE Transactions on Visualization and Computer Graphics*, 30, 2024. doi:10.1109/TVCG.2024.3372047.

Yuliya Patotskaya, Ludovic Hoyet, Anne Hélène Olivier, Julien Pettré, and Katja Zibrek. Avoiding virtual humans in a constrained environment: Exploration of novel behavioural measures. *Computers & Graphics*, 110:162–172, 2 2023. doi:10.1016/J.CAG.2023.01.001.

Florian Pausch, Lukas Aspöck, Michael Vorländer, and Janina Fels. An Extended Binaural Real-Time Auralization System With an Interface to Research Hearing Aids for Experiments on Subjects With Hearing Loss. *Trends Hear.*, 22:1–32, 2018. doi:10.1177/2331216518800871.

Rasmus Lundby Pedersen, Lorenzo Picinali, Nynne Kajs, and François Patou. Virtual-Reality-Based Research in Hearing Science: A Platforming Approach. *Journal of the Audio Engineering Society*, 71:374–389, 6 2023. doi:10.17743/jaes.2022.0083.

David Peeters. Virtual reality: A game-changing method for the language sciences. *Psychonomic Bulletin and Review*, 26:894–900, 2019. doi:10.3758/s13423-019-01571-3.

Jonathan Peirce, Rebecca Hirst, and Michael MacAskill. *Building experiments in PsychoPy*. Sage, 2022.

Tomislav Pejsa, Sean Andrist, Michael Gleicher, and Bilge Mutlu. Gaze and attention management for embodied conversational agents. *ACM Trans. Interact. Intell. Syst*, 5:3:1–34, 2015. doi:10.1145/2724731.

Catherine Pelachaud. Modelling multimodal expression of emotion in a virtual agent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364:3539–3548, 12 2009. doi:10.1098/RSTB.2009.0186.

Sönke Pelzer, Bruno Masiero, and Michael Vorlaender. 3D Reproduction of Room Acoustics using a Hybrid System of Combined Crosstalk Cancellation and Ambisonics Playback. In *Proceedings of International Conference on Spatial Audio*, page 297, 2011. doi:10.14279/depositonce-33.

Daniel Pimentel and Charlotte Vinkers. Copresence With Virtual Humans in Mixed Reality: The Impact of Contextual Responsiveness on Social Perceptions. *Frontiers in Robotics and AI*, 8:634520, 4 2021. doi:10.3389/FROBT.2021.634520.

Sandra Poeschl and Nicola Doering. The German VR Simulation Realism Scale-Psychometric Construction for Virtual Reality Applications with Virtual Humans. *Annual Review of Cybertherapy and Telemedicine*, 11:33–37, 2013. doi:10.3233/978-1-61499-282-0-33.

Sandra Poeschl and Nicola Doering. Measuring co-presence and social presence in virtual environments - Psychometric construction of a german scale for a fear of public speaking scenario. *Annual Review of CyberTherapy and Telemedicine*, 13:58–63, 2015. doi:10.3233/978-1-61499-595-1-58.

Sandra Poeschl, Konstantin Wall, and Nicola Doering. Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. In *2013 IEEE Virtual Reality (VR)*, pages 129–130. IEEE, 2013. doi:10.1109/VR.2013.6549396.

Ronald Poppe, Khiet P. Truong, and Dirk Heylen. Backchannels: Quantity, type and timing matters. *International Conference on Intelligent Virtual Agents (IVA)*, pages 228–239, 2011. doi:10.1007/978-3-642-23974-8_25.

Barteld N. J. Postma and Brian F. G. Katz. Dynamic Voice Directivity in Room Acoustic Auralizations. In *Germ. Ann. Conf. Acoustics (DAGA)*, pages 352–355, 2016. URL `https://pub.dega-akustik.de/DAGA_2016/data/articles/000357.pdf`.

Barteld N. J. Postma, Hugo Demontis, and Brian F. G. Katz. Subjective Evaluation of Dynamic Voice Directivity forA uralizations. In *Acta Acustica united with Acustica*, pages 181–184, 2017. doi:10.3813/AAA.919045.

Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H. Bulthoff, and Michael J. Black. The Virtual Caliper: Rapid Creation of Metrically Accurate Avatars from 3D Measurements. *IEEE Transactions on Visualization and Computer Graphics*, 25:1887–1897, 2019. doi:10.1109/TVCG.2019.2898748.

Astrid M. Von Der Pütten, Nicole C. Krämer, Jonathan Gratch, and Sin Hwa Kang. ”It doesn’t matter what you are!” Explaining social effects of agents and avatars. *Computers in Human Behavior*, 26:1641–1650, 11 2010. doi:10.1016/J.CHB.2010.06.012.

Livia Qian and Gabriel Skantze. Joint Learning of Context and Feedback Embeddings in Spoken Dialogue. In *Interspeech 2024*, page arXiv:2406.07291, 2024. doi:10.48550/ARXIV.2406.07291.

Lingyun Qiu and Izak Benbasat. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems*, 25:145–182, 4 2008. doi:10.2753/MIS0742-1222250405.

Lingyun Qiu and Izak Benbasat. A study of demographic embodiments of product recommendation agents in electronic commerce. *International Journal of Human-Computer Studies*, 68:669–688, 10 2010. doi:10.1016/J.IJHCS.2010.05.005.

R-Core-Team. R: A Language and Environment for Statistical Computing, 2015. URL `http://www.r-project.org/`.

Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. From a User-created Corpus of Virtual Agent’s Non-verbal Behavior to a Computational Model of Interpersonal Attitudes. In *13th International Conference on Intelligent Virtual Agents*, pages 263–274, 2013. doi:10.1007/978-3-642-40415-3_23.

Brian Ravenet, Angelo Cafaro, Beatrice Biancardi, Magalie Ochs, and Catherine Pelachaud. Conversational behavior reflecting interpersonal attitudes in small group interactions. In *International Conference on Intelligent Virtual Agents*, pages 375–388, 2015. doi:10.1007/978-3-319-21996-7_41.

Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. Automating the Production of Communicative Gestures in Embodied Characters. *Frontiers in Psychology*, 9:1144, 7 2018. doi:10.3389/fpsyg.2018.01144.

Matthias Rehm and Elisabeth André. Where Do They Look? Gaze Behaviors of Multiple Users Interacting with an Embodied Conversational Agent. In *Intelligent Virtual Agents*, pages 241–252, 2005. doi:10.1007/11550617_21.

Rim Rekik, Stefanie Wuhrer, Ludovic Hoyet, Katja Zibrek, and Anne-Hélène Olivier. A Survey on Realistic Virtual Human Animations: Definitions, Features and Evaluations. *Computer Graphics Forum*, 43, 2024. doi:10.1111/cgf.15064.

Rutger Rienks, Ronald Poppe, and Dirk Heylen. Differences in Head Orientation Behavior for Speakers and Listeners: An Experiment in a Virtual Environment. *ACM Trans. Appl. Percept*, 7:1–13, 2010. doi:10.1145/1658349.1658351.

Carina Riest, Annett B. Jorschick, and Jan P. de Ruiter. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in Psychology*, 6:62–75, 2015. doi:10.3389/fpsyg.2015.00089.

Jens H. Rindel, Felipe Otondo, and Claus L. Christensen. Sound Source Representation for Auralization. In *Int. Symp. Room Acoustics: Design and Science*, 2004. URL `https://www.odeon.dk/pdf/RADS04-Rindel.pdf`.

David A. Robb, José Lopes, Muneeb I. Ahmad, Peter E. McKenna, Xingkun Liu, Katrin Lohan, and Helen Hastie. Seeing eye to eye: trustworthy embodiment for task-based conversational agents. *Frontiers in Robotics and AI*, 10, 2023. doi:10.3389/frobt.2023.1234767.

Thomas Robotham, Olli S. Rummukainen, Miriam Kurz, Marie Eckert, and Emanuel A.P. Habets. Comparing Direct and Indirect Methods of Audio Quality Evaluation in Virtual Reality Scenes of Varying Complexity. *IEEE Transactions on Visualization and Computer Graphics*, 2022. doi:10.1109/TVCG.2022.3150491.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs. In *International Conference on Multimodal Interaction*, pages 186–190, 2018. doi:10.1145/3242969.3242997.

Daniel Roetenberg, Henk Luinge, and Per Slycke. Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors. Technical report, 2013. URL `www.xsens.com`.

Daniel Roth, Jean-Luc Lugrin, Julia Buser, Gary Bente, Arnulph Fuhrmann, and Marc Erich Latoschik. A simplified inverse kinematic approach for embodied VR applications. In *2016 IEEE Virtual Reality (VR)*, pages 275–276, 2016. doi:10.1109/VR.2016.7504760.

Daniel Roth, Peter Kullmann, Gary Bente, Dominik Gall, and Marc Erich Latoschik. Effects of Hybrid and Synthetic Social Gaze in Avatar-Mediated Interactions. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct*, pages 103–108, 2018a. doi:10.1109/ISMAR-Adjunct.2018.00044.

Daniel Roth, David Mal, Christian Felix Purps, Peter Kullmann, and Marc Erich Latoschik. Injecting Nonverbal Mimicry with Hybrid Avatar-Agent Technologies: A Naïve Approach. In *Proceedings of the Symposium on Spatial User Interaction*, pages 69–73, 2018b. doi:10.1145/3267782.3267791.

Daniel Roth, Gary Bente, Peter Kullmann, David Mal, Chris Felix Purps, Kai Vogeley, and Marc Erich Latoschik. Technologies for social augmentations in user-embodied virtual reality. *Proceedings of the ACM Symposium on Virtual Reality Software and Technology, VRST*, 11 2019a. doi:10.1145/3359996.3364269.

Daniel Roth, Carola Bloch, Josephine Schmitt, Lena Frischlich, Marc Erich Latoschik, and Gary Bente. Perceived Authenticity, Empathy, and Pro-social Intentions evoked through Avatar-mediated Self-disclosures. In *MuC'19: Proceedings of Mensch und Computer 2019*, pages 21–30, 2019b. doi:10.1145/3340764.3340797.

K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum*, 34:299–326, 2015. doi:10.1111/CGF.12603.

J. P. De Ruiter, Holger Mitterer, and N. J. Enfield. Projecting the end of a Speaker's Turn: A Cognitive Cornerstone of Conversation. *Language*, 82:515–535, 2006. doi:10.1353/LAN.2006.0130.

Najmeh Sadoughi, Yang Liu, and Carlos Busso. Meaningful Head Movements Driven by Emotional Synthetic Speech. *Speech Communication*, 95:87–99, 2017. doi:10.1016/j.specom.2017.07.004.

Eva-Lotta Sallnäs. Haptic Feedback Increases Perceived Social Presence. In *Proceedings of the 2010 international conference on Haptics-generating and perceiving tangible sensations: Part II*, pages 178–185, 2010. URL `https://dl.acm.org/doi/abs/10.5555/1893760.1893788`.

David Sander, Didier Grandjean, Susanne Kaiser, Thomas Wehrle, and Klaus R Scherer. Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion. *European Journal of Cognitive Psychology*, 19:470–480, 2007. doi:10.1080/09541440600757426.

Motoaki Sato, Takahisa Uchida, Yuichiro Yoshikawa, Celso M de Melo, Jonathan Gratch, and Kazunori Terada. People negotiate better with emotional human-like virtual agents than android robots. In *12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 89–98, 2024.

Isabel S Schiller, Lukas Aspöck, Carolin Breuer, Jonathan Ehret, and Andrea Bönsch. Hoarseness among university professors and how it can influence students' listening impression: an audio-visual immersive VR study. In *Proceedings of the 1st AUDICTIVE Conference*, pages 134–137, 2023a. doi:10.18154/RWTH-2023-08885.

Isabel S. Schiller, Lukas Aspöck, and Sabine J. Schlittmeier. The impact of a speaker's voice quality on auditory perception and cognition: a behavioral and subjective approach. *Frontiers in Psychology*, 14, 2023b. doi:10.3389/FPSYG.2023.1243249.

Isabel S. Schiller, Carolin Breuer, Lukas Aspöck, Jonathan Ehret, Andrea Bönsch, Torsten W. Kuhlen, Janina Fels, and Sabine J. Schlittmeier. A lecturer's voice quality and its effect on memory, listening effort, and perception in a VR environment. *Scientific Reports*, 14:1–12, 2024. doi:10.1038/s41598-024-63097-6.

Sabine Janina Schlittmeier, Chinthusa Mohanathasan, Isabel Sarah Schiller, and Andreas Liebl. Measuring text comprehension and memory: A comprehensive database for Heard Text Recall (HTR) and Read Text Recall (RTR) paradigms, with optional note-taking and graphical displays. page 7, 2023. doi:10.18154/RWTH-2023-05285.

A. Schmitz. Ein neues digitales Kunstkopfmeßsystem. *Acta Acustica united with Acustica*, 81:416–420, 1995. URL `https://www.ingentaconnect.com/content/dav/aaua/1995/00000081/00000004/art00016`.

Dirk Schröder and Michael Vorländer. RAVEN: A real-time framework for the auralization of interactive virtual environments. In *Forum Acusticum*, pages 1541–1546, 2011. URL `https://www2.users.ak.tu-berlin.de/akgroup/ak_pub/seacen/2011/Schroeder_2011b_P2_RAVEN_A_Real_Time_Framework.pdf`.

Dirk Schröder, Frank Wefers, Sönke Pelzer, Dominik Rausch, Michael Vorländer, and Torsten Kuhlen. Virtual Reality System at RWTH Aachen University. In *Proceedings of the International Symposium on Room Acoustics (ISRA)*, 2010. URL `https://www.vr.rwth-aachen.de/media/papers/VirtualRealitySystem.pdf`.

Marc Schröder, Marcela Charfuelan, Sathish Pammi, and Ingmar Steiner. Open source voice creation toolkit for the MARY TTS Platform. In *12th Annual Conference of the International Speech Communication Association*, pages 3253–3256, 2011. URL `https://inria.hal.science/hal-00661061/`.

Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments*, 10:266–281, 6 2001. doi:10.1162/105474601300343603.

Immo Schuetz, Harun Karimpur, and Katja Fiehler. vexptoolbox: A software toolbox for human behavior studies using the Vizard virtual reality platform. *Behavior Research Methods*, 55:570–582, 2 2023. doi:10.3758/S13428-022-01831-6.

Philipp Schäfer, Pascal Palenda, Lukas Aspöck, and Michael Vorländer. Virtual Acoustics - A real-time auralization framework for scientific research. 2023. doi:10.5281/zenodo.13744554.

Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. Voice in Human–Agent Interaction. *ACM Computing Surveys*, 54:1–43, 2021. doi:10.1145/3386867.

Etienne De Sevin, Sylwia Julia Hyniewska, and Catherine Pelachaud. Influence of personality traits on backchannel selection. In *10th International Conference on Intelligent Virtual Agents (IVA)*, pages 187–193, 2010. doi:10.1007/978-3-642-15892-6_20.

Katrina Sewell, Violet A Brown, Grace Farwell, Maya Rogers Id, Xingyi Zhang Id, and Julia F Strandid. The effects of temporal cues, point-light displays, and faces on speech identification and listening effort. *PLoS ONE*, 18:e0290826., 2023. doi:10.1371/journal.pone.0290826.

Mike Seymour, Lingyao Yuan, Alan Dennis, and Kai Riemer. Have We Crossed the Uncanny Valley? Understanding Affinity, Trustworthiness, and Preference for Realistic Digital Humans in Immersive Environments. *Journal of the Association for Information Systems*, 22: 591–617, 2021. doi:10.17705/1jais.00674.

Noam R. Shabtai, Gottfried Behler, Michael Vorländer, and Stefan Weinzierl. Generation and Analysis of an Acoustic Radiation Pattern Database for Forty-one Musical Instruments. *J. Acoust. Soc. Am.*, 141:1246–1256, 2017. doi:10.1121/1.4976071.

Katharine A Shapcott, Marvin Weigand, Iuliia Glukhova, Martha N Havenith, and Marieke L Schölvinck. DomeVR: A setup for experimental control of an immersive dome virtual environment created with Unreal Engine 4. In *bioRxiv* , 2022. doi:10.1101/2022.04.04.486889.

Ari Shapiro. Building a Character Animation System. In *Motion in Games*, pages 98–109, 2011. doi:10.1007/978-3-642-25090-3_9.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018. doi:10.1109/ICASSP.2018.8461368.

Thomas B Sheridan. Musings on Telepresence and Virtual Presence. *Presence: Teleoperators & Virtual Environments*, 1:120–126, 1992. doi:10.1162/pres.1992.1.1.120.

Mincheol Shin, Stephen W Song, Se Jung Kim, and Frank Biocca. The effects of 3D sound in a 360-degree live concert video on social presence, parasocial interaction, enjoyment, and intent of financial supportive action. *International Journal of Human-Computer Studies*, 126:81–93, 2019. doi:10.1016/j.ijhcs.2019.02.001.

Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to Body Dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7574–7583, 2018. doi:10.1109/CVPR.2018.00790.

John Short, Ederyn Williams, and Bruce Christie. *The social psychology of telecommunications.* Wiley, 1976.

Ludwig Sidenmark and Hans Gellersen. Eye, Head and Torso Coordination During Gaze Shifts in Virtual Reality. *ACM Trans. Comput.-Hum. Interact*, 27:4:1–40, 2019. doi:10.1145/3361218.

E. H. Siegel, J. Wei, A. Gomes, M. Oliviera, P. Sundaramoorthy, K. Smathers, M. Vankipuram, S. Ghosh, H Horii, J. Bailenson, and R. Ballagas. HP Omnicept cognitive load database (HPO-CLD)–developing a multimodal inference engine for detecting real-time mental workload in VR. In *Technical Report*. HP Labs, 2021. URL `https://developers.hp.com/omnicept/omnicept-open-data-set-abstract`.

Paul Skalski and Ron Tamborini. The Role of Social Presence in Interactive Agent-Based Persuasion. *Media Psychology*, 10:385–413, 2007. doi:10.1080/15213260701533102.

Gabriel Skantze. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech and Language*, 67, 2021. doi:10.1016/J.CSL.2020.101178.

Richard Skarbez, Frederick P. Brooks, and Mary C. Whitton. A Survey of Presence and Related Concepts. *ACM Computing Surveys*, 50:96:1–39, 2017. doi:10.1145/3134301.

Mel Slater. How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence*, 13:484–493, 2004. doi:10.1162/1054746041944849.

Mel Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364: 3549–3557, 2009. doi:10.1098/RSTB.2009.0138.

Mel Slater, Bernhard Spanlang, and David Corominas. Simulating Virtual Environments within Virtual Environments as the Basis for a Psychophysics of Presence. *ACM Transactions on Graphics (TOG)*, 29, 7 2010. doi:10.1145/1778765.1778829.

Mel Slater, Domna Banakou, Alejandro Beacco, Jaime Gallego, Francisco Macia-Varela, and Ramon Oliva. A Separate Reality: An Update on Place Illusion and Plausibility in Virtual Reality. *Frontiers in Virtual Reality*, 3, 2022. doi:10.3389/FRVIR.2022.914392.

Harrison Jesse Smith and Michael Neff. Understanding the impact of animated gesture performance on personality perceptions. *ACM Transactions on Graphics*, 36:1–12, 2017. doi:10.1145/3072959.3073697.

Harrison Jesse Smith and Michael Neff. Communication Behavior in Embodied Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 289, 2018. doi:10.1145/3173574.3173863.

Alec Solway, Jonathan F. Miller, and Michael J. Kahana. PandaEPL: A library for programming spatial navigation experiments. *Behavior Research Methods*, 45:1293–1312, 12 2013. doi:10.3758/S13428-013-0322-5.

Yang Song, Jingwen Zhu, Xiaolong Wang, and Hairong Qi. Talking Face Generation by Conditional Recurrent Adversarial Network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019. doi:https://dl.acm.org/doi/abs/10.5555/3367032.3367163.

Bernhard Spanlang, Jean-Marie Normand, David Borland, Konstantina Kilteni, Elias Giannopoulos, Ausiàs Pomés, Mar Gonzáez-Franco, Daniel Perez-Marcos, Jorge Arroyo-Palacios, Xavi Navarro Muncunill, and Mel Slater. How to Build an Embodiment Lab:

Achieving Body Representation Illusions in Virtual Reality. *Frontiers in Robotics and AI*, 1, 2014. doi:10.3389/frobt.2014.00009.

Charles Spence. Audiovisual multisensory integration. *Acoustical Science and Technology*, 28: 61–70, 2007. doi:10.1250/ast.28.61.

Michael J. Starrett, Andrew S. McAvan, Derek J. Huffman, Jared D. Stokes, Colin T. Kyle, Dana N. Smuda, Branden S. Kolarik, Jason Laczko, and Arne D. Ekstrom. Landmarks: A solution for spatial navigation and memory experiments in virtual reality. *Behavior Research Methods*, 53:1046–1059, 6 2021. doi:10.3758/S13428-020-01481-6.

Mark A. Steadman, Chungeun Kim, Jean-Hugues Lestang, Dan F. M. Goodman, and Lorenzo Picinali. Short-term effects of sound localization training in virtual reality. *Scientific Reports*, 9, 2019. doi:10.1038/s41598-019-54811-w.

Anthony Steed, Lisa Izzouzi, Klara Brandstätter, Sebastian Friston, Ben Congdon, Otto Olkkonen, Daniele Giunchi, Nels Numan, and David Swapp. Ubiq-exp: A toolkit to build and run remote and distributed mixed reality experiments. *Frontiers in Virtual Reality*, 3, 10 2022. doi:10.3389/frvir.2022.912078.

Radosław Sterna and Katja Zibrek. Psychology in Virtual Reality: Toward a Validated Measure of Social Presence. *Frontiers in Psychology*, 12:705448, 2021. doi:10.3389/fpsyg.2021.705448.

Radosław Sterna, Artur Cybulski, Magdalena Igras-Cybulska, Joanna Pilarczyk, Natalia Segiet, and Michał Kuniecki. How Behavioral, Photographic, and Interactional Realism Influence the Sense of Co-Presence in VR. An Investigation with Psychophysiological Measurement. *International Journal of Human-Computer Interaction*, 2023. doi:10.1080/10447318.2023.2285641.

Catherine J Stevens, Bronwyn Pinchbeck, Trent Lewis, Martin Luerssen, Darius Pfitzner, David M W Powers, Arman Abrahamyan, Yvonne Leung, and Guillaume Gibert. Mimicry and expressiveness of an ECA in human-agent interaction: familiarity breeds content! *Computational cognitive science*, 2:1, 2016. doi:10.1186/s40469-016-0008-2.

Paweł M. Strojny, Natalia Dużmańska-Misiarczyk, Natalia Lipp, and Agnieszka Strojny. Moderators of Social Facilitation Effect in Virtual Reality: Co-presence and Realism of Virtual Agents. *Frontiers in Psychology*, 11:503209, 6 2020. doi:10.3389/FPSYG.2020.01252.

Stef Van Der Struijk, Hung-Hsuan Huang, Maryam Sadat Mirzaei, and Toyoaki Nishida. FACSvatar: An Open Source Modular Framework for Real-Time FACS based Facial Animation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 159–164, 2018. doi:10.1145/3267851.3267918.

Ningyuan Sun and Jean Botev. Technological Immersion and Delegation to Virtual Agents. *Multimodal Technologies and Interaction 2023, Vol. 7, Page 106*, 7:106, 11 2023. doi:10.3390/MTI7110106.

Supasorn Suwajanakorn, StevenM. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Transactions on Graphics*, 36:95, 2017. doi:10.1145/3072959.3073640.

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio G. Rodriguez, Jessica Hodgins, and Iain Matthews. A Deep Learning Approach for Generalized Speech Animation. *ACM T. Graphic.*, 36:1–11, 2017. doi:10.1145/3072959.3073699.

Marcus Thiebaux, Stacy Marsella, Andrew N. Marshall, and Marcelo Kallmann. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, pages 151–158, 2008. URL `http://dl.acm.org/citation.cfm?id=1402409`.

Sean Thomas, Ylva Ferstl, Rachel McDonnell, and Cathy Ennis. Investigating how speech and animation realism influence the perceived personality of virtual characters and agents . In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 11–20, 2022. doi:10.1109/VR51125.2022.00018.

Raphaël Thézé, Mehdi Ali Gadiri, Louis Albert, Antoine Provost, Anne Lise Giraud, and Pierre Mégevand. Animated virtual characters to explore audio-visual speech in controlled and naturalistic environments. *Scientific Reports 2020 10:1*, 10:1–12, 9 2020. doi:10.1038/s41598-020-72375-y.

Kshitij Tiwari, Ville Kyrki, Allen Cheung, and Naohide Yamamoto. DeFINE: Delayed feedback-based immersive navigation environment for studying goal-directed human navigation. *Behavior Research Methods*, 53:2668–2688, 12 2021. doi:10.3758/s13428-021-01586-6.

Ilaria Torre, Emma Carrigan, Rachel McDonnell, Katarina Domijan, Killian McCabe, and Naomi Harte. The effect of multimodal emotional expression and agent appearance on trust in human-agent interaction. *Proceedings - MIG 2019: ACM Conference on Motion, Interaction, and Games*, 2019. doi:10.1145/3359566.3360065.

Ilaria Torre, Emma Carrigan, Katarina Domijan, Rachel McDonnel, and Naomi Harte. The effect of audio-visual smiles on social influence in a cooperative human–agent interaction task. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28, 12 2021. doi:10.1145/3469232.

Tomas Trescak, Anton Bogdanovych, and Simeon Simoff. Populating virtual cities with diverse physiology driven crowds of intelligent agents. In *Social Simulation Conference*, 2014. URL `https://ddd.uab.cat/record/127935`.

Xoana G. Troncoso, Stephen L. Macknik, and Susana Martinez-Conde. Microsaccades counteract perceptual filling-in. *Journal of Vision*, 8:15–15, 10 2008. doi:10.1167/8.14.15.

Laura C. Trutoiu, Elizabeth J. Carter, Iain Matthews, and Jessica K. Hodgins. Modeling and animating eye blinks. *ACM Transactions on Applied Perception*, 8:17:1–17, 2011. doi:10.1145/2010325.2010327.

Chih-Hsiung Tu and Marina McIsaac. The Relationship of Social Presence and Interaction in Online Classes. *American Journal of Distance Education*, 16:131–150, 2002. doi:10.1207/S15389286AJDE1603_2.

Fabian Unruh, David Vogel, Maximilian Landeck, Jean-Luc Lugrin, and Marc Erich Latoschik. Body and Time: Virtual Embodiment and its Effect on Time Perception. *IEEE Transactions on Visualization and Computer Graphics*, 29:2626–2636, 2023. doi:10.1109/TVCG.2023.3247040.

Jakob T. Valvoda, Torsten W. Kuhlen, and Christian H. Bischof. Influence of exocentric avatars on the sensation of presence in room-mounted virtual environments. In *Proceedings of the 10th Annual International Workshop on Presence*, pages 59–69, 2007. URL `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e15c1274918bf19e00fc0f6f38abd7b106546c2a`.

Madis Vasser, Markus Kängsepp, Murad Magomedkerimov, Kälver Kilvits, Vladislav Stafinjak, Taavi Kivisik, Raul Vicente, and Jaan Aru. VREX: An open-source toolbox for creating 3D virtual reality experiments. *BMC Psychology*, 5:1–8, 2 2017. doi:10.1186/S40359-017-0173-4.

Julie Vercelloni, Jon Peppinck, Edgar Santos-Fernandez, Miles McBain, Grace Heron, Tanya Dodgen, Erin E. Peterson, and Kerrie Mengersen. Connecting Virtual Reality and Ecology: A New Tool to Run Seamless Immersive Experiments in R. *PeerJ Computer Science*, 7: 1–14, 6 2021. doi:10.7717/PEERJ-CS.544.

Roel Vertegaal, Robert Slagter, Gerrit Van Der Veer, and Anton Nijholt. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. *Conference on Human Factors in Computing Systems - Proceedings*, pages 301–308, 2001. doi:10.1145/365024.365119.

Michelle C. Vigeant, Lily M. Wang, and Jens H. Rindel. Objective and Subjective Evaluations of the Multi-channel Auralization Technique as Applied to Solo Instruments. *Appl. Acoust.*, 72:311–323, 2011. doi:10.1016/j.apacoust.2010.10.004.

Torben Volkmann, Daniel Wessel, Nicole Jochems, and Thomas Franke. German Translation of the Multimodal Presence Scale. In *Mensch und Computer*, pages 475–479, 2018. doi:10.18420/muc2018-mci-0428.

Astrid M. Rosenthal von der Pütten, Carolin Straßmann, and Nicole C. Krämer. Robots or Agents-Neither Helps You More or Less During Second Language Acquisition. In *International Conference on Intelligent Virtual Agents (IVA)*, pages 256–268, 2016. doi:10.1007/978-3-319-47665-0_23.

Michael Vorländer. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality.* Springer Berlin Heidelberg, 2008. doi:10.1007/978-3-540-48830-9.

Hendric Voß and Stefan Kopp. Augmented Co-Speech Gesture Generation: Including Form and Meaning Features to Guide Learning-Based Gesture Synthesis. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023. doi:10.1145/3570945.3607337.

Nadine Wagener, Mareike Stamer, Johannes Schöning, and Johannes Tümler. Investigating Effects and User Preferences of Extra- and Intradiegetic Virtual Reality Questionnaires. In *26th ACM Symposium on Virtual Reality Software and Technology (VRST '20)*, 2020. doi:10.1145/3385956.3418972.

Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014. doi:10.1016/j.specom.2013.09.008.

Thomas Waltemate, Irene Senna, Felix Hülsmann, Marieke Rohde, Stefan Kopp, Marc Ernst, and Mario Botsch. The impact of latency on perceptual judgments and motor performance in closed-loop interaction in virtual reality. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, pages 27–35, 2016. doi:10.1145/2993369.2993381.

Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE Transactions on Visualization and Computer Graphics*, 24: 1643–1652, 2018. doi:10.1109/TVCG.2018.2794629.

Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. Progress in Outlier Detection Techniques: A Survey. *IEEE Access*, 7:107964–108000, 2019a. doi:10.1109/ACCESS.2019.2932769.

Isaac Wang and Jaime Ruiz. Examining the Use of Nonverbal Communication in Virtual Agents. *International Journal of Human–Computer Interaction*, pages 1–26, 2021. doi:10.1080/10447318.2021.1898851.

Isaac Wang, Jesse Smith, and Jaime Ruiz. Exploring Virtual Agents for Augmented Reality. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*. ACM, 2019b. doi:10.1145/3290605.3300511.

Ning Wang and Jonathan Gratch. Don't just stare at me! In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1241–1250, 2010. doi:10.1145/1753326.1753513.

Yuxuan Wang, Rj Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A Saurous. Tacotron: Towards End-to-End Speech Synthesis. *arXiv preprint arXiv:1703.10135*, 2017. doi:10.48550/arXiv.1703.10135.

Zhiyang Wang, Jina Lee, and Stacy Marsella. Multi-party, multi-role comprehensive listening behavior. In *Autonomous Agents and Multi-Agent Systems*, volume 27, pages 218–234, 2013. doi:10.1007/s10458-012-9215-8.

Marcus R. Watson, Benjamin Voloh, Christopher Thomas, Asif Hasan, and Thilo Womelsdorf. USE: An integrative suite for temporally-precise psychophysical experiments in virtual environments for human, nonhuman, and artificially intelligent agents. *Journal of Neuroscience Methods*, 326:108374, 10 2019. doi:10.1016/j.jneumeth.2019.108374.

Theresa F. Wechsler, Andreas Mühlberger, and Franziska Kümpers. 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, 2019. doi:10.3389/fpsyg.2019.01758.

Stefan Weinzierl, Michael Vorländer, Gottfried Behler, Fabian Brinkmann, Henrik von Coler, Erik Detzner, Johannes Krämer, Alexander Lindau, Martin Pollow, Frank Schulz, and Noam R. Shabtai. A Database of Anechoic Microphone Array Measurements of Musical Instruments. 2017. doi:10.14279/depositonce-5861.2.

Jonathan Wendt, Benjamin Weyers, Andrea Bönsch, Jonas Stienen, Tom Vierjahn, Michael Vorländer, and Torsten W. Kuhlen. Does the Directivity of a Virtual Agent's Speech Influence the Perceived Social Presence? In *Virtual Humans and Crowds for Immersive Environments (VHCIE), IEEE*, 2018.

Jonathan Wendt, Benjamin Weyers, Jonas Stienen, Andrea Bönsch, Michael Vorländer, and Torsten W. Kuhlen. Influence of Directivity on the Perception of Embodied Conversational Agents' Speech. In *Proc. Int. Conf. Intell. Virtual Agents*, pages 130–132. ACM, 2019. doi:10.1145/3308532.3329434.

Mark West, Rebecca Kraut, and Han Ei Chew. I'd blush if I could: closing gender divides in digital skills through education. 2019. doi:10.54675/RAPC9356.

Robert Whelan. Effective Analysis of Reaction Time Data. *The Psychological Record*, 58: 475–482, 5 2008. doi:10.1007/BF03395630.

Markus Wieland, Michael Sedlmair, and Tonja-Katrin Machulla. VR, Gaze, and Visual Impairment: An Exploratory Study of the Perception of Eye Contact across different Sensory Modalities for People with Visual Impairments in Virtual Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6. ACM, 2023. doi:10.1145/3544549.3585726.

Shawn M. Willett, Sarah K. Maenner, and J. Patrick Mayo. The perceptual consequences and neurophysiology of eye blinks. *Frontiers in Systems Neuroscience*, 17:1242654, 2023. doi:10.3389/fnsys.2023.1242654.

Alexander Winkler, Jungdam Won, and Yuting Ye. QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars. In *SIGGRAPH Asia*, 2022. doi:10.1145/3550469.3555411.

Bob G. Witmer and Michael J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, 7:225–240, 1998. doi:10.1162/105474698565686.

Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *IEEE Transactions on Human-Machine Systems*, 52:379–389, 6 2022. doi:10.1109/THMS.2022.3149173.

Pieter Wolfert, Gustav Eje Henter, and Tony Belpaeme. Exploring the Effectiveness of Evaluation Practices for Computer-Generated Nonverbal Behaviour. *Applied Sciences*, 14:1460, 2024. doi:10.3390/APP14041460.

Yanxiang Wu, Sabarish V. Babu, Rowan Armstrong, Jeffrey W. Bertrand, Jun Luo, Tania Roy, Shaundra B. Daily, Lauren Cairco Dukes, Larry F. Hodges, and Tracy Fasolino. Effects of virtual human animation on emotion contagion in simulated inter-personal experiences. *IEEE Transactions on Visualization and Computer Graphics*, 20:626–635, 2014. doi:10.1109/TVCG.2014.19.

Matthias Wölfel, Daniel Hepperle, Christian Felix Purps, Jonas Deuchler, and Wladimir Hettmann. Entering a new Dimension in Virtual Reality Research: An Overview of Existing Toolkits, their Features and Challenges. In *Proceedings - 2021 International Conference on Cyberworlds, CW 2021*, pages 180–187, 2021. doi:10.1109/CW52790.2021.00038.

Ioannis Xenakis, Damianos Gavalas, Vlasios Kasapakis, Elena Dzardanova, and Spyros Vosinakis. Nonverbal Communication in Immersive Virtual Reality through the Lens of Presence: A Critical Review. *PRESENCE: Virtual and Augmented Reality*, pages 1–41, 8 2023. doi:10.1162/PRES_A_00387.

Yuyu Xu, Andrew W Feng, Stacy Marsella, and Ari Shapiro. A Practical and Configurable Lip Sync Method for Games. In *Proceedings of Motion on Games*, pages 131–140, 2013. doi:10.1145/2522628.2522904.

Yuyu Xu, Catherine Pelachaud, and Stacy Marsella. Compound Gesture Generation: A Model Based on Ideational Units. In *International Conference on Intelligent Virtual Agents*, pages 477–491, 2014. doi:10.1007/978-3-319-09767-1_58.

Amal Yassien, Passant ElAgroudy, Elhassan Makled, and Slim Abdennadher. A Design Space for Social Presence in VR. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–12. ACM, 2020. doi:10.1145/3419249.3420112.

Nick Yee and Jeremy Bailenson. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Communication Research*, 33:271–290, 2007. doi:10.1111/j.1468-2958.2007.00299.x.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. TransPose: Real-time 3D Human Translation and Pose Estimation with Six Inertial Sensors. *ACM Transactions on Graphics*, 40:1–13, 2021. doi:10.1145/3450626.3459786.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics*, 39:222:1–16, 11 2020. doi:10.1145/3414685.3417838.

Margaret Zellers, David House, and Simon Alexanderson. Prosody and hand gesture at turn boundaries in Swedish. In *Speech Prosody*, pages 831–835, 2016. doi:10.21437/SpeechProsody.2016-170.

Huiyu Zhou and Huosheng Hu. Human motion tracking for rehabilitation—A survey. *Biomedical Signal Processing and Control*, 3:1–18, 1 2008. doi:10.1016/J.BSPC.2007.09.001.

Yang Zhou, Zhan Xu, Chris Landreth, Avangelos Kalogerakis, Subhransu Maji, and Karan Singh. VisemeNet: Audio-Driven Animator-Centric Speech Animation. *ACM T. Graphic.*, 37, 2018. doi:10.1145/3197517.3201292.

Katja Zibrek and Rachel McDonnell. Social presence and place illusion are affected by photorealism in embodied VR. In *MIG '19: Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–7, 2019. doi:10.1145/3359566.3360064.

Katja Zibrek, Elena Kokkinara, and Rachel McDonnell. Don't Stand So Close To me: Investigating the effect of control on the appeal of virtual humans using immersion and a proximity-based behavioral task. In *Proceedings of the ACM Symposium on Applied Perception*, 2017. doi:10.1145/3119881.3119887.

Katja Zibrek, Elena Kokkinara, and Rachel McDonnell. The Effect of Realistic Appearance of Virtual Characters in Immersive Environments - Does the Character's Personality Play a Role? *IEEE Transactions on Visualization and Computer Graphics*, 24:1681–1690, 2018. doi:10.1109/TVCG.2018.2794638.

Katja Zibrek, João P Cabral, and Rachel McDonnell. Does Synthetic Voice alter Social Response to a Photorealistic Character in Virtual Reality? In *MIG '21: Proceedings of the 14th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2021. doi:10.1145/3487983.3488296.