Aachener Beiträge zur Hörtechnik und Akustik

Cosima Antonia Ermert

Simulating complex listening scenarios in audiovisual virtual reality

The impact of audiovisual cues, spatial noise, and room acoustics on listening performance

IHTA Institute for Hearing Technology and Acoustics
RWTH AACHEN UNIVERSITY

Simulating complex listening scenarios in audiovisual virtual reality

The impact of audiovisual cues, spatial noise, and room acoustics on listening performance

Von der Fakultät für Elektrotechnik und Informationstechnik der
Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades einer

DOKTORIN DER INGENIEURWISSENSCHAFTEN

genehmigte Dissertation

vorgelegt von

M.Sc.
Cosima Antonia Ermert

aus Aachen, Deutschland

Berichter:

Universitätsprofessorin Dr.-Ing. Janina Fels
Professor (apl.) Dr. phil. Stephan Getzmann

Tag der mündlichen Prüfung: 10. Juli 2026

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

You never know what is going to happen.
Lemony Snicket

Für Marius, Mama und Papa.

Abstract

Real-life listening is inherently complex; scenarios often contain multiple partially masking sound sources, varying spatial configurations, and room acoustic reflections. Moreover, listeners typically rely not only on auditory but also on visual cues, such as lip movements, gestures, gaze, and embodiment.

Laboratory listening experiments often fail to capture this complexity. For example, many studies employ simplified environments when examining listening performance, which may result in findings that do not generalise to real-life listening. Recent technological advances enable the simulation of complex auditory and visual environments, e.g., through virtual reality (VR). By utilising these techniques, listening experiments can be designed to imitate real-life listening more closely and to increase complexity. Consequently, the primary objective of this thesis was to assess how increasing complexity influences listening performance, conceptualised in terms of working memory and listening effort.

In seven experiments, reported in five publications, systematic increments in auditory and audiovisual complexity were evaluated. More precisely, the experiments introduced spatial audio reproduction and spatial target and background position changes, varied the semantic content and temporal structure of background noise, examined the influence of display device and audiovisual incongruences in virtual environments with embodied conversational agents (ECAs), and finally investigated the influence of room acoustic conditions in a simulated audiovisually complex environment. Furthermore, throughout the thesis, task complexity was increased from the classical serial recall task with isolated digits as target items towards a joint analysis of memory and listening effort by employing the heard-text recall (HTR) task in a dual-task design.

The findings suggest a task-dependent relevance of these complexity increments and highlight the interdependence of environment, target source, task, and receiver. In conclusion, the thesis proposes a framework of audiovisual complexity that extends existing acoustic models by explicitly incorporating visual, receiver-related, and technical dimensions. In addition, a conceptual distinction between descriptive and generative approaches to complexity is introduced, which might facilitate more differentiated discussions and interpretations of future findings. In this context, this thesis highlights that current simulation-based approaches

and experimental paradigms may not capture real-life listening to a sufficient degree, and argues for a more holistic perspective on complex listening scenarios that incorporates behaviour- and experience-based dependent variables.

Together, these contributions aim to support the design and interpretation of future listening experiments in complex audiovisual environments and to move towards models that are better suited to predict everyday listening. In this way, listening experience can ultimately be improved, particularly for vulnerable listener groups, such as people with hearing impairment.

Contents

Glossary **xi**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Descriptors of Virtual Audiovisual Scenarios . . . 3
1.2.1 Realism and Authenticity . . . 3
1.2.2 Plausibility . . . 4
1.2.3 Immersion and Presence . . . 5
1.2.4 Internal and External Validity . . . 7
1.2.5 Ecological Validity . . . 7
1.2.6 Complexity . . . 9
1.3 Source Complexity . . . 10
1.3.1 Technical Background . . . 10
1.3.2 State of the Art . . . 12
1.4 Environment Complexity . . . 14
1.4.1 Technical Background . . . 14
1.4.2 State of the Art . . . 15
1.5 Source-Environment Interactions . . . 16
1.6 Visual Complexity . . . 17
1.6.1 Technical Background . . . 17
1.6.2 State of the Art . . . 18
1.7 Receiver's Task Complexity . . . 26
1.7.1 Working Memory . . . 27
1.7.2 Listening Effort . . . 32
1.8 Thesis Outline and Objectives . . . 36

2 Spatial target source alternations in the serial recall **39**
2.1 Abstract . . . 39
2.2 Introduction . . . 40
2.3 Experiment 1 . . . 43
2.3.1 Method . . . 44
2.3.2 Results . . . 49

2.3.3 Discussion 50
2.4 Experiment 2 53
2.4.1 Method 54
2.4.2 Results 56
2.4.3 Discussion 56
2.5 Limitations 58
2.6 Conclusion 58

3 Spatial-changing irrelevant speech in visuo-spatial and visuo-verbal serial recall 61
3.1 Introduction 62
3.1.1 Present Study 65
3.2 Materials and Methods 66
3.2.1 Participants 66
3.2.2 Serial Order Reconstruction Tasks 68
3.2.3 Irrelevant Sound Sequences 70
3.2.4 Procedure 71
3.2.5 Data Analysis 72
3.3 Results 73
3.3.1 Visual-Verbal Task 73
3.3.2 Visuo-Spatial Task 74
3.4 Discussion 75
3.5 Conclusion 79

4 Audiovisual angle and voice incongruence 81
4.1 Abstract 81
4.2 Introduction 82
4.3 Experiment 1: Audiovisual Spatial Incongruence 84
4.3.1 Method 84
4.3.2 Results 93
4.3.3 Discussion 95
4.4 Experiment 2: Audiovisual Semantic Incongruence 97
4.4.1 Method 98
4.4.2 Results 100
4.4.3 Discussion 103
4.5 Limitations 104
4.6 Conclusion 104

5 Heard-text recall and listening effort under irrelevant speech and pseudo-speech in virtual reality 107
5.1 Abstract . . . 107
5.2 Introduction . . . 108
5.3 Method . . . 111
5.3.1 Participants . . . 111
5.3.2 Dual-Task Paradigm . . . 111
5.3.3 Masker Conditions . . . 112
5.3.4 Audio-Visual Virtual Reality Setup . . . 113
5.3.5 Procedure . . . 114
5.3.6 Data Analysis . . . 115
5.4 Results . . . 116
5.5 Discussion . . . 117
5.6 Limitations . . . 120
5.7 Conclusion . . . 121

6 Listening Effort In Complex Audiovisual Scenes Under Varying Simulated Room Acoustic Conditions 125
6.1 Abstract . . . 125
6.2 Introduction . . . 126
6.2.1 Present Study . . . 130
6.3 Methods . . . 131
6.3.1 Participants . . . 131
6.3.2 Dual-Task Paradigm . . . 132
6.3.3 Audiovisual Environment . . . 133
6.3.4 Procedure . . . 137
6.3.5 Data Analysis . . . 138
6.4 Results . . . 139
6.4.1 Primary Task Performance . . . 139
6.4.2 Secondary Task Performance . . . 141
6.4.3 Secondary Task Response Times . . . 141
6.5 Discussion . . . 143
6.6 Limitations . . . 146
6.7 Conclusion . . . 147

7 Discussion and Conclusion 151
7.1 Acoustic Complexity Framework . . . 151
7.1.1 Source Complexity: Multiple Sources . . . 152
7.1.2 Source Complexity: Source Diversity . . . 154
7.1.3 Source Complexity: Source-Source Interactions . . . 154

7.1.4 Environmental Complexity 155
7.1.5 Visual Complexity 155
7.1.6 Receiver's Task Complexity 156
7.2 The Role of Visual Cues in Listening Research 158
7.2.1 Visual Cues as a Subcomponent of Auditory Information 158
7.2.2 Visual Cues in Relation to other Factors of Complexity . 158
7.3 Audiovisual Complexity Framework 160
7.4 Descriptive versus Generative Complexity 163
7.5 Conclusion and Outlook 165

Appendices **169**

A Copyright Clearances **171**
A.1 Chapter 2 171
A.2 Chapter 3 172
A.3 Chapter 4 172
A.4 Chapter 5 173
A.5 Chapter 6 174

Bibliography **177**

Curriculum Vitae **221**

Publications **223**

Acknowledgements **227**

Aachener Beiträge zur Hörtechnik und Akustik **229**

Glossary

Symbols

C_{50}	Clarity index for speech	dB
M	Arithmetic mean	
SD	Standard Deviation	
T_{30}	Reverberation time estimate based on a on the sound pressure level (SPL) decay from $-5\,\mathrm{dB}$ to $-35\,\mathrm{dB}$	s
f	Frequency	Hz

Acronyms

ACALES	Adaptive Categorical Listening Effort Scaling
ARTE	ambisonic recordings of typical environments
avVSR	audiovisual verbal serial recall
CAVE	Cave Automatic Virtual Environment
CI	credible interval
CoSS	common sound scenarios
ECA	embodied conversational agent
elpd	expected log pointwise predictive density
ELU	Ease of Language Understanding
EMA	Ecological Momentary Assessment
FUEL	Framework for Understanding Effortful Listening
GLMM	generalized linear mixed model
HDI	highest density interval
HMD	head-mounted display
HRTF	head-related transfer function
HTR	heard-text recall
IHTA	Institute for Hearing Technology and Acoustics
ISE	irrelevant sound effect
ISI	interstimulus interval

LOO	leave-one-out
MARC	Multidimensional Assessment of Research in Context
MoLE	Model for Listening Engagement
NASA-TLX	NASA Task Load Index
O-OER	object-oriented episodic record
OLSA	Oldenburger sentence test
PD	probability of direction
pDTC	proportional dual-task cost
PPD	posterior probability distribution
RAVEN	Room Acoustics for Virtual ENvironments
RIR	room impulse response
ROPE	region of practical equivalence
RT	reverberation time
SCaLAr	Spherical Cap Loudspeaker Array
SD	standard deviation
SI	speech intelligibility
SNR	signal-to-noise ratio
SPL	sound pressure level
SRM	spatial release from masking
SRT	speech reception threshold
SSAE	spatial-source alternation effect
SSQ	Speech, Spatial and Qualities of Hearing Scale
STI	speech transmission index
SUS	Slater, Usoh, and Steed presence questionnaire
TVE	talker-variability effect
VA	Virtual Acoustics
VR	virtual reality

1

Introduction

1.1 Motivation

Real-life listening is inherently complex. In everyday scenarios, listeners rarely encounter clear speech signals in anechoic conditions. Instead, they are exposed to complex acoustic scenarios in which multiple sound sources overlap in time, and background sound[1] and room acoustic reflections are present. In his classic description of the "cocktail party problem", Cherry [85] also lists spatial information, voice pitch, speech rate, accents, as well as semantic and prosodic information as important cues for navigating complex acoustic scenarios. Much of the research on listening performance has simplified this complexity. Experimental paradigms in listening experiments often use short, isolated stimuli in quiet or together with stationary maskers, and they frequently focus on non-spatial presentation. While such investigations have been helpful in establishing fundamental mechanisms of listening performance, they only provide a limited insight into how listeners operate in realistic listening scenarios. Recently, attempts have been made to more systematically characterise the complexity of auditory scenarios. For example, Weisser et al. [484] proposed a framework that decomposes the complexity of auditory scenarios into four dimensions, such as properties of the sound sources, the acoustic environment, their interaction, and the participant's task (see Section 1.2.6). This framework highlights that complexity is not a straightforward attribute of a scene but depends on multifaceted factors.

Further, many studies neglect the additional influence of the visual domain. Visual cues can help listeners segregate and follow a target talker in multi-talker

[1] In this thesis, the term *background* (e.g., background sounds, background stimuli, background speech) is used as an umbrella term for all sounds or stimuli other than the task-relevant target stimuli / target speech.

situations and have been shown to improve speech intelligibility, particularly in the presence of background sound ([260] for a review). Listening in real-world scenarios requires the simultaneous processing and integration of acoustic and visual cues, as well as the flexible allocation of attention to the relevant sources. Consequently, listening performance in complex scenarios cannot be predicted from auditory conditions alone [165]. Despite the considerable progress in understanding audio-only speech perception, relatively little research focusses on the relevance of visual cues, particularly in complex scenarios [165, 333]. For example, both Cherry [85] and Weisser et al. [484] note visual cues as a relevant aspect of scenario complexity, but they still occupy a secondary position compared to auditory cues in both models and also in studies on listening performance.

Thus, when examining listening performance in complex audiovisual scenarios, both auditory and visual components must be taken into account. Systematic variations of isolated components may help shed a light on how to appropriately simulate scenarios for research. This motivates the present thesis: based on the complexity framework proposed by Weisser et al. [484], this thesis extends purely auditory to complex audiovisual scenarios.

In this thesis, a focus is set on virtual scenarios, i.e., non-physical scenarios that are reproduced auditorily, visually, or audiovisually by technical systems rather than being encountered directly in the real world. From a technical perspective, virtual environments can either be recorded or simulated in VR[2], e.g., through visualisation and "auralisation" [474] of computer-based audiovisual scenarios. Together, these techniques enable the creation of complex audiovisual virtual scenarios in which, e.g., source characteristics, spatial distributions of the background sound, room acoustics, as well as audiovisual cues, can be controlled and, particularly in simulated scenarios, manipulated independently and systematically. To establish common terms used in the context of virtual scenarios, Section 1.2 defines realism and authenticity, plausibility, immersion, presence, and different aspects of validity (internal, external, and ecological), and introduces the framework of complexity according to Weisser et al. [484]. Building upon this framework, the following sections discuss the state of the art for the different factors of complexity defined by Weisser et al. [484]: source complexity (Section 1.3), scenario complexity (Section 1.4), source–environment interactions (Section 1.5), and visual complexity (Section 1.6). Paradigms to evaluate listening

[2] virtual reality (VR) is widely used in literature and different definitions exist. Some definitions regard both recorded and simulated environments as VR (e.g., [149]), while others understand VR as computer-generated, non-recorded environments (e.g., [217, 496]). In the present thesis, VR scenarios are understood as computer-generated (simulated), while the term *virtual* scenarios is used to describe both video- and simulation-based environments. VR is further understood as audiovisual, if not specified otherwise (e.g., visual VR, auditory VR).

performance, specifically listening effort and memory, and approaches to increase their complexity are presented in the section on task complexity (Section 1.7). Finally, Section 1.8 summarises the objectives and structure of the thesis.

1.2 Descriptors of Virtual Audiovisual Scenarios

Virtual (simulated or pre-recorded) audiovisual scenarios are increasingly used to present complex scenarios in participant studies. In contrast to assessing listening performance in the real-world they allow for reproducibility and experimental control. However, it is unclear how well such experiments generalise beyond the laboratory. When looking back at early VR studies, it becomes clear that the quality of simulated scenarios is defined by advances in VR technology (see, e.g., [83, 496]). Early VR is hardly comparable to today's systems; likewise, what is the current best possible quality may seem dated to future readers, such as contemporary virtual humans[3] and computational power. Likewise, comparisons between laboratories are challenging as many different technologies are employed and systematic contrasts, e.g., in the form of round-robin experiments, are rare. Consequently, results obtained with current technology should be assessed relative to quality constructs, such as realism, authenticity, or plausibility, to enable sustainable research and meaningful comparison across studies.

However, research on virtual audiovisual scenarios is inherently interdisciplinary, spanning subjects like acoustics, computer science, and psychology, leading to inconsistent terminology. Therefore, in this section, common terms of (perceptual) quality are introduced and influencing factors as well as common measurement approaches are listed.

1.2.1 Realism and Authenticity

The ground truth of any consideration is the real-life. Many considerations in virtual scenarios concern themselves with how closely real-life is imitated. A central metric is therefore *realism*, "the degree to which the materials and procedures involved in an experiment are similar to events that occur in the real world" [329, p. 599]. Realism often concerns physical or rendering realism, i.e., detailed modelling of scenario geometry and materials, photometry (e.g., illumination, shadows), and acoustics (e.g., room acoustics, spatial audio), as well as accurate reproduction

[3] Virtual humans are here understood as simulated "human-like characters on a computer screen or agents with embodied life-like behaviours" [377, p. 290]. For conversational purposes, sometimes the term embodied conversational agent (ECA), is used, which are "not just interfaces in which human or animal bodies appear lifelike or believable in their actions and their reactions to human users. [They] are specifically conversational in their behaviours and specifically humanlike in the way they use their bodies in conversation." [83, p. 72].

(e.g., via displays and playback devices) [430]. How actors and objects move or respond to collisions or other external influences also plays a role; in the case of virtual humans, this includes animation and general movement patterns (e.g., [45, 112, 114, 116, 115]). It is further important to provide consistency of cues across modalities (e.g., audiovisual, see Section 1.6.2, as well as Chapter 4 [127]).

An attribute commonly used to describe the realism of a given scenario or stimulus is the *authenticity*. In the context of audio reproduction, Blauert [44] defines authenticity as follows: "If the reproduced signals in the ear canals correspond exactly to those during recording, the spatial attributes of the auditory events also correspond exactly to those of the auditory events that the subject had during the recording. An *authentic* spatial reproduction thus takes place." [44, p. 358] Based on this definition, Meyer-Kahlen et al. [311] subdivide authenticity into perceptual and physical authenticity and postulate that if physical authenticity is achieved, e.g., by correctly reproducing the sound, perceptual authenticity follows. They define: "A virtual source is authentic if it is indistinguishable from its real reference" [311, p. 800] (see also [64, p. 1784]). While these definitions already incorporate perception to a certain degree, an alternative concept with a stronger focus on perception is *plausibility*.

1.2.2 Plausibility

Lindau and Weinzierl [274] define acoustic plausibility as "a simulation in agreement with the participant's expectation towards a corresponding real event" [274, p. 1]. They emphasise that an immediate comparison between simulation and reality is often not experimentally possible and, moreover, not required in applications where participants do not have access to any external real-world reference. Instead, they argue that it is sufficient that participants are familiar with the corresponding real-world scenario and use this as the basis of their judgment. In his comprehensive overview about the role of plausibility in acoustics, Kuhn-Rahloff [250] introduces the concept of an internal reference, which is derived from prior experience with real-world object or stimuli rather than from direct comparison [250, p. 99], [311, p. 799]. Many simulated audio(visual) scenarios take place in practically motivated, but theoretical scenarios. Thus, exact real-world references do not exist. Consequently, experiments often focus on plausibility rather than authenticity (e.g., [64, 118, 311, 340]).

In the literature, plausibility is mostly assessed via subjective evaluation (e.g., Likert or continuous sliders [118, 311], or as a binary variable [274, 340]). Plausibility has been shown to depend on a variety of technical and perceptual factors, such as system latency, artefacts, and source localisation [274]. Spatial

reproduction techniques and room acoustic simulations have also been shown to influence plausibility [118, 340].

While plausibility definitions are conceptually coherent in the field of acoustics, the term becomes much more ambiguous when adding a visual component. In visual VR, plausibility, often termed as *plausibility illusion*, is closely related to the concept of presence [431]; which will be discussed in more detail in the following section.

1.2.3 Immersion and Presence

Immersion and presence describe complementary aspects of the participants' experience in virtual scenarios.

Immersion is often related to the presentation system's capabilities to "envelope" the participant [4, 493, for a review]. Murray [330] compares immersion to the feeling of submerging oneself into water (see also [368]): like water, an immersive experience surrounds the participant. Immersion is typically related to the structure and fidelity of the sensory input provided by the system [368, 431, 433]. Multiple articles argue that immersion is directly influenced by how much the system obscures the physical reality [378, 368, 433]. The lower the salience of the physical world, the greater the immersion [4, 169, 378]: visual presentation can be fully immersive (e.g., Cave Automatic Virtual Environment (CAVE) systems [96, 249] or head-mounted displays (HMDs) [60, 115, 399], cf. Chapters 5 [125] and 6 [124]), semi-immersive (e.g., projection screens [352]), and non-immersive (e.g., standard computer monitor [112], cf. Chapter 4 [127]). Haptic cues and naturalistic movement and interaction possibilities can further increase immersion [82, 86].

Quantifying immersion is not a straightforward task. There are approaches employing behavioural and physiological measures like skin impedance, neural imaging, eye movement, and interaction metrics [287, 355, 450, 492]. Questionnaires exist [214, 462], also with regard to the acoustic aspects of the scenario [16, 91], but due to their perception-based and highly subjective nature, they are often not clearly separated from the concept of *presence*.

Presence [98, 493] explicitly incorporates the subjective experience. It has been defined as the "illusion of being there [in the presented scenario]" [432, p. 432], or as the "extent to which a person's cognitive and perceptual systems are tricked into believing they are somewhere other than their physical location" [352, p. 479] (see [4, 493, for a review]). Presence is often subdivided into two contributing illusions: (i) the "place illusion" (the feeling of being in the depicted place) and (ii) the "plausibility illusion" (illusion that the virtual scenario is really happening, even though participants know it is not real) [431]. According to Slater [431] the

plausibility illusion arises when a scenario "provides correlations between external events not directly caused by the participant and his/her own sensations" [431, p. 3553]. An example they give is automatically smiling back at an ECA [83] that smiles at the participant - as the participant would instinctively do if a real person smiled at them. Importantly, Slater [431] emphasises that such a plausibility illusion does not necessitate high realism in a physical sense; rather, it depends on the perceived coherence of events.

Presence has been shown to be influenced by traditionally immersion-related technical factors [216, 272, 331, 348], but also by human-related factors like emotions, arousal, personality traits, and attention [211, 270] or contextual factors like environmental coherence, narrative plausibility, consistency of events, and social behaviour [216, 331].

The subdivision of *presence* into the place and plausibility illusion in VR literature [431] differs from plausibility definitions in acoustic-related literature (see Section 1.2.2). Thus, in interdisciplinary studies, clear mentioning of the assumed definition is important when an experiment is set out to investigate the effect of experimental variations on plausibility (e.g., [39, 208]). Otherwise, interpretation of results is challenging. An example for this is the study by Bergström et al. [39], who aim to examine the plausibility illusion of a VR scenario featuring a virtual string quartet [431] by asking participants to compare scenarios to an explicitly presented reference scenario. It is debatable whether such an approach truly measures plausibility in either of the previously introduced definitions or instead tests for reference-matching abilities. This example highlights how easily different constructs can become confused in practice, and underlines the importance of clear and consistent definitions across disciplines.

A sub-construct of presence is *social presence* or *co-presence*, which is especially relevant for experiments with ECAs. Harms and Biocca [176] define co-presence as "the degree to which the [participant] believes he/she is not alone and secluded, their level of peripheral or focal awareness of the other, and their sense of the degree to which the other is peripherally or focally aware of them." [176, p. 1]. Co-presence is modulated by the ECA's appearance and animation (e.g., with regard to co-verbal and non-verbal signals like gaze behaviour, lip movement, facial expressions), behavioural realism and responsiveness (e.g., turn-taking [114]), and the degree to which the ECA signals mutual awareness [45, 114, 116].

As a subjective construct, presence is commonly evaluated with questionnaires, such as the igroup[4] presence questionnaire [455], the Slater, Usoh, and Steed presence questionnaire (SUS) [462] (which is also employed in this thesis in Chapter 4 [127]), the Witmer–Singer presence questionnaire [495], or the sense

[4] www.igroup.org

of presence inventory [268]. Physiological and behavioural measures are also used to complement self-report [289]. Co-presence is also evaluated through questionnaires (e.g., [369]) and/or behavioural indicators (e.g., interpersonal distance regulation, gaze-following, task behaviour) [46, 47, 114, 116].

1.2.4 Internal and External Validity

From the stimulus-based constructs defined above, this section now turns to implication-based considerations, more specifically to aspects of *validity*. According to Roe and Just, "validity [. . .] is generally concerned with whether a particular conclusion or inference represents a good approximation to the true conclusion or inference" [387, p. 1266]. Validity is most commonly subdivided into *internal* and *external* validity. Internal validity addresses the experiments methods and design. For example, Bortz and Döring state that: "Internal validity refers to the unambiguity with which a research result can be related to the hypothesis in terms of content" [48, p.33] (see also [387, p. 1266] and [451, p. 24]). Typical threats to internal validity are ranging from instrumentation over participant selection to statistic analysis (e.g., [48, 451]). Such risks can be counteracted to a certain degree by careful study design, e.g., by applying randomisation or balancing (see, e.g., [48, 451]).

External validity, on the other hand, concerns the generalisability of findings "to other persons, times, and settings" [387, p. 1267] (see also [48, 451]). External validity is at risk when, amongst others, measures may not capture the intended construct, sampling errors occur, or when participants' behaviour changes merely because they know they are being studied ([48, 451]). External validity can be strengthened by appropriate sampling, multidimensional testing (e.g., in the lab and field, with varying tasks and stimulus sets), or replication studies.

In many laboratory experiments, internal and external validity contradict each other [451, p. 31]. For example, to increase external validity, experimenters may increase the complexity of their experiments as a step towards real-life scenarios, e.g., by using real-world recordings as stimuli. This creates a tension between control, which contributes to internal validity, and generalisability, which contributes to external validity.

1.2.5 Ecological Validity

Another validity construct is *ecological validity*, which has gained substantial prominence in recent years, particularly in hearing-related research. Historically, ecological validity was introduced by Brunswik [75, 76] based on object perception theory; ecological validity is "the reliability of cues received at the sense organs in terms of the correlation between those cues, and objects or events in the

scenario" [31, p. 1396], which shows clear parallels to authenticity definitions. In many definitions, however, ecological validity is related to a different concept introduced by Brunswik [75, 76], which is the concept of "representative design". For representative designs, experimenters should not only sample a variety of participants, but also situations and tasks such that they reflect the variability of the real world [334]. The term ecological validity is thus prevailingly used to address how well the experimental situation and task represent the context in which the behaviour of interest normally occurs (see also [143, 180]).

For example, in their comprehensive review paper on ecological validity in hearing science, Keidser et al. [232] propose the following definition of ecological validity: "the degree to which research findings reflect real-life hearing-related function, activity, or participation". Keidser et al. [232] relate ecological validity to five methodological dimensions: the source of stimuli (content, multimodality, etc.), the scenario (presentation of stimuli), the context of participation (motivation, preparation, etc., of the participant), the task (paradigm, complexity, freedom, etc.), and the individual (personality, health, abilities, etc.).

Ecological validity is often used as a synonym for realism and complexity [232]. This is seen as problematic by Beechey [31], who argue that generalizability of experimental findings, and thus external validity should always be the goal - not realism. According to their view, realism *can* be a tool to improve generalisability, but should not be taken as a proxy for it. Researchers should strive for *representative* rather than *realistic* conditions. This differentiation between realism and ecological validity is supported by a number of papers [6, 191, 334] which assume that a highly realistic experimental set-up can still be ecologically invalid if it does not capture the specific real-world situations or behaviours of interest, whereas a seemingly artificial paradigm may have high ecological validity if it reliably predicts performance in the relevant everyday contexts. These considerations show strong overlap with aspects of external validity, also noted by Beechey [31].

Since ecological validity is not uniformly defined [232], it is also difficult to measure or assess it in practice. With recent technological advancements, VR is often proposed as a tool to increase the ecological validity of studies as the direct comparison of real-life and the laboratory scenario seemingly becomes possible. Many researchers follow the approach of introducing participants to both a scenario in real-life and to a replica of it in VR and compare their behaviour [189, 350, 409, 442], but results may be highly influenced by the perceived quality of the VR system [232].

An established tool to gather data directly in the field (and thus with assumed high ecological validity) is the Ecological Momentary Assessment (EMA) [400]. For example, with a mobile phone application, participants can enter ratings and experiences in a diverse set of real-life situations. At the same time, the phone

can collect and record relevant acoustic parameters. However, this approach has also been called into question due to the high risk of noisy data [376]. For laboratory experiments, ecological validity is often proposed to be evaluated based on the assessment performed by the experimenter. Keidser et al. [232], Naumann et al. [334] and Akeroyd [6] proposed checklists that experimenters can fill out prior to conducting their study. In these lists, a variety of factors, such as the stimuli or the participant sample, have to be rated with regard to their closeness to real-life scenarios [334] or with regard to the risk they pose for the ecological validity of the experiment [6]. In addition to an evaluation of the experiment, these tools also encourage researchers to critically assess their experiment design and potentially refine factors that hinder ecological validity. While these checklist-based approaches promote transparency and structured reflection, they rely on judgment by few people (the experimenters) and are therefore subject to individual biases while offering limited reliability. Such ecological validity assessments should thus be interpreted with caution and, where possible, complemented by data from participants or real-world measurements.

1.2.6 Complexity

A factor which is often mentioned in relation to realism or ecological validity is *complexity*. While no clear definition of complexity in the scope of audiovisual scenarios exists, it is often understood as the opposite of "simple, quiet scenarios" [55, p. 16] and used as a synonym for adverse and challenging situations which approximate real-world conditions [484] without claiming high realism or authenticity. A framework for describing complex *acoustic* scenarios was proposed by Weisser et al. [484]. They divided acoustic complexity into four subtypes with a total of nine characteristics:

1. Source: (1) number and spatial distribution of sound sources, (2) acoustic source diversity, and (3) interactions between sources.

2. Environment: (4) Reverberation, reflections, scattering, diffraction, and diffusion, and (5) non-uniform medium for sound propagation (for example underwater acoustics, temperature gradients).

3. Source–environment interactions: (6) loudspeaker amplification systems (such as public address systems or, e.g., television sets), (7) source-environment adaptation (e.g., Lombard speech), and (8) cues of other modalities (such as visual information)

4. Task: (9) receiver's task.

Weisser et al. [484] proposed a questionnaire to evaluate the complexity of acoustic scenarios. For a given auditory scenario, participants are first asked to identify and categorise all audible sound sources. Secondly, subjective attributes of the scenarios are evaluated both with and without target speech; attributes include familiarity, loudness, busyness, annoyance, and pleasantness. Weisser et al. [484] evaluated the questionnaire in several pre-recorded acoustic scenarios and found that busyness most prominently contributes to perceived complexity. Acoustic complexity thus goes beyond challenging signal-to-noise ratios (SNRs). Although the framework of Weisser et al. [484] was originally developed for acoustic scenarios, it explicitly includes cross-modal cues, such as visual information. Visual information is hereby understood as part of the scenario with which the perception of the acoustic source may interact, for example through audiovisual integration.

In the following, the review of literature on virtual complex audio(visual) scenarios is structured following the four categories proposed by Weisser et al. [484]. As the key interest in the present thesis is extension of audio-only scenarios to audiovisual scenarios, a special focus is set on visual cues as a type of source-environment interactions (characteristic 8: cues of other modalities [484]). Therefore, a separate section reviews the influence of visual complexity (Section 1.6). However, audiovisual experiments may be discussed in sections focussing on auditory-only aspects, such as source or environment complexity, if the respective research question is focussed on the contribution of the auditory information.

1.3 Source Complexity

According to Weisser et al. [484], a major component of acoustic complexity is the nature of the target and the background stimuli (source complexity), which includes spatial sound reproduction, stimulus content, and source-source interactions. The following section provides the technical background on how such stimuli can be auralised.

1.3.1 Technical Background

Human hearing is inherently spatial. For sound localisation and segregating multiple simultaneous sound sources, the auditory system exploits differences of acoustic features between both ears (binaural cues) or spectral variations apparent in the signal of one ear (monaural cues) [44, 54, 275]. Binaural cues include differences in time of arrival, level, and phase between the two ears [44, 275]. Monaural cues are direction-dependent spectral colourations caused by asymmetries in the participants' physiology [44, 275]. Both binaural and monaural

cues arise from the interaction of the incident sound field with the participant's anatomy (pinna, head, torso). Together, they act as a direction-dependent filter, which is commonly described as the head-related transfer function (HRTF) [119, 301].

To reproduce spatial sound in listening experiments, these directional filtering effects must be taken into account. The conceptually simplest way to reproduce sound from a given direction is to place a loudspeaker at that direction and play back the respective signal. In such a setup, the participant's anatomy naturally filters the sound (e.g. [95, 105]). However, this approach is limited: a separate loudspeaker is required for each desired source direction, which quickly becomes infeasible when many directions (e.g., in the case of room acoustic simulations, see Section 1.4), high angular resolution, or moving sources are needed.

Another possibility could be to obtain spatial recordings of real-life scenarios, e.g., with artificial heads or microphone arrays. While such recordings might yield realistic material, they are tied to specific situations, because they are difficult to modify in retrospective, and thus offer only limited flexibility for small, controlled variations in acoustic parameters, which are often of interest for studies. A more flexible alternative is binaural reproduction over headphones using HRTFs. To simulate a sound source at a particular azimuth and elevation, the raw signal is filtered with the left- and right-ear HRTF corresponding to that direction and presented via headphones. If the participant's head movements are tracked, dynamic auralisation frameworks, such as Virtual Acoustics (VA) [209], can update the selected HRTF in real time, maintaining stable sound source positions as the participants turn their heads. While individual HRTFs offer the most accurate spatial cues, their measurement is complex and time-consuming [119, 384]. In many applications, non-individual HRTFs measured on artificial heads (e.g., [408, 53]) or from generic databases provide sufficiently good spatial impression for most participants [341].

Loudspeaker-based spatial audio reproduction extends on simple one-loudspeaker setups by using loudspeakers arrays and signal processing to create virtual acoustic scenarios. Examples include cross-talk cancellation and vector-based amplitude panning [242]. Ambisonics is another technique in which the sound field is encoded in spherical-harmonic functions and decoded for a specific loudspeaker array (or also headphone playback) [242, 506]. With these techniques, raw signals can be rendered at a large variety of positions, which supports source complexity as defined by Weisser et al. [484] by providing flexibility with regard to the number and spatial distribution of sound sources and, depending on the stimulus material, acoustic source diversity and interactions between sources.

All studies presented in this thesis use spatial sound reproduction via VA. Loudspeaker-based reproduction is used in Chapter 3 [126], while Chapters 2 [128],

4 [127], 5 [125], and 6 [124] use HRTF-based headphone reproduction. Background sounds are employed in Chapters 2 [128], 3 [126], 5 [125], and 6 [124]. Chapter 6 [124] presents the highest acoustic source diversity and interaction between sources.

1.3.2 State of the Art

When listening to speech in complex scenarios, parts of the target signal might be masked. Masking generally refers to "the process by which the threshold of hearing for one sound is raised by the presence of another" [11]. If speech signals, are (partially) masked by the presence of another sound, speech intelligibility is decreased. Distinct challenges arise depending on the type of masking: energetic masking occurs when a target signal and the interfering sound overlap in time and frequency [73, 74, 234, 466, 469]. Informational masking describes the increased difficulty in understanding a target signal due to the similarities between the target and the masker [233, 234, 466]. Many everyday signals give rise to both energetic and informational masking [233, 234]. Speech maskers represent a special case in masking due to their temporary structure, which enables "in the dip" listening where glimpses of the target signal are audible [233]. The differential effect of different maskers on listening performance is the research intent of a number of studies (e.g., [14, 66, 381, 382]), also in the present thesis (see Chapters 2 [128] and 5 [125]).

A spatial separation between maskers and the target leads to distinct spatial cues for both signals (see Section 1.3.1). Spatial separation thus can facilitate stream segregation and reduce masking effects via spectro-temporal separation, and increase speech intelligibility. This effect is called spatial release from masking (SRM) [66, 312, 466]. Studies on the influence of SRM on speech intelligibility are numerous [229, 275, 312, for a review] and show that accurate auditory reproduction is necessary to benefit from SRM effects.

Many studies in listening research use background stimuli that are stationary pink, white, and speech-shaped noise or simple multi-talker babble together with non-spatial reproduction, which disregards the complexity characteristics 1 (spatial distributed sources), 2 (variety of sources), and 3 (source-source interactions) by Weisser et al. [484]. Incorporating spatial reproduction, multiple time varying background sources, and conversations between sources (e.g., signified by talker-alternations and prosody [172]) in studies is important if they aim to reflect real-life conditions. Wolters et al. [497] characterise the acoustic scenarios and listening situations that people encounter in everyday life and derive a structured set of prototypical scenarios, the common sound scenarios (CoSS). These scenarios can inform experiment designs by suggesting which everyday contexts, tasks,

and background conditions should be represented in more realistic laboratory tests.

Multiple studies have attempted to extend source complexity. Some follow the approach of investigating people directly in real-life scenarios, such as through soundwalks [411]. These, however, lack reproducibility. Another approach to present a high level of source complexity in a controlled way is through recordings of real-life soundscapes. For example, Thiemann et al. [453] provide a database of 18 scenarios in the categories domestic, office, public, transportation, nature, and street, which are all five minutes in length and recorded with a microphone array. Heittola et al. [183] published ten binaurally recorded urban sound scenarios with a total duration of six hours. Both these databases are however primarily used for algorithmic speech enhancement. A database which is more commonly used for participant studies is the ambisonic recordings of typical environments (ARTE) database by Weisser et al. [485]. ARTE includes 14 scenarios, each about two minutes long, ranging from relatively quiet scenarios (e.g, library) to busy scenarios (e.g., food court). The ARTE scenarios have been used in studies of communication behaviour in normal-hearing and hearing-impaired listeners [32, 33, 284, 315, 483], and in investigations of listening effort [464] and speech intelligibility [20, 78, 313]. Only few of these studies directly contrast these complex recordings with other (synthetic) background stimuli. For example, Badajoz-Davila and Buchholz [20] compared ARTE scenarios with standard babble noise while examining speech intelligibility in cochlear implant users. They found that speech intelligibility was lowest in the realistic background stimulus and concluded that traditional speech-in-noise tests overestimate performance, highlighting the relevance of complex stimuli.

While such recordings provide high realism, they offer limited possibilities for manipulating individual scenario components (for example turning a particular source on or off, or varying SNR conditions). To allow systematic control over acoustic complexity, simulated background sound scenarios are often employed instead (e.g., [27]). Breuer et al. [62] developed a tool to facilitate the generation of sound scenarios consisting of both constant and varying background sound components with pseudo-random timing. These scenarios can be rendered in real time or pre-rendered and enable controlled manipulation of background sound complexity (for example number of sources, temporal variability, spatial distribution). This tool has also been used in the scope of this thesis in Chapter 6 [124] to provide a complex open-plan office noise, which moves away from simplified background sounds in a controlled manner.

Similar time-varying simulated background sounds have been employed by Leist et al. [264] to investigate speech comprehension and serial recall performance with either spatial or non-spatial presentation of the background sound.

The complex background sound significantly reduced performance compared to a silence baseline, while spatialised sound presentation improved performance. In a similar study, Yadav et al. [500] investigated serial recall performance in silence and with pre-recorded spatialised time-varying open-plan office sounds which either included speech or not. They found a significant performance decrement caused by speech in the background signal. While spatial reproduction (in the case of Leist et al. [264]) and background speech (in the case of Yadav et al. [500]) surely contribute to source complexity as conceptualised by Weisser et al. [484], neither of these studies compared their complex background stimuli to synthetic background stimuli, such as white or pink noise (but see [501]). Such an investigation was recently made by Breuer et al. [61] with regard to auditory selective attention. They investigated auditory selective attention in a VR classroom and compared spatial time-varying simulated background sound with white noise. Performance differences between these acoustic conditions could be determined, highlighting that simplified noise stimuli might underestimate the effects of realistic background sounds.

In their study, Breuer et al. [61] did not only simulate a complex background sound, but also room acoustic aspects of the virtual classroom. Such effects are part of the complexity type 2 (environment) by Weisser et al. [484] and are discussed in the following section.

1.4 Environment Complexity

Weisser et al. [484] contribute environment complexity primary to room acoustics and the medium of propagation. As the propagation medium is air in most studies on listening performance, the latter is negligible in the context of the present thesis. Thus, this section will focus on studies which examined room acoustic influences.

1.4.1 Technical Background

As discussed in Section 1.3.1, the human body acts as an acoustic filter which provides spatial information. The acoustic environment itself also acts as a filter. Sound waves are reflected, absorbed, scattered, and diffracted by room boundaries and objects, depending on their geometry and material properties. As a result, not only the direct sound reaches the participant, but also early reflections and late reverberation. An established measure to quantify the room influence is the reverberation time (RT). RT is commonly quantified with T_{60}, i.e., the time required for the SPL to decay by 60 dB. To account for different SNR conditions, different methods exist; T_{30}, for example, is based on doubling the decay time

from −5 dB to −35 dB. For a fixed source–receiver pair, the influence of the room can be described by the room impulse response (RIR), i.e., the impulse response from the source position to the receiver position in that particular room. [251]

As room acoustic cues contribute to sound localisation and distance perception, omitting room influences in auralisation, i.e., reproducing anechoic conditions, might diverge from natural auditory processing. One option to include room acoustics is to measure RIRs in an existing room for discrete combinations of source and receiver positions and to use these RIRs for auralisation by convolution. However, this approach is limited to existing spaces and does not facilitate the examination of isolated factors. As a result, room acoustic simulations have become increasingly employed in research. Here, RIRs are predicted based on a virtual model of the room using either wave-based methods, which provide high accuracy at the drawback of requiring high computational power, or geometry based methods (e.g., image-source models, ray tracing), which are computationally efficient at the cost of not accurately representing low frequency and wave based effects [251, 474].

To account for the relevance of room acoustic influences, this thesis examines the effect of simulated RIR on listening performance in Chapter 6 [124]. The room acoustic simulations used in this thesis are carried out with the software Room Acoustics for Virtual ENvironments (RAVEN) [410], which implements real-time capable geometric room acoustic modelling.

1.4.2 State of the Art

Acoustic scenarios that include room acoustics, compared to anechoic or free-field conditions, have been shown to influence multiple metrics of listening performance. For example, speech intelligibility suffers from reverberation. Au et al. [19] evaluated speech intelligibility using measured RIRs of real rooms (living room, lecture room, church). They found that intelligibility with reverberation was worse than with an anechoic baseline. Utilizing simulated RIRs, Best et al. [42] examined speech intelligibility in anechoic conditions and a simulated cafeteria. Speech intelligibility was better in the anechoic scenario.

SRM benefits have also been shown to depend on the room acoustic conditions, as reverberation can decrease source separation [275, 473]. Room acoustics can further influence localisation behaviour in complex scenarios. Comparing three room acoustic conditions, Ahrens et al. [5] found that participants located a target talker in multi-talker VR scenarios both faster and more accurately in shorter compared to longer RTs.

Studies focusing on cognitive performance metrics also found influences of room acoustics. For example, Doggett et al. [109] examined cognitive performance

in a simulated open-plan office and varied the acoustic treatment of the walls as well as background sound. They found that the performance decrement caused by the ambient background sounds could be ameliorated by the acoustic treatment. Muhammad et al. [328] found that increasing the simulated isolation between adjacent rooms can improve working memory. Breuer et al. [59] examined auditory spatial attention with simulated RIRs and found that participants made more errors in reverberant compared to anechoic conditions.

Evidence for an effect of reverberations on listening effort are mixed, as discussed in more detail in Chapter 6 [124]. For example, Picou et al. [363] could not find differences in listening effort for different RIRs. Both with and without VR, Fichna et al. [133] found higher listening effort in reverberant scenarios compared to anechoic conditions. This thesis presents a study in which the presence of reverberation reliably affected listening effort compared to anechoic conditions, while small RT variations did not show significant effects (see Chapter 6 [124]).

Taken together, these results indicate that in real-life scenarios, which naturally include reverberations, listening performance differs from that in anechoic scenarios in many aspects. It is therefore commendable when studies incorporate measured or simulated RIRs, even if they are not of key interest for the research objective, both with and without VR (e.g., [115, 189, 243, 442]), or indirectly by using ambisonic or binaural recordings obtained in real rooms (e.g., [185, 313]).

1.5 Source-Environment Interactions

The third complexity type proposed by Weisser et al. [484] are source-environment interactions. Such interactions can be loudspeaker amplification systems, source-environment adaptions, and cues of other modalities. The latter are discussed in Section 1.6.

Importantly, loudspeaker amplification systems are not understood as a reproduction system of complex acoustic scenarios (as described in Section 1.3.1), but as a component thereof. For example, when investigating a party as a complex acoustic scenario, music is likely played back via loudspeakers within the scenario, which might alter the source directivity and spectral content compared to live acoustic music. Investigating loudspeaker amplification systems *within* complex scenarios has not been a key research topic in recent years. Within the ARTE database [485], some scenarios contain loudspeaker-based sources [484], such as television sound in the living room scenario, loudspeaker-based music in the party and the café scenarios, and announcements in the train station scenario. However, there is no study directly evaluating the influence of these scenario components.

The second aspect of source-environment interactions, source-environment adaption, is most commonly understood in the context of so-called Lombard

speech [505], where a talker involuntary adapts their speech to adverse situations. Such an adaption is not only reflected in an increased SPL, but also a rise in fundamental frequency, a flattening of spectral slope and increased signal duration [505]. Even though many experiments present speech in adverse listening scenarios, the SPL of the speech signals is often artificially controlled without incorporating other aspects of Lombard speech, which may influence speech intelligibility [94]. A paradigm with Lombard speech was developed by Miles et al. [313]: during recording of the speech material, talkers were presented with scenarios from the ARTE database. The talkers adapted their speech accordingly. Other databases with Lombard speech exist [8, 491]. Still, few experiments incorporate Lombard speech. Since some talkers have been shown to not only adapt auditory cues, but also lip movements, which represents a shift in visual cues, this could be particularly interesting for studies using audiovisual presentation [151]. Such audiovisual presentation is the focus of the following section.

1.6 Visual Complexity

In everyday conversation, speech is rarely perceived only auditory. Visual information can actively convey spatial information (e.g., source locations), information about the source signal (e.g., lip movement), source-interaction cues (e.g., turn-taking behaviour), and environmental cues (e.g., room dimensions and material properties). There are multiple ways to present such visual cues, which will be presented in the following.

1.6.1 Technical Background

As discussed in Section 1.3, *acoustic* complexity can be achieved by investigating real-life scenarios, using recordings of real scenarios, or by simulating acoustic scenarios. *Visual* complexity, similarly, can be achieved by investigating participants in real-world scenarios or replicas of real-world scenarios, presenting pre-recorded videos (2D or 3D), or showing simulated visual cues in VR.

VR, in the sense used here, refers to computer-generated three-dimensional scenarios that allow participants to perceive and interact with virtual scenarios in real time [496]. Visual VR content is typically created using game engines such as Unity (Unity Technologies, United States of America) or Unreal Engine (Epic Games, Inc., United States of America), which allow for modelling of objects, lighting, animation, and physics-based interaction. Several types of display devices are used for presenting VR scenarios. The most common ones are HMDs, which provide stereoscopic images directly in front of the participants' eyes and track head-movements to update the presented scenario accordingly; and CAVE

systems, which are room-sized setups where images are projected onto multiple surrounding walls (and sometimes the floor) [96, 249]. But VR scenarios may also be presented on computer screens.

VR has the potential to create highly immersive experiences (see Section 1.2.3), where participants feel surrounded by and embedded in the virtual scenario. At the same time, it can cause simulator sickness or cybersickness, and insufficient frame rates or high delays can break the illusion [272]. This must be carefully considered in experimental designs to avoid confounding factors. For evaluating listening performance in complex audiovisual scenarios, visual VR can provide co-verbal cues such as lip movements via ECAs, i.e., animated virtual characters designed to engage in spoken interactions [87].

In the present thesis, visual VR was employed in several studies to provide visual information, including Chapters 4 [127], 5 [125], and 6 [124]. The corresponding visual scenarios were created in Unreal Engine (Epic Games, Inc., United States of America) and presented via HMDs. These scenarios include a virtual laboratory, a living room, and an open-plan office, in which one or more ECAs, simulated with the MetaHuman Creator (Epic Games, Inc., United States of America) delivered speech material while participants are exposed to different acoustic and visual conditions (e.g., varying background sound types, room acoustics, or details of co-verbal cues).

1.6.2 State of the Art

With the rise in accessible simulation-based VR technology, but also through video recordings, visual information in the form of source and / or scenario visualisation have been incorporated in many listening experiments. The following sections discuss the effects of how visual information is presented and how auditory and visual information interact, as well as findings of studies incorporating source and scenario visualisation. While a multitude of experiments examining the effect of audiovisual complexity on, e.g., immersion and presence exist, this chapter reviews existing work with a particular focus on listening-related performance, working memory, and listening effort in accordance with this thesis' objectives.

Presentation Technique

Many studies have investigated if the method with which visual information is presented (i.e., as a video or simulation, with computer screens or HMDs) influences performance and perception.

Some studies go through the effort of creating a digital twin of a real-world scenario in VR. A database of simulated audiovisual scenarios was introduced by van de Par et al. [465]. They provide a visual model and an acoustic geom-

etry model with acoustic surface properties for an underground station, a pub, and a living room scenario. As real-life twins of these scenarios exist, van de Par et al. [465] also provide room acoustic measurements. These scenarios have been used in a number of studies [3, 134, 154, 190]. Especially the living room scenario has been compared to its real-life twin with regard to speech intelligibility [413, 414] and has been shown to yield comparable results. Also in other scenes and across several audio-related dependent variables, such as loudness perception or listening effort, results between HMD-based VR and real-world scenarios are comparable (e.g., [133, 189, 409]). Marginal differences have been shown in localisation and distance perception [204, 442]. Taken together, these findings suggest that many acoustic metrics behave similarly in simulated VR scenarios and real-life scenarios.

Instead of comparing real-world scenarios with simulations, some studies evaluate the influence of the device with which the simulated scenario is presented (e.g., computer screen and HMD). In visually simulated scenarios, using a HMD has been shown to benefit acoustic sound quality [380] and auditory-based speech perception [397]. Other VR studies replicate established laboratory studies and compare performance quantitively instead of directly, again with comparable results between methods (e.g., [60, 328]). Also evaluating the influence of the presentation device, in Chapter 4 [127] of this thesis, working memory performance was compared between HMD and computer monitor presentation. No difference between presentation techniques was found.

Simulation is only one way to obtain virtual visual scenarios. Pre-recorded scenarios, similar to those in the acoustic ARTE database [484], also exist for audiovisual stimuli. Two-dimensional videos are a common method in traditional studies to present co-verbal cues (see Section 1.6.2). Higher immersion is generally achieved through 360° videos. Robotham et al. [386] introduced a database of 360° videos with accompanying ambisonic audio of everyday scenarios (e.g., parks and forest scenarios, traffic, conference venues), which have been used to evaluate quality-related measures such as immersion, quality of experience, and visual comfort [147, 427]. In a recent publication, Fremerey et al. [146] compared audiovisual scene analysis performance in a pre-recorded 360° classroom video to a simulated VR classroom, both presented via an HMD. The 360° videos yielded better performance than the virtual VR scenario. However, it is unclear how much this effect was dominated by the difference in co-verbal cues provided in the video and the animation (see Section 1.6.2 for details on co-verbal cues). Seol et al. [416] performed a speech intelligibility test under three conditions: in a 360 ° video of a cafeteria with a target talker sitting right in front of the participant, presented via an HMD, the same scenario presented on a traditional computer monitor, and lastly, and audio-only condition. Unfortunately no statistical comparison

between the three conditions was performed, but subjective evaluations revealed a preference for the availability of co-verbal visual cues.

In summary, the presentation of visual cues in audiovisual complex scenarios can be achieved through many different techniques which have differential advantages and disadvantages with regard to controllability and realism as well as their effect on task performance. The VR based experiments in this thesis rely on simulated scenarios due to their high customisability and control (see Chapters 4 [127], 5 [125], and 6 [124]).

Audiovisual Interaction

If auditory and visual information are processed together, as in the previously mentioned studies, they can be integrated into a single object or event [259]. Such multisensory processing can lead to improved performance for bimodal (e.g., audiovisual) stimuli compared to the unimodal (e.g., audio- or visual-only) baselines [439]. Whether audiovisual signals are bound together depends on a multitude of factors. Integration does not require perfect synchrony; there is an asymmetric window of tolerance for temporal offsets, typically broader for visual-leading than for visual-lagging conditions [299, 305]. Spatial co-location consistently facilitates tasks that involve spatial attention, with weaker effects for non-spatial tasks [437]. Temporal and spatial coherence may also help match a voice to the correct face, aiding stream segregation in crowded scenarios through object-based selective attention [97, 285, 419], which can help to navigate complex audiovisual scenarios. For example in multi-talker scenarios, seeing faces improves selective attention only when faces and voices are spatially aligned, with stronger benefits at lower SNRs and more complex scenarios in which multiple multisensory streams compete [136]. Another obvious benefit arises from the availability of lip movements as an audiovisual co-verbal cue. This is discussed in more detail in Section 1.6.2.

A further discussion point in literature is whether audiovisual binding is performed automatically or requires attention (e.g., [71]). Grant and Bernstein [165] highlight that when stimuli demonstrate an audiovisual benefit, a high integration efficiency it is often assumed. However, they propose that the audiovisual benefit is related to the spectral relationship of target and background, resulting in a different audiovisual benefit in different background sound conditions, while the efficiency stays constant.

As multisensory processing is complex and multifaceted this overview is far from exhaustive. A multitude of studies has focussed on the cognitive processes involved in multimodal processing, the differences between multimodal integration and binding, as well as object-formation and its role in attention [259]. An

overview over these research areas is given in the book by Lee and Wallace [260]. This thesis, however, focuses on more behavioural aspects of audiovisual stimulus presentation.

Audiovisual Incongruence

As mentioned above, several preconditions have to be fulfilled in order for an auditory and visual signal to be integrated. Mismatches can lead to stimuli being perceived as incongruent, which may require additional resources to resolve conflicts [71].

If auditory and visual stimuli are mismatched in meaning, they are semantically incongruent [436]. A well-known example of semantic incongruence is the *Stroop task* [444], where participants see colour words printed in a different coloured ink (e.g., the word "red" printed in green), and are asked to name the colour of the ink. Semantic incongruences have also been evaluated in virtual environments. For example, Tromp et al. [458] investigated semantic mismatches in a VR scenario. Virtual humans which had a plate in front of them and made a statement about their food, e.g., "I just ordered this salmon". The content either matched (a salmon was displayed on the plate) or not (pasta is displayed on the plate). Using neural imaging, Tromp et al. [458] detected a difference in the processing of congruent and incongruent stimuli. Breuer et al. [63] evaluate auditory selective attention and examined semantic audiovisual incongruence. Participants had to categorise an auditory target word into a certain category (e.g., an animal into flying or non-flying) among a mixture of voices. Simultaneously, they are presented with a matching or mismatching picture of an animal on a virtual classroom blackboard, which they were told to ignore. Breuer et al. [63] found that the presentation of the animal pictures generally increased response times, with incongruent pictures increasing response times more than congruent ones. Further, fewer errors were made with congruent compared to incongruent pictures. Semantically congruent stimuli have also been shown to enhance short-term memory performance in a VR free recall task [10] and are evaluated in this thesis in Chapter 4 [127].

Under certain circumstances, the incongruence between the auditory and visual information can lead to a so-called audiovisual illusion, where the combined percept differs from the information conveyed by either modality. A well-known audiovisual illusion is the ventriloquism effect [195]. In this effect, the perceived position of a beeping sound is diverted towards a synchronously blinking visual source [195]; i.e., the visual location influences the perceived auditory location. The strength of this bias mainly depends on movement and temporal correlation (see also [84, 259]). Such spatial incongruences are evaluated in Chapter 4 [127].

Another established illusion is the McGurk effect [306], where lip movements conflict with the heard syllable. As a result, participants percepts might be fused into a syllable that differs from either input (see also [259, 461]). Such incongruent stimuli have been shown to diminish SRM benefits [165]. However, Brown and Strand [71] demonstrated that the decisive component was not the congruence of McGurk stimuli but whether these stimuli were integrated. Contrasting congruent and incongruent stimuli did not show a difference between response times, but responses were faster in illusion trials compared to non-illusion trials.

Audiovisual incongruences are particularly important for VR presentation (see [290, 297]), as they may impact immersion and can occur easily due to scenario design (e.g., with inadequate generated lip movement) and real-time computation (leading to, e.g., temporal lagging) [425]. Providing coherent user interaction in VR, e.g., by synchronised audiovisual virtual hand illusions [86] or sound feedback according to the user's movements, e.g., walking [193], has been shown to increase immersion.

Further, many studies focussing on audiovisual interaction and incongruences are done in simple scenarios, e.g., without background sound, but it is hypothesised that multimodal processing in these scenarios differs from that in more complex scenarios (e.g., [82, 136]). Illusions based on incongruent lip movements, such as the McGurk effect, have been replicated in VR with animated ECAs [425, 452] and with 360° videos [160]. The ventriloquism illusion has been also been examined in VR, but opens up the research question whether the realism of the visual stimulus influences the illusion [142, 205]. However, no clear relationship could be found [205], with some research even indicating stronger effects for simpler visualizations [142]. This calls into question the applicability of such illusions for more realistic audiovisual scenarios, with, e.g., ECAs as embodied sound sources, and is thus addressed in Chapter 4 [127].

Source Visualisation

This section lists several aspects of source visualisation, including embodiment, co-verbal cues and source-source interactions.

Embodiment. Providing visual information creates the opportunity to show a visual resemblance of a source, i.e., a visual *embodiment* of the source.

In VR scenarios with speech, target sources are often embodied through ECAs (e.g., [114, 243, 399], see also Chapter 4 [127], 5 [125], and 6 [124]) or, in case of multiple participants conversing, virtual avatars as embodiment of the participants [208]. Some studies also use "virtual appliances" to avoid the need to embody and animate virtual talkers. For example, Moore et al. [324] implemented a speech intelligibility task with real talker videos displayed on a

virtual computer screen in a virtual cafeteria, but without visually embodied background talkers (see also [421]). Mohanathasan et al. [322] used static virtual faces as target talker embodiment when examining listening effort and memory.

Mixed techniques exist also for background sound sources. Citing computational constraints, confounding effects of visual distractions, or a limited field of view, some studies do not embody their background sources (e.g. [61], see also Chapter 5 [125]). To provide realistic co-verbal cues some studies embody the background as virtual screens (e.g. [421]) or as virtual loudspeakers (e.g., [133]). Other studies reduce animation demands: investigating room acoustic influences on cognitive performance, Doggett et al. [109] showed stylised embodied background talkers, but these were not animated. Himmelein et al. [189] compared a laboratory scenario with a VR scenario and embodied the experimenter in VR, but do not mention any animation technique. Shavit-Cohen and Zion Golumbic [420] used a virtual café scenario to study spatial attention. Their ECAs were animated only with lip movement.

However, it has been shown that animated embodiment is preferred by participants: Ehret et al. [115] presented participants with a virtual seminar room with a virtual lecturer and peers. A dynamic, time-varying background sound, comparable to that used by Breuer et al. [61, 62], was employed. Different visualisation conditions for the background sources (no visualisation, static visualisation, animated visualization) were tested. Animated background visualisation was subjectively preferred and perceived as more coherent than static or no visualisation, but it did not significantly affect objective listening effort measures (see also [116]). Animated embodied background sound sources are thus provided in Chapter 6 [124].

Embodiment can provide spatial information as well as co-verbal cues. Both of these cues have been shown to facilitate listening performance: while visual spatial information can improve auditory scene analysis (see Section 1.6.2), co-verbal cues can improve listening performance, as discussed in the following section.

Co-verbal Cues. Speech-related visual stimuli conveyed in the visual domain are called *co-verbal cues*. Co-verbal cues include lip and articulatory movements, facial expressions (e.g., conveying emotions), head movements, and gaze. They convey both timing and content information, are central to natural listening [155, 260], and are important for navigating adverse listening scenarios [85].

Videos of lip movement have been shown to improve speech intelligibility compared to audio-only conditions [163, 259, 446]. Benefits tend to be larger at lower SNRs or in the presence of reverberation, where the auditory information is masked by the background sound [259, 283, 445]. The contribution here is mul-

tifaceted: auditory information can be supported both by phonetic information (e.g., place-of-articulation visible on the lips) and temporal information (e.g., the visual amplitude/mouth-opening) [164, 184]. Visual speech can be particularly beneficial for listeners with hearing impairment, where lip movement can compensate for distorted acoustic input [121, 136, 137]. While these studies often use stationary background sound, the audiovisual benefit of providing co-verbal cues has been extended to more complex audiovisual scenarios: using background sound scenarios from the ARTE database [484], Buchholz et al. [78] investigated speech intelligibility with video recordings presented on projector screens. Performance reliably increased in audiovisual compared to audio-only conditions. Deaglau et al. [99] examined audiovisual benefits in unscripted speech with background sound. They either presented videos of talkers on projector screens only, audiovisual videos, audiovisual videos with masked lip movements, or audio-only stimuli. The results revealed a significant audiovisual benefit, which was carried primarily by the availability of lip movement information.

The studies above use pre-recorded videos, but speech intelligibility benefits have also been demonstrated for animated ECAs. However, the speech intelligibility benefit provided by virtual lip movement tends to be less than that provided by video-recorded lip movements. For example, Nirme et al. [337] demonstrated speech comprehension benefits from seeing lip movement conveyed by virtual faces compared to audio-only, even when the virtual faces are not perceived as realistic. The benefit provided by video recordings was, however, significantly larger than that provided by virtual faces. Cooper et al. [95] investigated speech intelligibility with simulated lip movements and ECAs in a projector-based VR restaurant scenario. They compared audio-only and audiovisual conditions with both behavioural and neural measures of speech intelligibility and found higher accuracy with audiovisual speech presentation (see also [105, 185]). Otten et al. [347] found that in ECAs, speech intelligibility benefits highly depend on the technique of lip-animation.

If lip movements do not match the auditory content, incongruences can appear and alter the perceived speech information and / or increase processing demands. This was discussed in terms of the McGurk illusion in Section 1.6.2 and holds true both for videos of lip movement and animated characters (e.g., [452]).

Lip movements have not only been evaluated with regard to speech intelligibility, but also listening effort. Evidence is mixed regarding how lip movement affects listening effort. Adding visual cues does not always reduce listening effort; individual differences in lip-reading abilities and working memory capacity have been shown to play a role [364, 365]. Comparing videos of real faces, virtual faces, and audio-only conditions, Nirme et al. [337] found no difference in perceived listening effort. Hendrikse et al. [185] detected decreased perceived listening effort

for animated and video presentation of talkers compared to audio-only conditions. Other studies use ECAs to examine listening effort but did not directly contrast different presentation techniques (e.g., [105, 399]). However, there is only limited research on the effect of co-verbal cues on working memory. Some studies report that audiovisual presentation can enhance working memory of non-verbal materials [102]. Interestingly, the main focus in working memory research has been the influence of background sounds on lip-read lists (i.e., only visual co-verbal cues) [108, 221], however, audiovisual speech presentation has also been shown to enhance working memory [148].

Besides information, auditory signals can also convey the talkers' emotions, such as happiness, fear, or anger [292, 366]. Studies show that emotions are less confused in audiovisual presentation compared to audio-only presentation, facilitating emotion detection [291]. Emotions have also been successfully integrated into ECA and show good recognition rates [291].

In summary, co-verbal cues provide important information which contributes to a holistic listening experiment. Thus, the VR-based studies in this thesis (Chapters 4 [127], 5 [125], and 6 [124]) employ ECA that are animated to speak the target stimuli.

Source-source Interactions. A third characteristic of source complexity, according to Weisser et al. [484], are source-source interactions, where "human talkers [...] form interactive communication networks" [484, p. 4]. Communication between human talkers relies on several factors, such as prosody and speech rate, but also visual factors such as gazing, gestures, and movement [172].

Hendrikse et al. [185] found, e.g., that when following conversations between multiple talkers (either as videos or as animated ECA), the gazing pattern can help participants identify the talker they should attend to, aiding localisation performance and speech intelligibility. They also argue that gaze patterns can direct attention. Using animated ECAs only, Ehret et al. [114] separately investigated the contribution of gestures, gazing, and breathing in turn-taking behaviour. Listening-only conditions as well as conditions in which participants had to take part in the conversation were examined. Ehret et al. [114] investigated social presence as well as speech gaps in interactive conditions. While social presence ratings did not differ between conditions, gestures were shown to reduce speech gaps. The same audiovisual turn-taking behaviour is implemented in Chapters 5 [125] and 6 [124] to provide participants with this aspect of visual complexity.

Environment Visualisation

When immersed in a VR scenario, scenarios almost always include a virtual room (or less often outdoor scenarios). As discussed in Section 1.4, rooms filter

sound through interaction with the walls, ceiling, and floor, as well as materials, resulting in a characteristic reverberation. For a congruent audiovisual experience, thus, visual rooms should be accompanied by acoustic RIRs. While some studies employ RIRs to provide congruent audiovisual cues [61, 413, 414], others present anechoic acoustic conditions anyway, e.g., because room acoustic conditions are not of key interest for the research question (e.g., [57, 63, 114]).

To evaluate the importance of coherent environmental cues, some studies actively manipulate the audiovisual congruence of the scenario by presenting reverberation which does not match the visual scenario. For example, incongruent visualisations have been shown to affect distance perception [156, 371, 394]. Perceived reverberation, however, seems to be insensitive to the size of the visually presented room [412]. Similarly, incongruent audiovisual scenarios have also been shown be tolerated in speech intelligibility tests, if they are not too extreme [370]. Further, auditory scene analysis is largely robust against incongruent audiovisual scenarios and more dominantly influenced by reverberation time and the number of background talkers [5] (see also [133]).

This thesis presents anechoic conditions in most Chapters, but specifically examines room acoustic influences in Chapter 6 [124], where variations in acoustic RIRs are reflected in changes of the visual model to counteract potential incongruence effects.

1.7 Receiver's Task Complexity

Finally, Weisser et al. [484] highlight the importance of the receiver's task for complexity. They state that the task influences how a participant interacts with and processes the scenario and thus defines which aspects of complexity are important to an investigation. The task is thus equally important as the design of the scenario itself.

Many indicators of listening performance exist and efforts have been made to increase complexity in related paradigms through task and/or stimulus complexity, e.g., speech intelligibility [52, 313, 314, 316], speech comprehension and understanding [58, 77, 78, 237], auditory selective attention [57, 60, 63, 285, 342], or auditory spatial analysis [5]. Instead of employing defined paradigms, some studies directly observe the behaviour of participants, e.g., while listening [185, 99], or during conversations [32, 33, 284, 313, 315], also in VR [208].

This thesis focusses on verbal working memory as an index of cognitive capacity relevant to speech processing, and listening effort as a multifaceted measure, which reflects the allocation of mental resources required to achieve a desired level of understanding under specific acoustic demands. The underlying constructs,

influencing factors, and paradigms of these two measures are introduced in the following.

1.7.1 Working Memory

In adverse listening situations, irrelevant stimuli might mask parts of a sentence. In such situations, participants often rely on memory performance: they temporarily store the (partially masked) sentence in their working memory, until they can fill in the intelligibility gaps through information deducted from the sentence structure [466]. Working memory describes the ability to temporarily store, maintain and manipulate information [25]. An example for this could be hearing "_oney", which could be interpreted as "honey" or "money". The correct interpretation is derived from the context, but for that matter sentence has to be held in working memory until the context suffices to resolve the ambiguity [466].

Therefore, when evaluating listening performance in adverse and complex scenarios, it is important to consider how working memory performance is affected. It is thus not surprising that working memory is evaluated in many noise-based studies, e.g., when investigating the influence of background sounds in open-plan offices [500]. Consequently, all studies in this thesis evaluate working memory (see Chapters 2 [128], 3 [126], 4 [127], 5 [125], and 6 [124]). In order to interpret the findings of the presented experiments in a meaningful way, it is important to understand what conceptualises working memory. The following sections thus give an overview over models of working memory, distractors, and paradigms.

Theoretical Frameworks of Working Memory

A detailed overview over theoretical frameworks of working memory is given in Chapters 2 [128] and 3 [126]. This section thus gives only a brief overview over the two most prominent models of working memory: the multi-component model and the object-oriented episodic record (O-OER) model.

The multi-component model proposed by Baddeley and Hitch [25] (see also [18, 21]) assumes separate short-term storage systems for verbal and visuospatial information (the phonological loop for verbal and the visuospatial sketchpad for visual information), coordinated by a domain-unspecific control system (the central executive) and, in later versions, complemented by an episodic buffer [24]. Each of these subsystems is responsible for distinct tasks.

The central executive is a limited-capacity attentional control system, that shifts attention between tasks [21, 22, 24]. The episodic buffer is a limited, temporary, and passive store for small portions of information, which may be multimodal, e.g., audiovisual [23]. The episodic buffer enables interaction of information and relation to the long-term memory [23]. The main focus of most studies and

present thesis, is, however, the nature and relationship of the phonological loop and the visuospatial sketchpad.

Verbal information is assumed to be stored in the phonological loop. But not only target speech or information of interest; because auditory information is processed automatically and without deliberate intent, task-irrelevant auditory information (i.e., background sound) is also assumed to gain access to the phonological loop [396]. This may result in interference between target and task-irrelevant information and can impede verbal working memory. An example for this is the irrelevant sound effect (ISE), where recall of target information is impaired by the presence of background sound. Visuospatial information is assumed to be stored in the visuospatial sketchpad, separate from verbal information. In other words, the multi-component model supports an interference-by-content principle [293]: irrelevant verbal information should disrupt tasks that rely on phonological coding, but not tasks that depend primarily on visuospatial working memory, and vice versa. Influences of task-irrelevant information on working memory are discussed in more detail in the next section.

The multi-component model has been constantly evaluated and extended (e.g. [24]). In one version of the model, for example, the visuospatial sketchpad is subdivided into subsystems specialised for visual object properties (e.g., shape, colour) and spatial attributes (e.g., locations, trajectories) [21, 277]. Building on this idea, Lehnert and Zimmer [262] proposed a modality–domain distinction in working memory. Here, modality refers to the sensory input channel (auditory vs. visual), whereas the processing domain reflects the type of information extracted from that input (verbal vs. non-verbal). The non-verbal domain can be further subdivided into object-related and spatial processing. Importantly, according to this account, spatial working memory is amodal: spatial information derived from auditory or visual input is represented in a common spatial store, whereas object-related information remains modality-specific. Consequently, spatial locations can be encoded and should be disrupted by both visual and auditory stimuli that carry spatial cues [100, 349, 101, 220, 265, 295, 457]. More information on the processing of spatial information in working memory is provided in Chapter 3 [126].

An alternative model of working memory is proposed by the O-OER model [219, 222]. In contrast to the multi-component framework, the O-OER model assumes a unitary short-term memory system in which inputs from different modalities are stored as amodal objects. Crucially, disruption is driven by interference-by-process rather than interference-by-content [228, 293]. That is, what matters is whether the irrelevant stimulus engages in the same ordering processes as the target, not whether they share verbal or visuospatial content. According to this view, any changing-state distractor that carries order information (e.g., vary-

ing letters, syllables, words, tones) can interfere with tasks that require order information, regardless of the input modality or domain.

Other models of working memory exist, but are not discussed in the scope of this thesis (see [295] for an overview). Evidence for the multi-component model and the O-OER is mixed. A detailed discussion is provided in Chapter 3 [126]. Both models, however, predict differential effects of distractors on working memory. An overview of stimuli shown to disrupt working memory is given in the following section.

Factors Influencing Working Memory

A well-studied example of distraction in verbal working memory is the ISE, originally described by [89] (see also [117]). The ISE refers to the robust finding that task-irrelevant background sound impairs performance in working memory tasks compared to a silent baseline condition.

The duplex-mechanism account of auditory distraction [199] distinguishes two main components of the ISE: a changing-state component and an attention capture component. The changing-state component is driven by variation in the irrelevant sound. Compared to steady-state sounds, such as continuous noise, pure tones, or repeated letters (e.g., "M-M-M-M"), are typically less disruptive than sequences of different syllables or letters (e.g. "M-U-L-S"), running speech, or in some cases also music [404]. These temporal changes convey order information that can conflict with the order information in the target sequence, disrupting serial memory performance [293] (see also O-OER model, Section 1.7.1). The attentional capture component, in contrast, arises when alerting signals in the task-irrelevant signal (e.g., one's own name, meaningful speech segments, or deviant items in an otherwise regular sequence, such as "M-M-G-M-M" instead of "M-M-M-M-M") temporarily divert attention away from the primary task [199].

Based on these assumptions, a wide range of task-irrelevant sounds have been shown to produce an ISE in verbal serial recall, including sequences of letters or syllables [81, 227, 401], unrelated words [174, 447], running speech in the participant's native or non-native language [238, 401], as well as music [404] and have been applied to scenarios with real-life relevance, such as open-plan offices and classrooms [170, 225, 328, 383, 500, 501, 502]. Two experiments in the scope of this thesis further highlight the relevance of semantic processing in verbal working memory: Chapters 2 [128] and 5 [125] demonstrate that stimuli carrying semantic meaning, such as intelligible sentences, disrupt performance in two different memory tasks more than pseudo-speech without meaning.

Not only the content of the task-irrelevant auditory signal, but also the spatial information conveyed by it may be of interest, especially when aiming for more complex presentation. While spatial-changing background stimuli do not

impair verbal working memory (see Chapter 3 [126]), detrimental effects of spatial changes in the target information could be demonstrated (see Chapter 2 [128]), given that the spatial changes occur at a fast rate. These findings build upon the study by Hughes et al. [200, 202] who found alternating the perceived location of successive target items disrupts serial recall with a monotic audio reproduction (e.g., left–right–left–right headphone channel). Spatial changes in the memory target are evaluated in Chapter 2 [128] while spatial changes in the task-irrelevant stimuli are evaluated in Chapter 3 [126].

Evidence for visual distraction on verbal working memory is lacking and there are only few studies investigating visual distractors in visuospatial memory tasks. Liebl et al. [271] investigated changes in workspace lighting and found no effect on memory task performance. Similarly, Lange [255] could not demonstrate an effect of flashing colour stripes during stimulus presentation on verbal memory performance. In visuospatial working memory, however, performance decrements could be demonstrated with visual noise and irrelevant pictures [302], and cross-modal interference exists [49] (see Section 1.6.2).

To meaningfully evaluate distractions on working memory, different paradigms of visuospatial as well as verbal memory have been proposed. The following section will give an overview over the memory paradigms most relevant to this work.

Paradigms

In classical research, verbal working memory is most commonly assessed with serial recall tasks [207]. In this paradigm, participants are presented with sequences of discrete items and are asked to reproduce them in the original order after a brief retention interval. Depending on the domain of the to-be-remembered items, a distinction is made between verbal serial recall and visuospatial serial recall, but also other forms, e.g., audiovisual verbal serial recall (avVSR) have been introduced (see Chapter 4 [127]). In visuospatial serial recall, common target items are pictures or dots on an invisible grid [206, 295] (see Chapter 3 [126]). In verbal serial recall, target items are usually words, letters, or digits [206, 401, 405]. Item sequences can be presented visually (as written text, pictures, or lip-read stimuli [102, 108, 179, 447]) or auditorily (as spoken items [404, 447] or object sounds [102]), with the latter most closely resembling real-world listening situations. Serial recall performance is typically quantified as the proportion of items recalled in the correct position, sometimes complemented by response times [206]. Typically, performance is best at the beginning of the sequence, the so-called *primacy effect* [206]. For auditory in contrast to visual presentation, performance has been shown to increase at the end of the sequence. This is called the *recency effect* and assumed to correspond to a mental "echo" of the last items.

The serial recall task is employed in Chapters 2 [128], 3 [126], and 4 [127] to quantify working memory because it allows for meaningful evaluation of complexity increments by comparing findings with established, less complex, publications.

Other paradigms for examining memory exist, such as the n-back task [92], free recall [206], the Corsi block task [417], and working memory span tasks [92]. All paradigms on this far from exhaustive list have one thing in common: the to-be-remembered targets are short, well-defined, and often unrelated items. This does not correspond to real-life, where memory targets are often more complex, e.g., running speech. Therefore, more paradigms with longer stimulus material have been developed. These paradigms are often not limited to examinations of working memory, but incorporate multiple components of information processing.

An example for a more complex memory paradigm, which integrates speech comprehension, understanding, and recall, is the HTR paradigm [406]. In this paradigm, the target signals are short narrations of family stories. Each HTR narration consists of ten sentences. After the presentation of the texts, participants are asked nine content-related questions. These questions address different details of the text, such as names, family relationships, ages, hobbies and professions, or other information, and are predefined for each text [406].

Based on the number of correct responses per text, memory performance is assessed. Although visual presentation is possible, e.g., as a written text, the presentation of the HTR text is most commonly auditory (cf. [135, 320]). Implementations in VR exist as well, where the texts are narrated by ECAs and the auditory presentation is accompanied by co-verbal cues (cf. [115, 114, 399], see also Chapters 5 [125] and 6 [124]). To enable such implementations, a database of audio and face-tracking data for the - currently 34 - HTR texts for one male and one female talker was created in the scope of the present work [129].

In comparison to the classic serial recall, the HTR paradigm represents a step towards more complex paradigms which potentially capture real-life behaviour better than simplified tasks. This, however, comes at the cost of experimental control and comparability across trials. The HTR texts differ in length, position of the item targets, and also in difficulty [323]. Still, such influences must be accepted, as they can also occur in real life and thus potentially increase generalisability and external validity of findings (cf. Section 1.2.4).

The HTR has been examined with regard to co-location or separation of two talkers narrating the stories [135, 321], with regard to target talker voice quality [399] and embodiment of background actors in VR [115], under different kinds of background sound [320], and in relation to the participants' cognitive capacities [321]. In the present thesis, the HTR served as an indicator for the influence of meaningful and non-meaningful background speech (see Chapter 5 [125]) and room acoustic simulations (see Chapter 6 [124]) on verbal memory. In most of

these studies, the HTR is not used to only evaluate memory, but also, together with a secondary task, to examine listening effort.

1.7.2 Listening Effort

Listening to speech, particularly in the presence of background sounds, does not only challenge speech intelligibility and memory, but can also give rise to listening effort. Even if speech intelligibility is constant in two different sound conditions, the effort needed to maintain performance might be different [466]. Listening effort is commonly understood as the amount of cognitive resources deployed to understand auditory information, especially under adverse conditions [150]. A key advantage of listening effort, compared to, e.g., speech reception threshold (SRT) tests is, thus, that it can be assessed in supra threshold conditions where speech remains largely intelligible. As participants spend the majority of their day in listening scenarios which might be complex, but do not have a low SNR (e.g. +5-+10 dB) [497] such investigations might resemble real-life listening performance more closely and thus influence the validity of the results (see Section 1.2). The underlying principles, influencing factors, and paradigms are presented in the following.

Theoretical Frameworks of Listening Effort

Several theoretical accounts provide a framework for understanding listening effort in complex acoustic scenarios. The two most widely recognised models are probably the Ease of Language Understanding (ELU) model by Rönnberg et al. [391] and the Framework for Understanding Effortful Listening (FUEL) by Pichora-Fuller et al. [361]. Other, more recent models exist [188, 34], but these will not be the focus of the present thesis.

The ELU model [389] provides a complementary perspective that links signal degradation to cognitive processing demands (see also [391, 390, 392]). According to the ELU, speech input is normally processed rapidly, automatically, and multimodally by matching the auditory signal to phonological representations stored in long-term memory. When the signal is distorted, e.g., due to background sounds, reverberation, or hearing loss, this automatic matching is compromised, leading to a mismatch that triggers explicit, attention-demanding processing supported by working memory. Listening effort here reflects the additional cognitive operations required to resolve these mismatches and reconstruct the received information. In this view, listening effort goes beyond speech intelligibility; words can for example be understood correctly in two different conditions, but the effort put into this understanding was different [390].

In the FUEL [361], effortful listening is embedded in a broader resource-allocation model that emphasises the interaction of task demands, available cognitive resources, and motivational factors. Here, listening effort is defined as the "deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task" [361, p. 10]. Within the FUEL, listening effort is not simply influenced by signal degradation. Rather, it is a strategic deployment of limited mental resources in response to situational challenges and personal goals. The model is rooted in classical limited-resource assumptions [65, 230], which suggest that the cognitive resources available to fulfil a task are limited. Performance may deteriorate when concurrent tasks draw on overlapping resources. The model thus predicts that in more difficult listening conditions (e.g., lower SNRs, higher reverberation, or more complex maskers), more cognitive resources will be allocated to the (target) speech task, at the expense of other processes. This assumption forms the basis of behavioural dual-task task measures of listening effort, which will be introduced in Section 1.7.2.

Factors Influencing Listening Effort

Listening effort has been shown to be influenced by numerous factors, e.g., participant-related factors or properties of the acoustic scenario.

One example for a participant-related factor is the presence or degree of hearing impairment [241, 346, 361, 440], hearing aid algorithms [40, 345, 365, 486], the participants' age [103, 415], or whether the content of interest was presented in the participants' first language or not [358]. Target speech which is unintelligible, e.g., due to distortion of the signal [245, 486] or - in presence of background sound stimuli - due to unfavourable SNR conditions [9, 194, 247, 345, 365, 443, 486, 494, 503], have been shown to increase listening effort. But not only the target intelligibility, but also the content of the background stimuli or masker may contribute to changes in listening effort: across a broad range of studies, speech maskers have been shown to elicit higher listening effort than non-speech maskers, even when SNRs are adjusted to equate intelligibility. This has been demonstrated for single competing talkers [241, 240, 346, 440, 469, 470, 487], two-talker babble [103, 141], and multi-talker babble with three or more voices [252, 344, 459]. When directly comparing different numbers of speech maskers at a matched SNR, effort tends to increase as the number of competing voices decreases [213, 487]. Fewer talkers preserve more coherent prosodic and linguistic information and thereby create stronger informational masking (see Section 1.3), while also offering more temporal "dips" for glimpsing the target (energetic benefit [233]). Both factors may increase the cognitive demands associated with selectively attending to and tracking the target speech. A more in-depth discussion of the effect of background

stimuli on listening effort is provided in Chapter 5 [125]. Room acoustic reverberation of the scenario can contribute to increased listening effort (e.g., [131, 197, 254, 276, 363, 382, 398, 415]). Such room acoustic effects are discussed in more detail in Chapter 6 [124]. This list of influencing factors, by no means exhaustive, illustrates the multifaceted nature of the concept of listening effort. Consequently, it follows that there are a multitude of methods by which listening effort can be measured.

Measurement Approaches

Multiple approaches exist to quantify listening effort, which can be categorised into subjective, physiological, and behavioural measures [303].

Subjective measures of listening effort involve questioning participants with regard to the perceived listening effort, i.e., the effort needed for a listening task in a certain scenario. Many studies use the NASA Task Load Index (NASA-TLX) [177] questionnaire, which was originally developed to examine subjective workload across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Even though it does not specifically evaluate listening effort, some dimensions of the NASA-TLX have been interpreted as an indicator of listening effort in hearing related studies (e.g., [72, 106, 282]). The NASA-TLX is also often related to more general cognitive load or task load in VR-based tasks (e.g., [29, 145, 189]). The Speech, Spatial and Qualities of Hearing Scale (SSQ) questionnaire [152] was developed to assess the influence of hearing impairment across domains such as competing contexts, spatial hearing, and stream segregation with regard to ease of listening, naturalness, and clarity. This questionnaire includes 49 items on three main subscales (Speech, Spatial, Qualities), but shorter versions exist (e.g., [327]). Excerpts of the original SSQ or the shorter versions have been taken as an indicator of listening effort (e.g., [9, 42, 72]). Seitz et al. [415] developed a listening effort questionnaire specifically for children, which can also be used to assess listening effort in adults. This questionnaire is based on the assumption that listening effort is a multidimensional concept, influenced not only by perceived effort, but also task success and general well-being (see [140]), and thus consists of three questions targeting listening effort, success, and well-being, respectively. Krueger et al. [244] developed the Adaptive Categorical Listening Effort Scaling (ACALES) procedure to assess listening effort on a rather straightforward scale from "no effort" to "extreme effort". The aspect which differentiates this procedure from other questionaries is that the given rating is used to directly inform the next condition presented to the participants, similar to adaptive SRT measurement procedures. Many studies use the main question of the ACALES in isolation to assess listening effort of

non-adaptive acoustic conditions (e.g. [186, 133]). Apart from other multi-item questionnaires (e.g., [9]), some studies directly ask for listening effort, similar to the ACALES questionnaire (e.g., [40, 192, 280]).

Subjective measures are straightforward to administer and have proven sensitive to changes in listening conditions. However, they rely on retrospection and self reflection, and may be influenced by individual reference frames and response biases. To eliminate such biases, an alternative to self-reports are physiological indicators of listening effort. Physiological indices can capture listening effort continuously (instead of at a singular time point after finishing a condition) and without requiring the participant to respond to - potentially immersion-breaking - tasks or questionnaires. One example for this is pupillometry, where the pupil dilation is monitored (e.g., [239, 240, 247, 346, 470, 486, 503]). Neural measures, such as neuroimaging, provide complementary information about cortical processing under effortful listening (e.g., [40, 106, 254, 470]). Other autonomic indices (e.g., heart rate [282, 375], skin conductance [282], or muscular activity [282]) have also been explored. Such approaches however often need elaborate equipment and calibration, not only of the apparatus itself, but also of the presented scenario. HMDs can, e.g., complicate the assessment of listening effort via neural measures (e.g., [442]) and changes in the lighting of the VR scenario can lead to responses by the pupil (e.g., [111]).

As a third approach, listening effort is evaluated based on the participants' behaviour. A common behavioural indicator of listening effort is the response time in the listening task [150, 168, 192]. Further, behavioural listening effort is often operationalised in dual-task paradigms ([150, for a review]). Here, participants perform a primary task which relies on listening (e.g., speech recognition or comprehension) while simultaneously engaging in a secondary, typically non-linguistic task. If possible, these tasks are administered in isolation (single-tasking) to obtain baseline performance, and simultaneously (dual-tasking). Participants are instructed to prioritise the primary task over the secondary task. Performance decrements or increased response times in the secondary task in dual-tasking relative to single-task baselines are interpreted as indicators of increased listening effort [150, 248]. This approach directly implements the limited-capacity assumption [65, 230] which postulates that the total cognitive resources available to a participant when performing a task are limited. As the primary listening task becomes more demanding, more resources are allocated to it, leaving fewer resources for the secondary task. Secondary tasks vary widely and include simple reaction-time tasks, visual or tactile discrimination, memory tasks, or decision-making tasks (e.g., [135, 336, 415], see [150, for a review]).

Most dual-task studies rely on relatively short speech materials (e.g., isolated syllables, words, or sentences [150]), which might evaluate processes different

from these involved in following continuous speech. An example for a dual-task paradigm which involves processing of running speech is using the HTR task [135, 406] (see Section 1.7.1) as the primary task. The HTR has been evaluated in combination with a secondary number categorisation task [135, 320] as well as a secondary vibrotactile pattern categorisation task [320, 399]. Because these secondary tasks do not involve linguistic processing, such tasks are assumed to be favourable because, while they may be subjected to general effects of background sounds due to stress and heightened cognitive effort [448], they allow for appropriate separation from the primary listening based task. The dual-task consisting of the HTR as the primary task and a vibrotactile secondary task is particularly well-suited for use in audiovisual and VR scenarios, as ECAs can be animated to speak the target texts, conveying co-verbal cues (e.g., lip movements, gaze, gestures), and the vibrotactile secondary task can be administered with handheld controllers while not interfering with the visually presented scenario [114, 115, 399].

The present thesis employs this implementation of the paradigm in Chapters 5 [125] and 6 [124] due to several reasons: the primary HTR task allows for examining memory effects in a more complex way than the serial recall. Further, the co-examination of listening effort and memory may shed a light on the multifaceted nature of listening performance.

Taken together, the considerations in the previous sections regarding source, environmental, visual, and task complexity, as well as the choice of memory and dual-task paradigms, define the methodological basis of this thesis. In the following, an overview is provided on how these elements are combined in this work to systematically increase audio(visual) complexity and evaluate its impact on listening performance.

1.8 Thesis Outline and Objectives

The primary objective of this thesis is to assess the impact of increased auditory and/or audiovisual complexity on listening performance. Building upon the complexity increments discussed in the previous section, this thesis focuses specifically on increasing complexity in working memory tasks. Starting from the classic serial recall task, which is traditionally examined without spatial audio reproduction and with stationary stimuli, this thesis implements different characteristics of complexity according to the framework by Weisser et al. [484]. This is done by building up to a complex audiovisual scenario integrating background sound complexity, spatial distribution of sources, reverberation, provision of spatial and co-verbal cues, and paradigm complexity by co-examining working memory and listening effort. Each of the chapters presents an experiment which varies one

component of complexity, assessing how the component influences the respective measure of listening performance.

Chapter 2 [128] presents two studies which extend upon the traditional serial recall task by introducing and examining spatial sound reproduction. The spatial location of the target talker changes between items, simulating the attentional demands of following a group discussion among spatially separated speakers in the presence of two static spatialised background talkers. Presentation rate and background sound are varied, specifically examining meaningful and pseudo-speech sounds.

Chapter 3 [126] introduces spatial position changes also to background sounds: background stimuli are presented from fixed or varying spatial locations and examined using verbal and visuospatial serial recall tasks to enhance understanding of cognitive representations of spatial stimuli.

Visual information is added in Chapter 4 [127] where the influence of the display device (computer monitor vs. HMD) and audiovisual incongruences (spatial and voice incongruence) are examined in an audiovisual version of the serial recall task with embodied talkers.

Building upon these aspects of complexity, more specifically by using VR presentation of embodied talkers as well as spatial target and background sounds, Chapter 5 [125] evaluates differences between meaningful and pseudo-speech sounds (see also Chapter 2 [128]). Further, the paradigm is changed from the classic serial recall towards the more complex HTR task as a dual-task paradigm in a VR.

Combining all these aspects of acoustic, visual, and task aspects, the most complex experiment is presented in Chapter 6 [124], which examines various room acoustic conditions within the HTR dual-task in a crowded, fully animated audiovisual scenario.

How these aspects of complexity contributed to the findings is assessed in Chapter 7, which summarises and discusses these multifaceted experiments. The studies are further contextualised within the complexity framework by Weisser et al. [484] and the central research question is addressed: How do these increments of audio(visual) complexity influence listening performance?

2

Spatial target source alternations in the serial recall

This Chapter was published as an original research article in the journal *Scientific Reports* as part of the collection *Auditory processing and perception*:
C. A. Ermert, M. Yadav, J. E. Marsh, et al. "Serial recall in spatial acoustic environments: irrelevant sound effect and spatial source alternations". In: *Scientific Reports* 15.1 (2025) [128].

2.1 Abstract

This study investigated serial recall performance in a complex acoustic scene that included spatialized background sounds and location changes within the target sequence to reflect real-life challenges. The focus is on two effects: the irrelevant sound effect (ISE) and the spatial-source alternation effect (SSAE). Both represent impairment in short-term memory performance of to-be-remembered items: the ISE due to irrelevant background sounds, and the SSAE due to location changes within the target sequence. Although distinct, these effects typically occur together in real-world settings, e.g., listening to multiple speakers in noise, but have not been investigated together yet. Building on the theoretical frameworks of these two effects, this study combines principles from both the ISE and the spatial-source alternation effect (SSAE) as a step towards enhancing acoustic complexity in established cognitive tasks. Experiment 1 examined auditory-verbal serial recall using spatially alternating target digits presented at a typical rate (1 item/1s), with either meaningful or meaningless background speech. Results showed an ISE, with meaningful speech causing greater disruption, but no SSAE - possibly due to either the presentation rate or the spatialized audio scene. To further clarify this, Experiment 2 was conducted with a faster presentation rate (1 item/350ms) consistent with a previous study, and more spatial target locations (monotic, $\pm 10^\circ, \pm 45^\circ, \pm 60^\circ, \pm 90^\circ$). An SSAE was revealed for all locations.

These findings suggest that the SSAE may mainly be modulated by the presentation rate - given the spatial separation is audible and only emergent at rapid location changes - questioning its applicability to naturalistic listening scenarios. As an attempt to bridge the gap between controlled laboratory settings and more complex listening tasks, these findings help explain how cognitive systems manage competing demands in real-world auditory environments, such as separating speech streams in noise.

2.2 Introduction

Auditory distractions are pervasive and can significantly disrupt our ability to process and remember information [26, 199, 281, 223], often without us fully realising it [17, 36]. When following multi-talker conversations in noise, listeners have to regularly refocus their attention on the active speaker and suppress potential background noise, including competing conversations, while simultaneously processing the spoken content of interest. Within cognitive psychology, auditory distraction has been widely studied and has provided key insights into the mechanisms of attention, memory, and perception [396, 224]. However, much of this research has focused on static auditory settings [117, 135], where the sound sources remain stationary and the effects of dynamic location changes that are typical in everyday settings - refocussing attention across active speakers - are not fully considered. Consequently, much less is known about the effect of changes in the location of sound sources on memory performance of spoken content. Understanding these dynamics is essential both from a theoretical point of view but also for designing environments and systems that minimise distraction and optimise performance.

A well-established paradigm in this context is the serial recall test, in which participants have to remember and recall a sequence of verbal items, e.g., digits or words, in the presentation order [207]. Serial memory of aurally presented items is often taken as a proxy for the ability to listen to speech and maintain the content for subsequent recall [404, 30]. However, the restricted dynamism of the auditory scenes in terms of the location and direction of the sound sources typical to studies of this paradigm [404, 227, 81, 401, 174, 447, 238] limits the generalization of findings to everyday communication settings, such as listening to conversations between spatially separated talkers under competing noise.

This study investigates how location switches in target speech and spatialised background noise impact short-term memory performance of aurally presented verbal information, as measured within the serial recall paradigm. The influence of background noise and location switches in target speech on memory performance can be modeled with the ISE and SSAE, respectively. The ISE [117] describes

how task-irrelevant background sounds disrupt the recall of to-be-remembered items. The SSAE [202, 200] (sometimes alternating-ear-effect) describes how location changes during the presentation of to-be-remembered items impedes performance compared to static presentation. While both the ISE and the SSAE represent disruptions in short-term memory performance for auditory material and are both usually examined using verbal serial recall tasks, they involve distinct mechanisms and contexts. The ISE arises from task-irrelevant auditory stimuli such as background speech, whereas the SSAE concerns task-relevant auditory stimuli, specifically focusing on spatial changes in the presentation of the to-be-remembered items. The theoretical concepts and findings underlying ISE and SSAE will be explained in more detail in the following.

The ISE [117] has proven to be a robust finding, enhancing our theoretical grasp of auditory cognition [224, 117, 89, 335] and providing practical implications for the acoustic optimization of learning and working environments like classrooms and open-plan offices, where cognitively demanding tasks have to be performed in the presence of noise [500, 264, 171]. According to the duplex-mechanism account [199], the ISE is characterised by two primary mechanisms: variations in the spectral content of the task-irrelevant sound over time (changing-state effect) and its capacity to capture attention. Changing-state sounds (e.g., sequences of syllables or letters like "A-B-C-D", words, running speech, backward speech), as opposed to more steady-state sounds (e.g., noise, pure tones, or repeated letters like "A-A-A-A"), contain order information arising from spectro-temporal changes in the auditory signal which competes with the order information in the target sequence of the serial recall, disrupting serial short-term memory performance [293]. The attentional capture mechanism, by contrast, occurs when a salient feature in the background sound, such as meaningful speech, alerting words like one's own name, or an unexpected deviant [199] (e.g., a unique item in a sequence like "A-A-B-A-A"), momentarily diverts the listener's attention away from the primary task [199].

Most studies investigating the ISE use serial recall of visually presented verbal items (e.g., words, letters, digits). Compared to visual item presentation, research using auditory presentation is limited. This can partly be attributed to the potential of target items being masked by the background noise when presented auditorily. It is hence crucial to ensure that background noise does not affect the perception or intelligibility of the presented items, as any performance decrement could otherwise be attributed to inaudibility/masking rather than memory-related effects [227]. Moreover, auditory presentation of target items is highly relevant to many everyday scenarios (e.g., communication settings), wherein both the target and noise occur in the auditory modality. Therefore, the effects of auditory presentation warrant deeper exploration in studies of the ISE,

in particular the effects of spatialised background noise and target items. Such investigations can bridge the gap between controlled experimental conditions and the complexity of real-world environments.

In the few studies where the ISE has been investigated using serial recall tasks with auditory presentation, several types of task-irrelevant sounds have been used. These have included sequences of letters or syllables [227, 81, 401], intelligible but unrelated words [174, 447], running speech in the listener's native [401] or non-native language [238], and music [404]. Most of these studies used nonspatial audio reproductions, such as headphones or mono/stereo loudspeaker setups. This differs from realistic scenarios, where sounds - both background and target - are typically accompanied by spatial information. Such information may include differences in intensity and time of arrival of a sound across the ears and direction-dependent changes in the spectral content of the sound, which are evaluated by higher auditory processing to analyse and navigate complex auditory scenes [44]. Spatial audio reproduction of background sounds has been employed in very few visual-verbal serial recall studies (e.g., with spatial loudspeaker setups [225, 170, 502, 501, 383] or virtual acoustic scenes presented via headphoness [328]) and even fewer auditory verbal serial recall studies (e.g., with virtual acoustic background scenes and non-spatialized target presented via headphones [500]). However, there are no previous studies that have included realistic spatial representation of background sounds and target items.

Another challenge in everyday listening environments is spatial changes in target signals, since target speakers are often spatially separated, requiring listeners to dynamically allocate attention across spatial locations [135]. This provides a strong motivation to investigate the effect of spatial changes in a target sequence on short-term memory performance, which is studied in the context of the SSAE [202, 200]. In this phenomenon, a spatial illusion has traditionally been created by presenting target items alternately to each ear via headphones (i.e., monotically), leading them to be perceived as originating from two distinct spatial locations (although still being localised in the head). While this reproduction technique has limitations in terms of spatial plausibility, it serves as an important initial step towards exploring the influence of spatial target signals in verbal serial recall.

Two prominent theoretical accounts have been proposed to explain the mechanisms underlying the SSAE: the item-encoding account and the perceptual-motor account. The item-encoding account argues that switching attention between spatial locations [456] (or voices [158, 296, 318, 338, 441]) during encoding - especially at fast presentation rates - places strain on cognitive resources, impairing recall. In contrast, the perceptual-motor account [200, 201] emphasises a mismatch between how listeners perceive and group incoming information and the demands

of the recall task. For instance, if digits alternate between left and right locations (e.g., left–right–left–right), listeners may group them into two separate perceptual streams based on location. However, the task still requires a single, linear reproduction of the sequence in its original order. This creates a "mapping conflict": the perceived structure of the input does not match the required structure of the output. The motor component of this account refers not to overt motor execution, but to the challenge of translating two competing perceptual streams into one ordered response. Both accounts predict disruption under spatial alternation, but they differ in where the disruption arises - at encoding or during output planning.

While our approach draws heavily from research on auditory distraction and verbal serial recall [199], the current study also intersects with broader findings in auditory cognition. For instance, spatial separation of competing sources has long been known to aid target identification in complex auditory scenes [110, 234, 67], and changes in spatial and acoustic cues can influence auditory streaming and attentional allocation [43, 285]. However, few studies have explored these dynamics in the context of short-term memory paradigms. By combining principles from auditory distraction research with more naturalistic spatial manipulations, our study aims to combine spatialized background and target sound sources within the serial recall paradigm to imitate more complex and realistic auditory scenes in cognitive research. The present paper reports two experiments: Experiment 1 examines serial recall performance in a simplified conversational scene with spatial target and background sources with background speech varying in semantical content as well as location changes, applying principles from both the ISE and SSAE. As a follow-up study designed to assess the mechanisms underlying the SSAE, Experiment 2 explored the emergence of the SSAE without background sounds, and with target sounds presented at faster rates and from a larger number of spatial locations.

2.3 Experiment 1

This experiment aims to investigate verbal serial recall performance with spatial auditory target and background sources, as well as location changes within a digit sequence; reflecting an analytic description of a natural conversation. Based on this, we designed a spatial setup of target and background sound sources to mimic a multi-talker conversation in noise, where two talkers are the target talkers and two separate talkers emit background speech. Three target location patterns were evaluated based on Experiment 3 by Hughes et al. [200] in the context of SSAE: no location changes, regularly alternating changes, and random changes between two fixed target locations. Instead of the monotic playback employed by Hughes et al. [200] (i.e., presenting the sequence to one ear at a time), sound

is spatialized using an HRTF. Monotic playback can be disadvantageous due to in-head localization where both target and noise stimuli seem to originate from "inside the head" with ambiguous directional separation. In contrast, simulating virtual spatial sound sources using HRTFs increases spatial resolution and enables externalization for listeners. This enhances sound source localisation and the ability to distinguish between competing auditory signals [44] and is an established procedure in cognitive research [135, 264, 500]. The three target patterns were combined with three background noise types based on the ISE study by Yadav et al. [500]: silence, semantically meaningful and intelligible speech, and speech that did not convey meaning to the listener.

The voice delivering the target sequences did not change with the location, as would occur in a real conversation. While having distinct voices for each target location more closely resembles real-life scenarios, it has been shown that voice changes within a target sequence disrupt verbal serial recall of spoken distractors (talker-variability effect (TVE)) [202]. Hence, if we had introduced distinct voices for each target location, the effect of alternating voices and locations could not have been disentangled [201]. Since a primary aim is to investigate the effect of spatialized target and background sounds within the serial recall, this trade-off in terms of realism was considered acceptable.

2.3.1 Method

Participants

$N = 27$ participants (16 female, 11 male), aged 21-39 years ($M = 26.34$, $SD = 4.47$) were recruited for the experiment via posters displayed at the authors' institutes and via mailing lists. All participants had German as their first language (self-reported) with normal hearing and (corrected-to) normal vision. Normal hearing was defined, according to the WHO criteria [498], as a maximum of 25 dBHL (decibels Hearing Level) between 250 Hz and 8 kHz. Hearing was examined using an AURITEC Ear 3.0 audiometer and Sennheiser HDA300 headphones, employing pulsed pure tone ascending audiometry. Vision, aided if necessary, was tested with a Snellen chart (20/30) [435]. The experiment procedure was pre-approved by the Ethics Committee at the Medical Faculty of RWTH Aachen University (EK396-19). Written informed consent was obtained from all participants before the experiment. Each participant received 10€ as compensation for their time and involvement. The study was conducted in accordance to the rules of conduct stated in the Declaration of Helsinki.

Paradigm

Short-term memory performance was examined with the verbal serial recall using auditory digit presentation. Each trial consisted of four phases: countdown, presentation, retention, and serial reconstruction. In the countdown phase, three rectangles decreasing in size were displayed sequentially on a computer monitor. Each rectangle was followed by a 500 ms pause, resulting in a total countdown time of 1.5 s. In the presentation phase, a sequence of spoken digits (1-9) was played through headphones at a rate of one digit per second. Each spoken digit had a duration of 600 ms, followed by a 400 ms interstimulus interval (ISI), resulting in a total presentation duration of 9 s. During this phase, the computer monitor displayed only a play button, with no visual representation of the digits. The digit sequences were designed such that each digit (1–9) appeared exactly once within a sequence. Furthermore, only two consecutive steps of one were permitted (e.g., 1–2–3 or 3–2–1), avoiding longer consecutive progressions (e.g., 1–2–3–4 or 4–3–2–1) [80]. Additionally, each digit appeared in each serial location approximately equally often across trials. The retention phase lasted 3 s during which the participants retained the presented sequence. In the reconstruction phase, the digits 1–9 were displayed in randomized order in a 3×3 grid on the computer screen (Calibri, 80pt$\hat{=}$ 1.4 ms). Participants were instructed to click the digits in the exact order they were presented. Once a digit was clicked, it disappeared from the grid and corrections were not permitted. After completing the sequence, the next trial began automatically. There was no time limit for completing the reconstruction phase.

The chosen presentation rate of one item per second is well established in serial recall studies [227, 81, 401, 174, 447, 238, 404], while slower presentation rates can also be found [221, 108]. In their study on the SSAE, Hughes et al. [200] used a presentation rate of 1 digit per 350 ms, to facilitate spatial stream segregation. Conversational alternations in naturalistic settings occur at a much slower rate than once every 350 ms [385, 269]. Thus, a presentation rate of one digit per second was selected instead.

Audio Reproduction

A simulated auditory scene replicating a conversational environment was created using Virtual Acoustics [209]. Binaural spatial signals were generated by convolving the auditory stimuli with a generic HRTF of an artificial head [408]. Two target sources, playing back digit sequences for the auditory serial recall task, were located at $\pm 45^\circ$ on the horizontal plane (see Fig. 2.1) in a virtual free field without room acoustics.

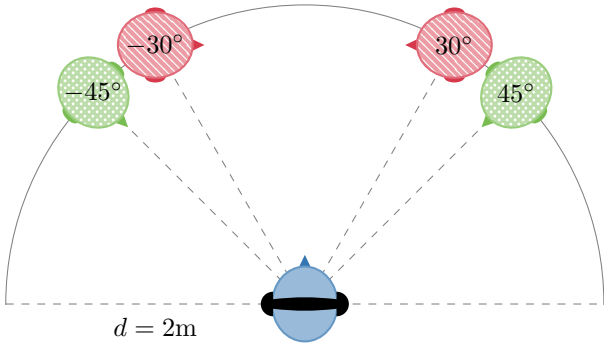


Figure 2.1: Setup in Experiment 1. Target sources ( ) were placed virtually at ±45°, noise sources ( ) at ±30° on the horizontal plane at a distance of $d = 2$ m from the receiver ( ) and played back via headphones.

The receiver (participant) and the two target sources (at ±45°) formed an equilateral triangle with two additional noise sources (at ±30°) on the horizontal plane, resulting in a 15° angular offset between each target and noise source pair. Based on the minimum audible angle, i.e., the smallest angular separation at which two sound sources are perceived as arising from two separate locations [317], for non-individual HRTFs [424], this spatial arrangement provides sufficient separation for the participant to perceive distinct target and noise sources. In this study, all virtual sound sources were located 2 m away from the receiver (i.e., listener location), a distance considered comfortable for interactions with unfamiliar individuals [173]. To simulate a conversational scenario, the target sound sources were oriented toward the receiver, while the noise sources were oriented toward each other. This arrangement mimicked a conversation between the noise sources.

Playback of all sounds was delivered through Sennheiser HD650 headphones with individualised headphone equalization [298], using an RME Fireface UC soundcard. The receiver's location was tracked using Optitrack and Motive software (v1.10.0), enabling dynamic reproduction to counteract head movements. This approach not only resembles real-life listening conditions but also enhances localization accuracy compared to static reproductions [341].

Stimuli

For the target stimuli, recordings of German digit words spoken by a female voice (female b) [339] were used. For the *noise*, three types of stimuli were created: *silence*, *meaningful speech*, and *pseudo-speech*.

The *meaningful speech* condition consisted of intelligible German sentences taken from the AuViST database [129], which contains 34 family stories narrated in 10 sentences each. Recordings are available for both female and male speakers. Sentences with digit words were excluded. The remaining sentences were presented in a randomised order [375]. While coherent narratives might have elicited higher ecological validity, this approach ensured stimulus control and variability, while preserving acoustic features of running speech, such as prosody and temporal dynamics. One background source used a female voice, and the other a male voice, balanced across participants to control for voice-related effects.

The *pseudo-speech* was designed to convey no meaning to the listener and be unintelligible (in contrast to *meaningful speech*) while maintaining the same temporal-spectral structure as the *meaningful speech*, to control for differences in energetic masking. For that matter, pseudo-speech was generated by cutting the sentences from the *meaningful speech* into syllables and randomly rearranging them to create unintelligible sequences for both female and male voices in Audacity (v3.1.3). A 10 ms cross-fading was applied between syllables to avoid artifacts.

Both noise sources emitted their signal simultaneously and continuously, without breaks between sentences. Although this approach differs from a realistic background conversation, it was chosen to avoid spatial-temporal changes that might exceed those caused by the temporal structure of the speech. Additionally, the continuous emission ensured a symmetrical design, maintaining consistent SRM benefits for the target source.

All stimuli were calibrated using the HEAD acoustics HMS III artificial head. Target stimuli were calibrated to 60 dB(A) and the background stimuli were adjusted to 57 dB(A), measured as the power sum across both ears at the receiver's location. Due to the spatial separation of the target and background sound sources, an additional spatial release from masking benefit [325, 66] of up to 6 dB can be expected [229].

Target Location Patterns

Three *patterns* were implemented for the target location changes: no location changes (*single*), regularly alternating changes (*alt-reg*), and random irregular alternations (*alt-irr*) between two fixed target locations. These were adapted from Experiment 3 in Hughes et al. [200], with the difference being that their study used eight digits, whereas this experiment used nine digits. In the *single* condition, all digits in the target sequence were played back from either the left or right side, with the side balanced across trials. In the *alt-reg* condition, the target source location alternated regularly between left and right for each

digit (e.g., LRLRLRLRL), with the starting direction (left or right) balanced across trials. In the *alt-irr* condition, the target source location alternated in an irregular, pseudo-random manner. A maximum of two consecutive digits could be presented from the same side, and the total number of digits presented from each side was balanced (five from one side, four from the other). The side emitting the majority of digits was also balanced across trials.

Procedure

Each participant completed the experiment in an individual session inside a sound-insulated hearing booth. The session began with audiometric testing, a Snellen vision test, and headphone equalization measurement. Only participants who passed these screening procedures proceeded to the main experiment. Participants were seated on a wooden chair (seat height: 0.44 m) located 1.0 m away from an LG Flatron L1710B, with the monitor's center at a height of 1.1 m. Written instructions were displayed on the monitor. Participants were explicitly instructed not to vocalise the digits, use their fingers as aids, or rely on other movements for support. To familiarise themselves with the procedure, participants first completed a practice block, consisting of six silent trials in the single condition. Three practice trials were presented from $+45^\circ$ and three from -45°, balanced across trials. Following this, participants listened to the *meaningful speech* and *pseudo-speech* and completed one practice trial each, with the order of conditions balanced. To ensure intelligibility, participants confirmed they could understand all the digits before proceeding to the main experiment. The main experiment comprised nine blocks of eight trials each, for a total of 72 trials. Each block included a unique combination of the variables *pattern* (*single*, *alt-reg*, *alt-irr*) and *noise* (*silence*, *meaningful speech*, *pseudo-speech*). In blocks with background noise, the noise signal played continuously without breaks throughout the block. The order of block presentations was counterbalanced across participants using a Latin square design. Participants were allowed to take voluntary breaks of no fixed length between blocks. Additionally, a mandatory 2 min break was administered after the fifth block. The total duration of the session, including the screening process, was approximately 60 min. The experiment was implemented using MATLAB (v2023a), with the ITA-toolbox [41].

Data analysis

The statistical analyses were conducted using R (version 4.2.2). The fixed effect, i.e., dependent variable, was the recall accuracy in the task, defined as the proportion of digits recalled at the correct serial location. The independent variables were the *noise* conditions (*silence*, *meaningful speech*, *pseudo-speech*), and the tar-

get *pattern* (*single*, *alt-reg*, *alt-irr*). Separate Bayesian generalised mixed-effects models were created using the R package brms (version 2.18) [79]. In each model, the random effects were incorporated as independently varying intercepts and independently varying and uncorrelated slopes across the serial locations, and the independent variables (*noise* and *pattern*) for the participants. A zero-one inflated Beta distribution with a logit link function was used, suitable for proportion data bounded between 0 and 1; several relevant distributions (e.g., Gaussian) were used for comparison.

Models were built incrementally, starting with an intercept-only baseline model, and progressively adding random effects, independent variables and their interactions. Weakly informative priors (in the logit scale) for the distributions per factor and their interactions were specified, and sampling included 16000 iterations (four independent chains of 5000 samples each, with the first 1000 discarded as warm-up). Model performance was compared using leave-one-out (LOO) cross-validation, evaluating differences in expected log pointwise predictive density (elpd) and the standard error of the elpd of the model [429]. Posterior predictive checks performed per model confirmed the models adequately captured the observed data distributions. Orthogonal contrasts were conducted across the factor levels, with results summarised using medians and 95% credible intervals (CIs), where statistically robust effects refer to the 95% CI not spanning zero. The Bayesian 95% CI, calculated using the highest density interval (HDI) of the posterior probability distribution (PPD), provides the interval within which 95% of the PPD lies. We use the proportion of the 95% CI inside the region of practical equivalence (ROPE) to determine whether an effect is meaningful/significant. ROPE is the range signifying an effect of negligible magnitude and is conceptually similar to the null hypothesis in frequentist statistics. The range for the ROPE was specified as $\pm 0.1 \cdot SD_{\text{dependent variable}}$ [246], and calculated using the R package bayestestR [288].

2.3.2 Results

The distributions of data are illustrated in Fig. 2.2. Based on the LOO comparisons across models, the model including only *noise* performed best, outperforming other models that also included *pattern* or an interaction between the predictor variables. This indicates that *pattern* was not a robust predictor of recall accuracy in the serial recall task. As summarised in Tab. 2.1, recall accuracy was significantly higher in the *silence* condition compared to the average accuracy in the *speech* conditions (*meaningful speech* and *pseudo-speech*). Furthermore, recall accuracy was significantly lower in the *meaningful speech* condition than in the *pseudo-speech* condition. Although the contrasts across the *pattern* condi-

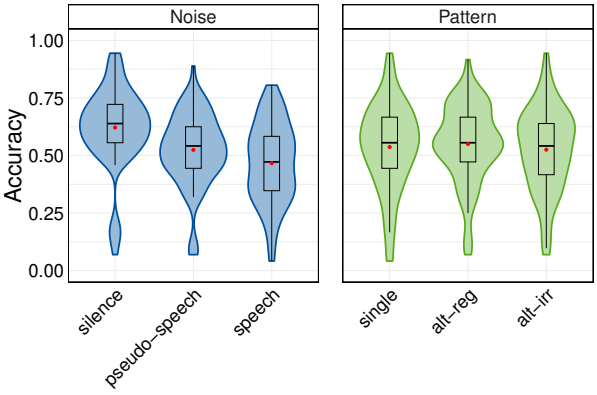


Figure 2.2: Results of Experiment 1. The Accuracy distribution (averaged over the serial location) is displayed for the independent variables *noise* and *pattern*. The mean is indicated with a red dot (·).

tions were not statistically significant, they are presented here for reference and potential use in future studies.

2.3.3 Discussion

Two main findings can be derived from the experiment, as shown in Tab. 2.1: both *meaningful speech* and *pseudo-speech* as background noise decreased recall performance (ISE), with a stronger impact of *meaningful speech*. Secondly, the *pattern* of location switches in the target signal did not affect recall performance (SSAE), and the factors *noise* and *pattern* did not interact.

Table 2.1: Orthogonal contrasts for recall accuracy across the *noise* conditions (top two rows; significant effects), and the *pattern* conditions (bottom two rows; non-significant effects).

Contrast	Median	95% CI	% in ROPE
silence − *speech*	0.13	[0.11,0.16]	0%
meaningful speech − *pseudo-speech*	-0.05	[-0.07,-0.03]	0%
single − *alternating*	0.003	[-0.01,0.02]	100%
alt-reg − *alt-irr*	0.02	[0.0,0.04]	80.5%

CI = Bayesian credible interval, ROPE = region of practical equivalence

The ISE for meaningful background speech was stronger than for pseudo-speech. Similar effects have been demonstrated for serial recall with visually [405, 460] and aurally [500] presented digits. One possible explanation of this effect is a difference in masking between the signals, potentially challenging the encoding of the items. Masking occurs if the audibility of a target signal is reduced when presented concurrently with an irrelevant sound and would therefore affect the perception or encoding of the items within the serial recall task. A distinction is made here between informational and energetic masking. Energetic masking is observed if the target and irrelevant signal contain energy in the same frequency range, causing parts of both signals to be inaudible, while informational masking is a higher-level cognitive process that occurs when both signals are comprehensible, making it harder to separate the target from the noise [325, 66, 110]. While it can be assumed that *pseudo-speech* primarily elicited energetic masking, in the *meaningful speech* condition, informational masking was present in addition [74]. It is, however, unlikely that the performance difference only stems from this masking difference, possibly impeding performance in the encoding phase of the serial recall, since intelligibility in all conditions was ensured and this "intelligibility" effect can be found in visual serial recall tasks as well [405, 460]. The duplex-mechanism account [199] justifies different effects of the noise signals in the recall phase: both signals elicited a changing-state effect, but the meaningful speech additionally had the capacity to capture attention. This finding underscores that the content's semantic meaning has a disruptive impact on recall of auditory verbal items, as noted in previous studies [500].

Splitting up the target sentence across two spatial locations did not impede memory performance. This implies that listeners were able to treat the spatially separated speech as a single coherent stream in the sequence reconstruction - likely because the presentation rate and consistent vocal characteristics may have facilitated re-grouping the items despite location changes. This stands in contrast to scenarios where either rapid presentation or multiple distinct talkers introduce greater perceptual load, making it more difficult to group items and potentially leading to reduced recall fidelity [422]. The traditional finding of the SSAE is often examined under different conditions with faster presentation rates (350 ms per digit [200] vs. 1 second per digit in this study) and without spatial audio (monotic [200] vs. spatialized in this study). From a theoretical point of view, both of these changes could have contributed to the absence of the SSAE. Both the item-encoding hypothesis [456] and the perceptual-motor account [200, 201] can be drawn upon to explain the absence of the SSAE due to the presentation rate. From the perspective of the item-encoding hypothesis, the slower rate may have reduced demands on encoding spatial changes, allowing sufficient time for processing and attenuating potential disruptions. The perceptual-motor account

suggests that slower rates might not create perceptual groupings strong enough to conflict with the planning and sequencing of a response; a stronger memory effect for increased rates can thus be predicted from both accounts. Although the dependency of SSAE on presentation rates has not been directly tested, slow location changes may not induce sufficient spatial disruption to elicit an SSAE. A rate dependency has, e.g., been found for alternating voices instead of alternating locations within a target sequence (TVE [318]). To explain the present findings comprehensively, the influence of rate on the SSAE should be investigated further.

The extent to which the spatial audio reproduction contributed to the absence of the SSAE is unclear since the SSAE has not been investigated together with spatial sound sources (background or target) so far. Contradictory predictions can be derived from the item-encoding hypotheses and the perceptual-motor account with regard to the effect of spatial sound and, consequently, the angle of target source separation on the SSAE. The item-encoding account attributes the SSAE to a cost in attention switching between the target source positions and/or between the ears as input channels in a more strongly-worded formulation of the account [456]. The cost of attention switching, which has been shown to be dependent on the switching angle [159], results in an angle-dependent SSAE. If switching the ear as the input channel is the driving factor, the SSAE would only emerge for monotic presentation, as spatialized audio presents the stimulus to both ears albeit with a level difference. In contrast, the perceptual-motor account would predict that perceivable spatial changes in the target sequence, even at minimal angles, would maintain SSAE effects due to "streaming-by-location", where sequences are processed as originating from distinct spatial locations [200].

While both factors - presentation rate and spatial audio reproduction - may have contributed to the absence of the SSAE in this study, their roles in increasing the auditory realism in serial recall tasks are distinct. As previously discussed, spatial audio is essential for naturalistic listening environments. If the use of spatial stimuli had led to the absence of the SSAE, this would have strong implications for designing experimental scenarios that better approximate real-world listening. However, listeners in everyday situations are unlikely to process alternations as rapid as every 350 ms. Consequently, if the SSAE is only observed at such fast presentation rates, its relevance for understanding or simulating typical real-life auditory experiences would be limited. Thus, to explain the effects of this study comprehensively, there is a need to investigate the SSAE with a faster rate and under varying spatial conditions, including larger variations in the extent of location changes, which is studied further in Experiment 2.

2.4 Experiment 2

In Experiment 1, a presentation rate of one item per second was chosen. Typically, faster rates are used to investigate the sequential integration of items into a perceptual stream [54]. To ensure that the SSAE is limited to high presentation rates and to further examine how spatial audio presentation contributed to the absence of the SSAE, Experiment 2 replicates the presentation rate of 1/350 ms from Hughes et al. [200] and extends the range of aural presentation angles. No background noise was included in this experiment to enable us to compare multiple target source locations within one experiment. The locations included the monotic condition (playing the raw digit stimulus one ear at a time), which was also employed by Hughes et al. [200]. The spatial auralization closest to the monotic reproduction places virtual sound sources at $\pm 90^\circ$ on the horizontal plane. In addition to this maximal angular change, we included a smaller separation of 20° ($\pm 10^\circ$ relative to frontal location), which is a small but audible change in location based on the expected minimum audible angle of the present system. Additionally, we chose $\pm 45^\circ$ for comparability with Experiment 1, and $\pm 60^\circ$ to expand the grid of angular sampling.

Different hypotheses can be formed for this experiment, depending on the underlying cognitive processes. The item-encoding account explains the SSAE as arising from the temporal cost associated with shifting attention, such as the left and right ears as distinct input channels [456]. When stimuli are presented at a fast rate, these attentional shifts may not keep pace with the item presentation rate. However, spatial, in contrast to monotic, presentation involves auditory information being available to both ears (with interaural level differences), so the ears no longer function as independent channels but information from both ears is evaluated. Consequently, the SSAE should only be observed in the monotic condition, where input channels remain distinct. However, it has been shown that the time required to shift auditory spatial attention between two target talkers depends on the magnitude of the angular change [159]. Thus, in a more general interpretation of the item-encoding account one could also expect the SSAE to vary with the angular distance between successive items based on the predictions of the item-encoding account.

Alternatively, if the SSAE is driven by streaming-by-location as proposed in the perceptual-motor account, then any perceived change in sound location - regardless of the size of the shift - could segment the target sequence into separate streams. While for each individual stream the order information is stored, this segmentation would disrupt the reconstruction of the original order from the different streams, potentially affecting both monotic and spatial conditions.

2.4.1 Method

Participants

$N = 30$ adults (16 female, 14 male), aged 19-44 years ($M = 27.83$, $SD = 6.05$) participated in the experiment. Recruitment, screening procedures, ethics approval, written consent, and compensation were consistent with Experiment 1.

Paradigm

The same paradigm as in Experiment 1 was used, with the exception that the item presentation rate was adjusted to match that of Hughes et al. [200]. Digits were presented for 250 ms with an ISI of 100 ms, resulting in a total item duration of 350 ms per digit. Consequently, a sequence duration was reduced from 9 s (Experiment 1) to 3.15 s.

Audio Reproduction

The methods of audio reproduction remained identical to those in Experiment 1, with the addition of more sound source locations in Experiment 2, as illustrated in Fig. 2.3. Five *angles* were examined: $\pm 10°, \pm 45°, \pm 60°, \pm 90°$, and *monotic* playback (lacking spatial information). The same dynamic playback and headphone equalization methods were applied as in Experiment 1.

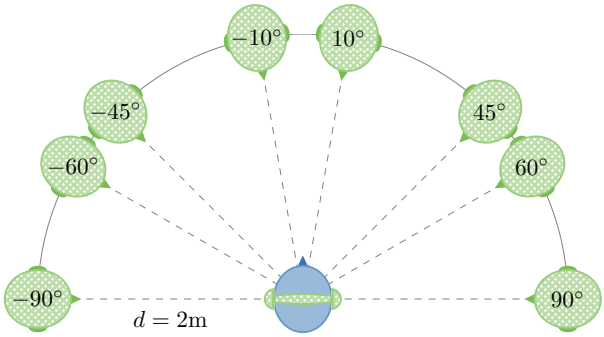


Figure 2.3: Setup in Experiment 2. Target sequences ( ) were reproduced spatially or *monotic* via headphones. Spatial target sources were placed virtually at $\pm 10°$, $\pm 45°$, $\pm 60°$, $\pm 90°$ on the horizontal plane at a distance of $d = 2$ m from the receiver ( ).

Stimuli

New stimuli were required to accommodate the faster presentation rate. Initially, the stimuli from Experiment 1 were shortened to 250 ms using Audacity's (v3.1.3)

time-compression algorithm, which preserved the original spectral distribution. However, this strong post-processing introduced substantial distortion. As a result, new recordings were made with a female speaker in a sound-isolated booth using a Rode NT5 microphone and a Zoom H6 Handy Recorder. A distance of approximately 1 m between the speaker's mouth and the microphone was maintained as in the stimuli from Experiment 1 [339]. The speaker was instructed to articulate quickly and clearly. Time-compressed recordings from Experiment 1 were used as a reference during the recording. Each digit was recorded multiple times in mixed orders to ensure consistent articulation. Post-processing was conducted in Audacity (v3.1.3), where recordings were cut and normalised following the European Broadcasting Union R128 standard, targeting an average loudness of -23 LUFS relative to Full Scale [132]. The final recordings were between 201 ms and 346 ms in duration ($M = 271.34$ ms, $SD = 37.22$ ms). These were further adjusted to 250 ms using Audacity's time-compression algorithm. Stimuli were calibrated to 60 dB(A) using the same procedure as in Experiment 1.

Target Location Patterns

To minimise participant fatigue and to focus on key contrasts between *single* and *alt-reg* conditions, the *alt-irr* condition from Experiment 1 was omitted. Only two *patterns* were employed: *single* and *alt-reg*.

Procedure

The equipment and general procedures mirrored those in Experiment 1. Adjustments were made to the training block and the main block conditions. The training block consisted of six trials with diotic sound presentation (i.e., identical stimuli presented to both ears without spatial information). The main experiment included 10 blocks, each corresponding to a unique combination of the variables *pattern* (*single*, *alt-reg*) and *angle* ($\pm10^\circ, \pm45^\circ, \pm60^\circ, \pm90^\circ$, *monotic*). The background noise as an additional factor were omitted. A balanced latin square design was used to control the order of conditions. Participants could take breaks between blocks, with a mandatory 2 min break after the fifth block. The total duration of the experiment was approximately 45 min.

Data analysis

For the two independent variables, *pattern* (*single*, *alt-reg*) and *angle* ($\pm10^\circ$, $\pm45^\circ$, $\pm60^\circ$, $\pm90^\circ$, *monotic*), the statistical analyses included the same steps as described in section Experiment 1 - Data analysis, except for using a Gaussian distribution as the family for the accuracy, and using random slopes only for the serial position. The latter two choices were based on the LOO comparisons,

where the Gaussian distribution provided a substantially better fit than the zero-one-inflated Beta distribution used in Experiment 1, and the model with random slopes for the independent variables did not improve the fit compared to the model with random slopes just for the serial position.

2.4.2 Results

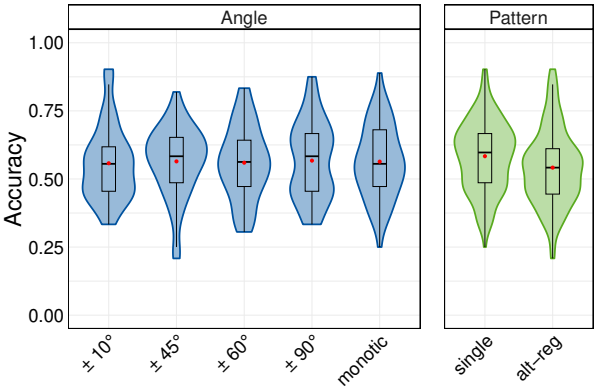


Figure 2.4: Results of Experiment 2. The accuracy distribution (averaged over the serial position) is displayed for the independent variables angle and pattern. The mean is indicated with a red dot (·).

The distribution of results is visualised in Fig. 2.4. Based on the LOO comparisons, the model including only pattern outperformed models that included angle or an interaction between the factors. The difference in accuracy between the *single* condition (all digits presented to one side) and the *alt-reg* condition (alternating locations with each digit) was statistically significant (Tab. 2.2). However, the angle did not result in statistically significant differences in accuracy.

2.4.3 Discussion

The analysis revealed that the *angle* ($\pm 10°, \pm 45°, \pm 60°, \pm 90°$, *monotic*) resulted in better performance for sequences where all digits were presented from one side (*single*) compared to sequences where the presentation direction alternated with each digit (*alt-reg*). This effect did not vary substantially across the levels of *angle*. The fact that SSAE was observed only in Experiment 2, where a faster presentation rate (1 item / 350 ms) was employed and not in Experiment 1, which used a slower rate (1 item / 1 s), demonstrates rate dependency. A similar

Table 2.2: Pairwise contrasts for recall accuracy across the *pattern* (top row; significant effect) and *angle* conditions (rows 2-10; non-significant effects).

Contrast	Median	95% CI	% in ROPE
single − alt-reg	0.04	[0.03,0.05]	0%
monotic − 10°	5.89e-03	[-0.01,0.02]	100%
monotic − 45°	-2.19e-03	[-0.02,0.02]	100%
monotic − 60°	4.40e-03	[-0.01,0.02]	100%
monotic − 90°	-3.43e-03	[-0.02,0.02]	100%
10° − 45°	-7.96e-03	[-0.03,0.01]	100%
10° − 60°	-1.42e-03	[-0.02,0.02]	100%
10° − 90°	-9.34e-03	[-0.03,0.01]	100%
45° − 60°	6.48e-03	[-0.01,0.02]	100%
45° − 90°	-1.34e-03	[-0.02,0.02]	100%
60° − 90°	-7.83e-03	[-0.03,0.01]	100%

CI = Bayesian credible interval, ROPE = region of practical equivalence

rate dependency has been detected for the varying voices between digits [318]. The emergence of SSAE at faster rates was predicted by both the perceptual-motor and item-encoding accounts. Based on this rate dependency, no significant decrease in performance due to shifts in source locations is to be expected in realistic conversational situations - relevant to the broader aim of this paper - as location changes typically occur at slower rates in everyday conversations (generally < 1/s [269]).

The lack of differences across angles aligns with the streaming-by-location explanation [200], where changes in angle are perceptually grouped into a single spatial stream when rates are fast. This finding challenges the item-encoding account's prediction that different angular separations change encoding demands and modulate SSAE effects; a strictly angle-dependent version of the item-encoding account would predict that variations in angular separations influence attentional switching demands and therefore differentially affect performance. However, more general formulations of the item-encoding framework, in which any reliable spatial alternation taxes encoding due to changes in co-verbal information, could still explain the observed lack of variation across angles [158]. In summary, since the SSAE was not significantly affected by presentation angle, it seems that once the prerequisites for SSAE are met, any perceivable source separation can be assumed to be sufficient to trigger streaming-by-location. Importantly, an SSAE can occur with both spatial audio reproduction and monotic reproduction at fast rates, reinforcing the importance of perceptual organisation.

2.5 Limitations

While the main aim of Experiment 1 was to increase the realism of the auditory presentation in the serial recall paradigm, it still represented a very controlled setting with multiple differences to real-world scenarios. Further improvements could include diversifying the target voices, incorporating room acoustic influences to increase acoustic complexity, and providing talker visualizations which would aid following the turn-taking. Furthermore, the realism of the target as well as background speech was limited. Background speech could be further improved by including coherent narratives and conversational dynamics between the background talkers, both of which were decided against in the present study to ensure control. The serial recall paradigm itself also limits the realism of the presented stimulus material (digits with pauses in between utterances). Other paradigms, e.g., with running speech [135], and partially overlapping utterances could possibly reflect naturalistic conversations better and lead to different findings. Still, in order to systematically translate established findings from memory research to more naturalistic contexts, a gradual increase in realism within an established and controlled task such as serial recall was considered an appropriate and effective approach. As for Experiment 2, an SSAE was observed for fast presentation rates. While a joint analysis of the ISE and SSAE at this faster rate would be warranted, particularly to establish a precise rate threshold at which the SSAE emerges, such an analysis was deemed beyond the scope of the current study's objectives of focusing on improving the realism of acoustic environments within the serial recall paradigm.

2.6 Conclusion

Taking inspiration from everyday conversations, which take place in the presence of background noise and where target speakers are often spatially separated, this study investigated serial recall performance applying principles from both the ISE and SSAE. Two experiments were conducted using serial recall with spatialised auditory target items, focusing on background speech (Experiment 1) and spatial alternation of the target speech stream (Experiments 1 and 2).

Our findings highlight two primary insights. First, irrelevant meaningful speech disrupts short-term memory performance more strongly than pseudo-speech, suggesting that semantic processing captures attention beyond changing-state characteristics in ways pseudo-speech does not. Second, spatial alternation within the target sequence was only shown to impair recall when items are presented at a sufficiently fast rate (here: 1 item per 350 ms), irrespective of how large the angular separation between alternating locations is.

From a practical standpoint, these results imply that in typical conversational settings - where speakers generally alternate turns at slower speeds - the impact of frequent spatial shifts on memory may be minimal. Although rapid switching between different locations can disrupt recall in a controlled lab task, such rapid alternations are uncommon in the real world. Background speech that carries semantic information reduces memory performance, underscoring the disruptive potential of meaningful speech regardless of spatial configuration.

On a theoretical level, our data suggest that once listeners detect a spatial alternation, the separation angle does not influence the degree of disruption at high item presentation rates. However, everyday conversations are rarely carried out at these fast rates. To better understand the temporal dynamics of auditory short-term memory and enhance predictions about memory performance in real conversational settings, future studies should investigate the precise cut-off rate for the emergence of the SSAE and whether further irregular alternation patterns can isolate the cognitive mechanisms behind the effect. Further, the realism of the experimental setting could be increased by incorporating factors like continuous speech, alternating voices, continuous spatial movement, room acoustics, and visual co-verbal cues. Collectively, our findings suggest that spatial separation of talkers is unlikely to affect short-term memory performance under normal presentation rates.

Acknowledgements

The authors would like to thank Chinthusa Mohanathasan and Jonathan Ehret for their valuable project collaboration, and Julia Seitz and Pascal Palenda for helping with the recordings for Experiment 2.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): SPP2236 - Project number 444724862: Listening to and remembering conversations between two talkers: cognitive research using embodies conversational agents in audiovisual virtual environments. Manuj Yadav was supported by the DFG: Project number 503914237: Effect of local architectural acoustics on dyadic conversations over short distances. John E. Marsh's contribution was supported by grants received from the Bial Foundation (Grant 201/20) and the German Academic Exchange Service (Deutscher Akademischer Austauschdienst [Grant 91875827]). Sabine Schlittmeier's contribution was supported by a grant from the HEAD-Genuit-Stiftung (Head-Genuit Foundation; P-16/10-W).

Author contributions

C.E: Conceptualization, Methodology, Software, Investigation, Data Curation, Writing (Original Draft, Review & Editing), Visualization; M.Y: Conceptualization, Methodology, Formal Analysis, Data Curation, Writing (Original Draft, Review & Editing); J.M: Conceptualization, Methodology, Validation, Writing (Original Draft, Review & Editing); S.S: Writing (Original Draft, Review & Editing), Funding Acquisition; T.K: Writing (Review & Editing), Funding Acquisition; J.F: Conceptualization, Resources, Methodology, Supervision, Writing (Review & Editing), Project Administration, Funding Acquisition

Competing interests

The authors declare no competing interests.

Additional information

Data availability: The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

3

Spatial-changing irrelevant speech in visuo-spatial and visuo-verbal serial recall

This Chapter was submitted for publication as an original research article in the journal *Memory & Cognition*:
C. A. Ermert, L. Vollmer, et al. "Spatial-changing irrelevant speech in visuo-spatial and visuo-verbal serial recall". In: *tbd* (under Review) [126].

Abstract

Real-world acoustic scenes often feature moving sources, yet research on auditory distraction often employs static auditory sources. Consequently, the effect of spatial-changing task-irrelevant sounds on both verbal- and visuo-spatial serial recall is not sufficiently understood. This paper investigates how spatially changing task-irrelevant speech affects visual–verbal and visuo-spatial serial recall. $N = 30$ normal-hearing adults performed both tasks under five background sound conditions: silence, spatial-static verbal-static, spatial-static verbal-changing, spatial-changing verbal-static, and spatial-changing verbal-changing speech. Irrelevant letters (repeated "A" in the verbal static condition and pseudo-random sequences "A"-"G" in the verbal-changing condition) were presented from either a fixed frontal loudspeaker (spatial-static) or pseudo-randomly across seven positions on the horizontal plane (spatial-changing). Verbal-changing speech reliably impaired visual–verbal serial recall relative to silence and verbal-static speech, with no meaningful additional decrement from spatially changing presentation. Visuo-spatial serial recall was unaffected across all background sound conditions. Our findings replicate the classic changing-state effect for verbal memory, but provide no evidence for a spatial changing-state effect on visuo-spatial serial recall. More generally, spatially varying irrelevant speech did not produce disruption beyond the established effect of verbal variation.

3.1 Introduction

Task-irrelevant background sound can impair short-term memory performance. A robust example is the ISE which refers to the disruption of shot-term memory performance by task-irrelevant sounds. The phenomenon was first reported by Colle and Welsh [89] using task-irrelevant background speech. Since then, irrelevant speech has been studied extensively and has proven to be a particularly potent auditory distractor (e.g., [407, 396, 175, 295, 388, 402, 238, 335, 447, 258]). The most commonly used paradigm for studying the ISE is verbal serial recall, which is a standard measure of verbal short-term memory. In this task, participants are asked to maintain and reproduce sequences of verbal items, such as letters or digits, in their original presentation order.

A central question in ISE research is which properties of the distractor determine its disruptive impact. Besides speech, non-speech sounds can impair verbal serial recall when they contain temporal-spectral variation across successive elements (so-called changing-state characteristic, i.e., variation across successive auditory elements as opposed to steady-state repetition). Thus, disruption is not determined solely by whether the distractor is speech, but also by its internal auditory-perceptive characteristics. Beyond these inherent properties of the auditory signal, another characteristic concerns how the distractor is distributed in space. In real life, for example, auditory distractions are distributed in space or often change in spatial location over time, leading to spatially complex background noise. Further, sources may move continuously (e.g., a car passing by) or in a discrete manner (e.g., multiple task-irrelevant talkers having a conversation). Thus, beyond its verbal and auditory-perceptive characteristics, irrelevant speech also varies in its spatial properties, which may be relevant for auditory distraction. Though potentially relevant, such spatial characteristics are less systematically studied than, e.g., verbal characteristics of distractors. In fact, most ISE studies have neglected spatial influences in the auditory presentation, e.g., by playing back the irrelevant sound from a fixed frontal location or via headphones devoid of meaningful spatial information. Consequently, the role of spatial information in auditory distraction has received comparatively little attention and is often limited to static background sounds (e.g., [383, 128, 500, 264, 225]). Since short-term memory is often used as a proxy for cognitive performance in applied settings such as classrooms (e.g., [265]) and open-plan offices (e.g., [170, 383, 328, 500]), understanding how spatial variations in background sound affect memory performance is therefore of both practical and theoretical importance.

This raises the question which mechanisms determine whether irrelevant sound disrupts serial recall, and how spatial variation in the distractor should be expected to affect performance. Different theoretical accounts of the ISE offer

different answers to this question. An influential one is the multi-component model by Baddeley and Hitch [25] (see also [21]), which proposes domain-specific short-term memory subsystems for verbal and visuo-spatial information, respectively. According to this model, verbal information is stored in the phonological loop. Due to automatic and involuntary processing of auditory information, irrelevant speech gains access to the phonological loop, where the target sequence of the verbal serial recall is stored [396], resulting in interference between the processing of the irrelevant stimulus and the serial recall sequence. Visuo-spatial information, in contrast, is stored in the visuo-spatial sketchpad, which is functionally separated from the phonological loop [25]. Importantly, because these subsystems are assumed to be distinct, task-irrelevant stimuli are expected to impair performance primarily if distractor and task draw on the same sub-systems. Thus, irrelevant speech varying across successive verbal elements is expected to selectively disrupt verbal serial recall, but not visuo-spatial serial recall (e.g., [295, 265, 153]). For spatial variation in the distractor (e.g., changes in sound location), the model implies potential interference with visuo-spatial serial recall if spatial information conveyed by auditory input is represented within the visuo-spatial sketchpad. However, this assumption is not explicitly stated and comparatively underexplored.

Building on the multi-component model, Lehnert and Zimmer [262] offer a more fine-grained account of spatial information by distinguishing between modality and processing domain. Within this framework, a distinction is made not only between verbal and visuo-spatial information, but also within the visuo-spatial sketchpad between object and spatial information [277, 21]. Lehnert and Zimmer [262] define modality as the sensory input through which information is presented (e.g., visual or auditory), whereas the processing domain refers to the high-order information extracted from that input (e.g., verbal or non-verbal). Within the non-verbal domain, a distinction can be made between object information (e.g., visual form or spectro-temporal information) and spatial information (e.g., location). Spatial information can be presented in the visual modality (e.g., via spatially arranged stimuli on a visual display; see [220, 265, 295]) or in the auditory modality (e.g., via sounds from different spatial locations using loudspeakers; see [349, 457, 101, 100, 128]). Importantly, the object domain is specific to the modality (i.e., auditory or visual input) whereas spatial information is assumed to be represented amodally, such that spatial information conveyed via auditory or visual input is processed within the same memory subsystem [261, 262, 438]. From the perspective of this account, irrelevant speech varying across successive verbal elements is expected to selectively disrupt verbal serial recall, but not visuo-spatial serial recall, in line with the predictions of the multi-component memory model. Critically, because spatial information derived from auditory

input (e.g., changes in sound location) is assumed to be represented within the same amodal spatial subsystem as visually encoded spatial information, such variation in the irrelevant speech should interfere with visuo-spatial serial recall. It should not interfere with verbal serial recall, because verbal serial recall relies primarily on verbal rather than spatial processing.

Whereas the multi-component model and its extension by Lehnert and Zimmer [262] retain a distinction between representational domains, this assumption has been challenged by findings suggesting domain-general disruption in serial recall. In a seminal study, Jones et al. [220] reported that irrelevant background comprising sequences of changing letters impaired performance not only in verbal serial recall but also on visuo-spatial serial recall. Performance in both serial recall tasks was more strongly reduced by changing-state sequences (e.g., "A-B-C-D-E", see [218]) than by steady-state sequences (e.g., repeated presentation of the same letter, "A-A-A-A-A"). These findings are accounted for by the O-OER model [219, 222], which assumes a unitary, amodal memory system in which information from different sensory modalities is represented as amodal items. Within this model, disruption arises from interference-by-process [228]. Specifically, if the irrelevant stimulus contains sequential order information, its automated and obligatory processing competes with the order information in the target sequence, thereby disrupting performance in tasks that rely on serial order processing, irrespective of whether the task involves verbal or visuo-spatial material. Accordingly, the O-OER model predicts that any form of sequential variation in the irrelevant speech - whether verbal (e.g., changing letters) or spatial (e.g., changes in sound source location) - should disrupt both verbal and visuo-spatial serial recall.

Evidence supporting the O-OER is mixed. While many studies support stronger disruption from changing-state than steady-state stimuli for verbal stimuli (e.g., [293, 407, 35], but see [37]), it is unclear to which degree verbal changing-state sequences affect both verbal and spatial serial recall (e.g., [220, 457], but see [265, 153, 253, 167]). Recently, [295] demonstrated in two high-powered experiments closely modeled after Jones et al. [220] that changing-state speech reliably impaired visual-verbal serial recall, but not visuo-spatial serial recall. These findings support the multi-component working memory model and challenge the O-OER model. Importantly however, these studies manipulated verbal but not spatial variation in the distractor. Thus, while the O-OER model clearly predicts that spatially varying irrelevant speech should also disrupt both tasks, this prediction has, to our knowledge, not yet been tested directly in a serial recall paradigm.

Initial empirical evidence that spatial variation in irrelevant sound may affect spatial memory comes from Schlittmeier and Fintor [403]. In their study, participants performed a virtual way-finding and landmark recall task while listening

to irrelevant speech, whose sound source location changed after each sentence. Compared to both silence and speech emitted from a constant spatial frontal location, performance was significantly impaired, indicating that spatial variation in the distractor can disrupt spatial memory performance. However, because this paradigm differs from standard visuo-spatial serial recall, it remains unclear whether such effects generalize to serial recall tasks and whether spatially varying irrelevant speech also affects verbal serial recall.

3.1.1 Present Study

The present study addresses this gap by investigating whether spatial variation in irrelevant speech produces domain-specific or domain-general disruption of serial recall by systematically manipulating both the verbal and spatial characteristics of irrelevant speech. Specifically, participants performed visual-verbal and visuo-spatial serial recall tasks under conditions that varied whether the irrelevant speech changed across successive verbal elements (verbal-changing) and whether it changed in spatial location (spatial-changing) over time. Both tasks (visual-verbal and visual-spatial serial recall) are therefore performed under five sound conditions: silence, spatial-static verbal-static speech (SS+VS), spatial-changing verbal-static speech (SC+VS), spatial-static verbal-changing speech (SS+VC), and spatial-changing verbal-changing speech (SC+VC).

Based on prior evidence, we expect to replicate the well-established finding that verbal variations (VC conditions) disrupt visual-verbal but not visuo-spatial serial recall [295, 265, 153], consistent with the predictions of the multi-component memory model.

The main research question, however, concerns the role of spatial variation (SC conditions) in the distractor. Specifically, the proposed study design allows us to test whether spatially varying irrelevant speech selectively disrupts visuo-spatial serial recall or also affects verbal serial recall. Competing theoretical predictions about this research question can be derived from the aforementioned models of working memory. The multi-component memory model [25, 21], as a content-based account, predicts that verbal variation in the distractor (VC conditions) should disrupt visual-verbal serial recall, but not visuo-spatial serial recall. With regard to spatial variation in the distractor (SC conditions), this model does not yield an equally strong and specific prediction. The extension of this framework proposed by Lehnert and Zimmer [262] predicts that spatial variation in the distractor (SC conditions) should disrupt visuo-spatial serial recall, but not visual-verbal serial recall. In contrast, the O-OER model [219, 222], as a process-based account, predicts that both verbal and spatial variation

in the distractor (VS and SC conditions) should disrupt both visual-verbal and visuo-spatial serial recall.

In light of the findings by Schlittmeier and Fintor [403], who found a spatial changing-state effect in a spatial memory task, as well as the study by Marsh et al. [295], who found a verbal changing-state effect for verbal serial recall but not for visuo-spatial serial recall, we derive domain-specific predictions: we expect the verbal-changing distractors (VC) to produce a greater disruption to the verbal serial recall compared to verbal-steady distractors (VS), and a greater disruption of visuo-spatial serial recall by spatial-changing distractors (SC) compared to spatial-static distractors (SS). Importantly, we do not expect a cross-impact of the spatial-changing distractors (SC) on verbal serial recall and vice versa (VC on visuo-spatial serial recall), as would be expected based on a strict interference-by-process account. In summary, we hypothesize that domain-specific changing-state effects exists, mirrored in (a) a verbal changing-state effect in verbal but not visuo-spatial short-term memory, and (b) a spatial changing-state effect in visuo-spatial but not verbal short-term memory.

3.2 Materials and Methods

To explore these hypotheses, we conducted a listening experiment employing two tasks: a verbal serial recall task and a visuo-spatial serial recall task (cf. [295, 220, 265]). Both tasks were performed in the presence of five sound conditions: silence, SS+VS, SS+VC, SC+VS, and SC+VC. Details are given in the following sections.

3.2.1 Participants

We recruited 31 participants, but one participant could not complete the session within the available time and was therefore not included in the final sample ($N = 30$, 15 male, 15 female, $22 - 35$ y.o., $M = 26.9$ y.o., $SD = 3.4$ y.o.). Participants had normal hearing and (corrected-to-)normal vision. Normal hearing below 25 dB HL between 250 Hz and 8 kHz was verified with an AURITEC Ear 3.0 audiometer and Sennheiser HDA300 headphones using an ascending pulsed pure tone audiometry. Participants received 10 € for their time and effort.

Power Analysis

We ran a pilot study without irrelevant speech to balance the difficulty of the visual-verbal and visuo-spatial serial recall tasks (see Section 3.2.2 for a detailed description of the tasks). Eight participants performed both tasks with sequence lengths of seven and nine items. Task order and sequence lengths were balanced

across participants in four experiment blocks. The first two blocks contained one task, while the last two contained the other, avoiding task switching between blocks. Sequence lengths were balanced within the same-task blocks. Each block consisted of twelve trials.

We examined recall performance by fitting a logistic mixed effects regression model to the data using the package glmmTMB (1.1.13, [69]) for R (4.4.3, [374]). The model formula included fixed effects for the two tasks and the sequence length. We further included a serial position effect nested within sequence length, and participant-level random intercepts and slopes for the interaction between task and sequence length. Given the small sample size, fitting the full covariance structure was infeasible. Instead, we fitted a rank-reduced covariance structure as described by McGillycuddy et al. [304] as a trade-off between model complexity and accounting for sources of individual variation. The resulting model formula was $(N_c, N_e) \sim \mathit{Task} * \mathit{Sequence\ Length}\ /\ \mathit{Serial\ Position} + (1 + \mathit{Task} * \mathit{Sequence\ Length} \mid \mathit{ID})$. Residual diagnostics for an acceptable model fit were performed with the R package DHARMa (0.4.7, [178]).

Because we were interested in the sequence lengths that resulted in the most similar task performances, we proceeded with pairwise comparisons of estimated marginal means for the task-sequence length interaction using the R package emmeans (2.0.0, [267]). All comparisons were multiplicity-adjusted using Holm's method. Participants performed best in the seven-item visual-verbal task ($M = 89.56\,\%$, $CI = 67.85\,\%$ to $100\,\%$) and worst in the nine-item visuo-spatial task ($M = 57.58\,\%$, $CI = 46.88\,\%$ to $68.28\,\%$), $OR = 29.10$, $z = 2.849$, $p = .0219$. Recall performance was least likely to differ statistically between the nine-item visual-verbal ($M = 75\,\%$, $CI = 62.43\,\%$ to $87.58\,\%$) and the seven-item visual-spatial task ($M = 80.75\,\%$, $CI = 71.35\,\%$ to $90.14\,\%$), $OR = 1.29$, $z = 0.642$, $p = .5206$. Therefore, we chose the nine-item visual-verbal and the seven-item visuo-spatial task configurations.

We obtained effect size estimates of the verbal changing-state effect from Marsh et al. [295]. They report a mean performance decline of 7 % for changing-state speech compared to silence and a 5 % decrease for changing- compared to steady-state speech. Our spatial changing-state effect size estimates were based on the results reported by Schlittmeier and Fintor [403], who found a 1-2 point performance decrease for spatial changing-state speech compared to spatial steady-state speech and silence. With a point score maximum of 16, these amount to 6.25 %-point to 12.5 %-point performance changes. Thus, for the purpose of the power analysis, and in the absence of directly comparable estimates for the spatial changing-state effect, we used the effect sizes reported by Marsh et al. [295] as pragmatic reference values for both the verbal and spatial changing-state comparisons. Reference performance levels in silence and participant-level

random effects were simulated based on corresponding estimates from the pilot study model.

The power analysis was then performed by simulating 1000 binomial datasets with mean correct response percentages as described above. Datasets varied from 20 to 50 participants and from ten to 14 trials per task and background auditory condition combination. The datasets did not contain serial position effects because their inclusion would strongly increase simulation complexity. We then fitted a logistic mixed effects regression model with the formula $(N_c, N_e) \sim \mathit{Task} * \mathit{Irrelevant\ sound} + (1 + \mathit{Task} \mid \mathit{ID})$ to the simulated datasets and obtained power curves for the critical changing-state effect comparisons. These were the 5 % performance decreases between verbal-static and verbal-changing speech for the visuo-verbal task, and between spatial-static and spatial-changing speech for the visual-spatial task, respectively. The proportion of models that detected these effects at a type-I error level of $\alpha = .05$ yielded the statistical power as a function of the number of participants N and the number of trials T. Statistical power exceeded $1 - \beta = .90$ at $N = 30$ and $T = 12$. Please refer to the corresponding data publication for more details.

3.2.2 Serial Order Reconstruction Tasks

There were two tasks: a visual-verbal and a visuo-spatial serial recall task (cf. [295]). The tasks were implemented in PsychoPy (v.2025.1.1, [357]) based on the implementation by Leist et al. [265]. Each trial started with a count-down consisting of three black rectangles decreasing in size presented for 0.5 s each. Afterwards, the visual-verbal or visuo-spatial sequences were presented with a 1 s display period and an ISI of 1 s, resulting in a presentation rate of 1 item per 2 s. There was no retention interval; the recall and input phase started directly after the presentation of the last digit. The items appeared (see below for task-specific details), and participants were asked to click on the items in the order in which they were presented. Once selected, the items turned gray, indicating that they had been selected. Corrections were not possible. All items had to be selected to move on to the next trial. After clicking on the last item, the next trial started automatically.

Visual-Verbal Serial Recall Task

The target stimuli were sequences of the digits $1 - 9$ in random order. Sequences were pre-generated according to three rules: each digit appeared exactly once within a sequence, no more than two consecutive steps of the same size were allowed (e.g., $1-2-3$ or $9-8-7$ within a sequence was possible, but $1-2-3-4$ or $9-8-7-6$ was not, cf. [80]), and each digit was presented at each serial

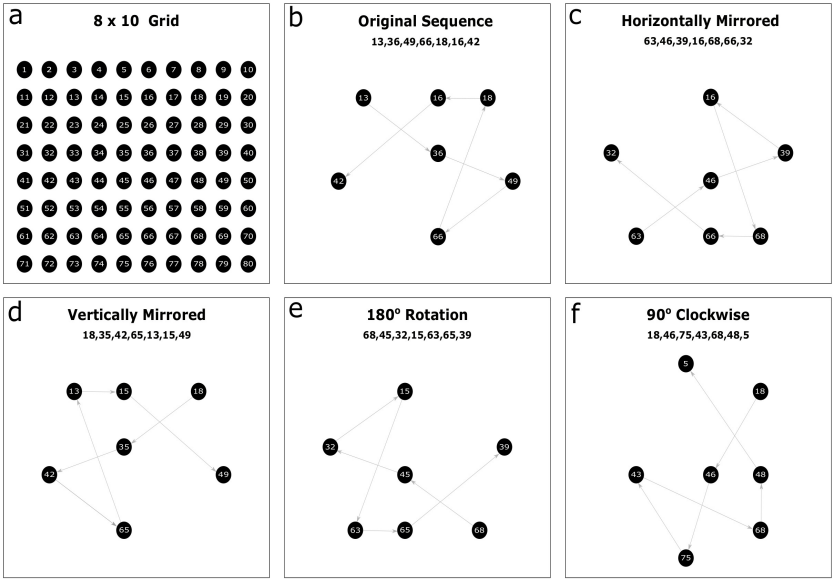


Figure 3.1: Generation of dot sequences for the five background sound blocks after Leist et al. [265]: (a) possible dot positions; (b) exemplary original sequence; (c) horizontally mirrored sequence; (d) vertically mirrored sequence; (e) sequence with 180° clockwise rotation; (f) sequence with 90° clockwise rotation.

position approximately equally often. The pre-generated digit sequences were presented to each participant in random order. No sequence was presented twice to a participant.

The digits were presented visually one after another in the center of the computer screen in black 44 pt bold-face Calibri on a white background. Immediately following the presentation of the digit sequence, all nine digits appeared on the screen within their own box in a jumbled 3×3 array to avoid visual memory techniques. Participants were instructed to click the numbers in the exact order they were presented.

Visuo-Spatial Serial Recall Task

The target stimuli were sequences of seven black dots with a diameter of 0.8 cm presented in an invisible 8×10 grid on an area of 304×183mm [265] on the computer screen.

Each dot stimulus was assigned a unique index position on a 8-row by 10-column grid (see Figure 3.1a). This index was then utilized to calculate the dot's x and y coordinates in PsychoPy relative to the center of the screen. As in the study by Leist et al. [265], the dot patterns were mirrored and rotated to obtain unique sequences of comparable difficulty for each experimental condition. As there were five experimental conditions (silence, SS+VS, SS+VC, SC+VS, and SC+VC), the original sequences were extended *via* transformations (see Figure 3.1): horizontal mirroring (c), vertical mirroring (d), 180° clockwise rotation (e), and 90° clockwise rotation (f). These transformations and rotations were balanced across auditory conditions. During the presentation period, the mouse cursor was made invisible, preventing the participant from following the positions of the dots. After the presentation period ended, the cursor reappeared. During the recall and input phase, all seven dots were displayed on the screen at once, and the participants were instructed to click the dots in the same order as they were originally shown.

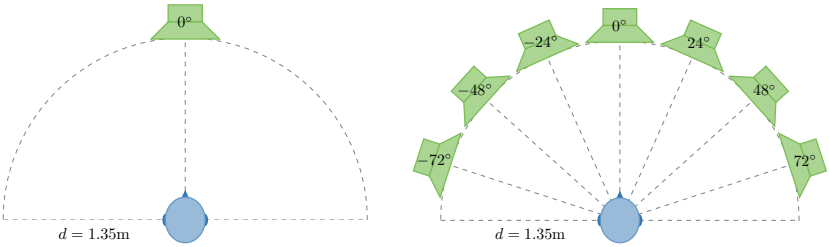
3.2.3 Irrelevant Sound Sequences

As stated above, the verbal and the visuo-spatial serial recall tasks were performed under five auditory conditions: silence, SS+VS, SS+VC, SC+VS, SC+VC.

The audio stimuli themselves were chosen to replicate Experiment 2 by Marsh et al. [295] as closely as possible. The same auditory stimuli as in the study by Marsh et al. [295] (Experiment 2) were used: recordings of the letters "A" through "G" spoken by a German male. As in the study by Marsh et al. [295], all stimuli were 250 ms in length and separated by an ISI of 250 ms, resulting in a presentation rate of two letters per second. The letter sequences were presented only during the encoding phase: sequences started simultaneously with the first to-be-remembered item and stopped at the start of the response phase. The sound pressure level was calibrated to 60 dB at the center of the loudspeaker array with a B&K Type 4134 1/2" diffuse field microphone, a B&K Nexus conditional amplifier type 2691, the ITA toolbox for Matlab [41], and Matlab 2023b.

Verbal-Static and Verbal-Changing Speech

The verbal-static (VS) sequence only repeated the letter "A". The verbal-changing (VC) sequence consisted of the letters "A" through "G". Contrary to the study by Marsh et al. [295], these letters were not played in alphabetical order, but pseudo-random order: all letters had to be played before they appeared again, and a letter could never be repeated directly. This was done to enable comparability to the spatial changing (SC) condition (see below), where the spatial positions changed



(a) Spatial-static condition: Auditory stimuli are played from the front of the participant.

(b) Spatial-changing condition: Auditory stimuli are played from seven loudspeakers in pseudo-randomized order.

Figure 3.2: Loudspeaker positions for spatial audio reproduction.

pseudo-randomly as well. The audio stimulus was played from a loudspeaker array [354] mounted in an anechoic chamber [353].

Spatial-Static and Spatial-Changing Speech

In the spatial-static (SS) condition, the letter sequences were played back from a frontal position (cf. Figure 3.2a). In the spatial-changing (SC) condition, there were seven possible loudspeaker positions on the horizontal plane (cf. Figure 3.2b): one at the front (0°), and loudspeakers to the left and right at ±24°, ±48°, and ±72°; with the angular distance of 24° defined by the loudspeaker array. Within the letter sequence, the loudspeaker position changed pseudo-randomly with each letter to enable comparability to Schlittmeier and Fintor [403] (Experiment 1): letters had to be played from all positions before a position reappeared, and a position could never be repeated directly.

3.2.4 Procedure

Each participant completed the experiment in an individual session inside an anechoic chamber. Each session began with an audiometry to ensure that the participant had normal hearing. Before the experiment started, the experiment supervisor used a 3-axis laser and an adjustable chair to ensure that the participant was seated precisely at the center of the loudspeaker array. A monitor (Samsung 2443BW) positioned at a distance of 1.55 m and a height of 1.3 m - right in front of the participant - was used to display written instructions and visual stimuli. Participants responded with a Logitec M325 wireless computer mouse.

Participants were instructed to refrain from saying the sequences out loud and from using their fingers or any physical movements to assist their memory. The experiment consisted of two parts: one for each task. The order of the tasks was counterbalanced over participants, such that participants either started with the verbal or the spatial task. They were informed that background noise would be present, but were asked to ignore it.

Each experiment part was structured as follows: Participants started with a practice block, which introduced them to the task and the background auditory conditions. The practice blocks included six trials that followed the same fixed sequence: first, two trials in silence were administered, which were followed by one trial per background noise (SS+VS, SC+VS, SS+VC, SC+VC). After completing the practice trials, participants proceeded to the five main experiment blocks. Each block consisted of twelve trials, or a total of 60 trials per task. Each block was assigned one of the five auditory conditions (silence, SS+VS, SC+VS, SS+VC, SC+VC), and the order of blocks was counterbalanced across participants using a Latin square design. After each block, participants could take a voluntary break of no fixed length, but they could not take a break between trials. A mandatory break of 4 min was administered halfway through the experiment after completing all blocks of one task before moving on to the next task. The entire experiment, including the audiometry, took approximately 100 min.

3.2.5 Data Analysis

We first applied the statistical model from the power analysis to the data, see Section 3.2.1. However, residual diagnostics suggested a misspecified model, which we suspected to result from ignoring serial position effects. Because the number of items differed between the verbal and the spatial task, we included the serial position as a fixed effect nested within the task effect. This nesting structure allowed us to estimate serial position effects independently for both tasks. Further, we added the chronological block number as a nuisance variable to control for condition imbalances and performance fluctuations with experiment duration.

The resulting model was fitted in R (4.4.3, [374]) using the package glmmTMB (1.1.13, [69]). It was formulated as

$(N_c, N_e) \sim \mathit{Task} \,/\, \mathit{Serial\ position} * \mathit{Irrelevant\ sound} + \mathit{Block\ number} + (1 + \mathit{Task} + \mathit{Irrelevant\ sound} \mid \mathit{ID})$,

where (N_c, N_e) are the counts of correct and incorrect responses across repetitions, denoting the binomial outcome variable. The default optimizer settings were adjusted to increase the number of iterations for the model fit and were unchanged otherwise.

Like for the pilot study model, residuals were examined using the R package DHARMa (0.4.7, [178]).

Factor significance was determined by parametric bootstrap likelihood ratio tests (see [304] for another application example). These tests are performed by fitting a null and an alternative model to the experimental data for each factor or interaction. While the null model excludes the factor of interest, it is included in the alternative model. Additionally, higher-level interactions with the factor of interest are excluded from both the null and the alternative model. By simulating data from the null model and refitting both the null and the alternative model to the simulated datasets, we obtain a bootstrapped likelihood ratio distribution under the null hypothesis that the factor of interest does not affect the outcome variable. If the likelihood ratio between the null and the alternative model fitted on the experimental data exceeds the 95th percentile of the bootstrapped likelihood ratio distribution, it is statistically significant at $\alpha = .05$.

Post-hoc comparisons were performed using the R package emmeans (2.0.0, [267]) with Holm's adjustment method for multiple comparisons. We report estimated mean correct response percentages, but pairwise comparisons were performed on odds ratios. The correct response percentage estimates were obtained by back-transforming the odds estimates from the link scale to the response scale and applying emmeans' bias-adjustment procedure for random effect variances [56, 273, 267]. Please refer to the corresponding data publication for more details.

3.3 Results

The results of the parametric bootstrap likelihood ratio tests are listed in Table 3.1. The task did not have a significant effect on recall performance, indicating that task difficulty was successfully matched between the verbal and the spatial task. In contrast, irrelevant speech did affect recall performance, and its effect interacted with the task. Block number and task-specific serial position effects were also found to be statistically significant, but were not of primary interest.

To test our hypotheses (see Section 3.1.1), we examined the interaction of task and irrelevant speech by comparing irrelevant speech conditions for each task. These comparisons are depicted in Figure 3.3.

3.3.1 Visual-Verbal Task

Recall was best in the silent condition ($M = 78.09\,\%$, $CI = [72.17, 84]\,\%$) and decreased with SS+VC speech ($M = 69.76\,\%$, $CI = [63, 76.52]\,\%$), relative to silence, $OR = 1.654$, $z = 4.813$, $p < .001$. Performance was likewise reduced in

Table 3.1: Results of the parametric bootstrap likelihood ratio test. df: degrees of freedom, $-2\Delta LL$: log-likelihood ratio. Significance level $\alpha = .05$. Significant comparisons are printed bold.

Factor	df	$-2\Delta LL$	p
Task	1	0.02	0.912
Irrelevant speech	**4**	**14.30**	**0.006**
Block number	**9**	**73.36**	**<0.001**
Task * Serial position	**14**	**1498.45**	**<0.001**
Task * Irrelevant speech	**4**	**36.71**	**<0.001**
Task * Serial position * Irrelevant speech	56	63.08	0.205

SC+VC ($M = 68.83\,\%$, $CI = [62.28, 75.39]\,\%$) compared to silence, $OR = 1.8$, $z = 4.018$, $p < .001$.

The VS speech condition performances (SS+VS, SC+VS) fell in between the silent and the two VC conditions (SS+VC, SC+VC). Performance in the SS+VS condition $M = 77.16\,\%$ ($CI = [71.53, 82.79]\,\%$) was not significantly different from the silent condition, $OR = 1.12$, $z = 1.119$, $p = .5266$. In contrast, the difference between the SC+VS condition ($M = 72.98\,\%$, $CI = [66.11, 79.84]\,\%$) and the silent condition was significant, $OR = 1.312$, $z = 2.626$, $p = .0432$.

Most importantly, we found evidence for a verbal, but not spatial, changing-state effect in the verbal task. Differences between VS and VC speech conditions were significant, whether the irrelevant speech source was, in both cases, static (i.e., SS, $OR = 1.477$, $z = 3.364$, $p = .006$) or changing (i.e., SC, $OR = 1.373$, $z = 2.882$, $p = .024$). Conversely, SS did not differ from SC conditions, neither when both were combined with VS speech ($OR = 1.172$, $z = 1.410$, $p = .4753$), nor with VC speech ($OR = 1.088$, $z = 0.773$, $p = .5266$).

Significant differences also emerged when both spatial and verbal content switched from static to changing or vice versa. Recall in SS+VS speech was better than in SC+VC speech ($OR = 1.608$, $z = 3.288$, $p = .007$), and performance in SS+VC speech was worse than in SC+VS speech ($OR = 0.793$, $z = -2.607$, $p = .0432$). Although these contrasts confound spatial and verbal changing-state characteristics, the pairwise pattern (significant VS–VC differences within SS and within SC, but no SS–SC differences within VS or within VC) indicates that they are primarily driven by the verbal changing-state manipulation.

3.3.2 Visuo-Spatial Task

For the visuo-spatial task, none of the irrelevant speech conditions differed from the silent condition. The largest difference between any two irrelevant speech con-

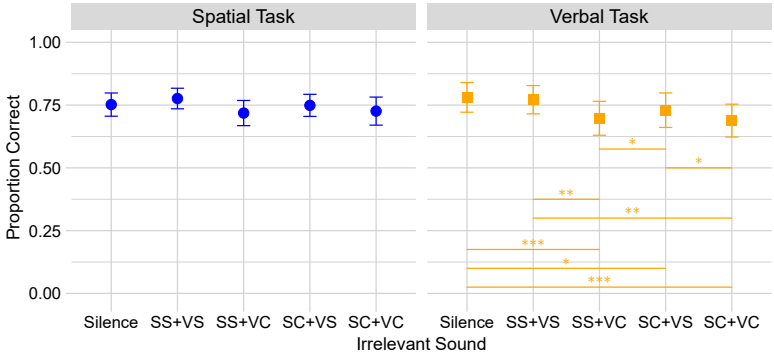


Figure 3.3: Recall performance in the visuo-spatial (left, blue circles) and visual-verbal (right, orange squares) serial recall tasks depending on the presented irrelevant sound. Abbreviations: SS+VS - spatial- and verbal static, SS+VC - spatial-static and verbal-changing, SC+VS - spatial-changing and verbal-static, SC+VC - spatial- and verbal-changing. Significance indicators: $* < .05$, $** < .01$, $*** < .001$.

ditions was observed between SS+VS and SS+VC speech ($\Delta M = 5.79$ %-points), but was not significant, $OR = 1.3575$, $z = 2.619$, $p = .0882$. Therefore, we did not find evidence for either verbal or spatial changing-state effect in the visuo-spatial task.

3.4 Discussion

The present study aimed to examine how spatially varying irrelevant speech affects visual–spatial short-term memory and visual–verbal short-term memory. We hypothesized that (a) a verbal changing-state effect exists in verbal but not visuo-spatial serial recall, and (b) a spatial changing-state effect exists in visuo-spatial but not verbal serial recall.

The results show a clear interaction effect between the task (visuo-verbal vs. visuo-spatial) and the auditory condition (silence, SS+VS, SS+VC, SC+VS, SC+VC), indicating a task-dependent effect of the irrelevant sounds. Crucially, this pattern was driven by a robust verbal changing-state effect alongside the absence of a corresponding spatial changing-state effect. The pairwise comparisons for the interaction paint an asymmetric picture. For the visuo-spatial task, none of the irrelevant speech conditions differed significantly from each other. In contrast, several comparisons were significant for the verbal task: verbal-static (VS) speech did not impair verbal serial recall more than silence, while verbal-changing (VC) speech produced a reliable ISE compared to both silence and verbal-static

(VS) speech. Spatial-changing speech did not add impairment: comparisons between spatial-static and spatial-changing conditions (SS vs. SC) did not produce differences, irrespective of the verbal characteristics (VS and VS). Differences between SS and SC conditions were observable only when the comparisons also contained a verbal changing-state effect (i.e., SS+VS vs. SC+VC).

An exception from these findings is the comparison between silence and spatial-changing verbal-static speech (SC+VS). This comparison was significant and cannot be attributed to a verbal changing-state effect. However, this does not necessarily indicate the existence of a spatial changing-state effect in verbal serial recall, since none of the comparisons involving spatial-static vs. spatial-changing stimuli (SS vs. SC) were significant. Given this inconsistent pattern, the isolated difference between silence and SC+VS is best interpreted with caution.

The lack of a significant difference between SS+VC and SC+VC might be explained by postulating that verbal and spatial changing-state effects are not linearly additive, but are both caused by a more general, domain-general (both spatial and verbal) changing-state effect. Still, if the existence of such a general changing-state effect in verbal memory is assumed, then the comparison between SS+VS and SC+VS should also be significant. This was not the case. These arguments speak against a reliable effect of spatial-changing noise in verbal serial recall, whether it is assumed to be independent of the verbal changing-state or not. Unless the effect of spatial-changing noise is smaller than assumed in the power analysis, the difference between silent and spatial-changing verbal-static irrelevant speech conditions may have occurred incidentally.

With regard to hypothesis (a), both the classic verbal changing-state effect in the ISE, as well as the domain-specificity of the verbal changing-state effect was replicated [295], supporting interference-by-content accounts and adding to the accumulating evidence that questions the finding by Jones et al. [220]. This is supported by the fact that irrelevant speech with verbal-changing-state characteristics reliably affected verbal serial recall. Crucially, the same manipulation did not reliably affect visuo-spatial serial recall. This confirms prediction (a) and closely aligns with the findings of Marsh et al. [295], who also reported a robust verbal changing-state effect for visual–verbal but not visual–spatial serial recall.

Revisiting hypothesis (b), which predicted a spatial changing-state effect in visuo-spatial serial recall, spatially changing irrelevant speech did not disrupt serial recall, neither in the visuo-spatial task nor the verbal task. Thus, the critical prediction derived from Lehnert and Zimmer [262] framework and the work by Schlittmeier and Fintor [403] was not supported under the present conditions; no reliable spatial changing-state effect could be found. Instead, it produced a modest but statistically reliable decrement in verbal serial recall (Silent vs. SC+VC), which is - as discussed - likely circumstantial. Taken together, these findings

show that in the present paradigm, irrelevant speech primarily disrupts verbal serial recall. Visuo-spatial serial recall proved robust, even when the irrelevant background contained pronounced spatial changes.

The absence of a clear effect of spatial-changing irrelevant speech on visuo-spatial memory performance is particularly noteworthy in light of the study by Schlittmeier and Fintor [403], who found that spatially changing irrelevant speech impaired performance in a visuo-spatial way-finding and landmark recall task. This discrepancy suggests that spatially varying irrelevant speech does not affect all forms of spatial memory in the same way. Several aspects of the paradigms used explain this discrepancy: Schlittmeier and Fintor [403] used a navigation-type task involving way-finding and landmark recall, which likely engages allocentric and route-based spatial coding. Our visuo-spatial task employed here relies more on positional coding (e.g., locations in a grid). Spatial changes in an auditory scene may align more naturally with route-based spatial representations than with static positional encodings. Further, in the way-finding task, the movement of a sound source might be perceived as the movement of an object in the environment, making its spatial dynamics more relevant to the processing of the spatial scene. In our paradigm, the spatial variation of the irrelevant speech may not have been connected to the reference frame used in the visuo-spatial task to the same degree, reducing interference between irrelevant and task-related spatial processes. In addition, Schlittmeier and Fintor [403] used short matrix sentences instead of isolated letters, and location changes were implemented between sentences and with a varying delay between sentences instead of regular-timed changes, as employed in the present study. Time-varying delays and sentences as background stimuli may adhere closer to real-life scenarios, e.g., conversations, where the active speaker changes at an unpredictable rate. The regularity of changes combined with the use of short letters facilitated a systematic evaluation of potential spatial changing-state effects in the present study and enhanced comparability with previous work. At the same time, however, these same characteristics may have reduced the likelihood that the distractor was perceived as a coherent auditory stream (see, e.g., [200, 226]). If so, spatial variation may not have been encoded as an ordered sequence and therefore may not have given rise to interference-by-process, which could explain the absence of a spatial changing-state effect in the present study. Future studies could therefore vary the speech material (e.g., using sentences instead of words) and structure of spatial changes (e.g., by varying spatial separation, movement continuity, or predictability) to determine under which conditions spatial variation reliably affects serial recall.

This need for more salient spatial distractors is also consistent with Marsh et al. [295]. Having found no effect of verbal changing-state speech on visuo-spatial serial recall, they hypothesize that, although visuo-spatial serial recall

may appear to be unaffected by any auditory distraction, the absence of an effect can be explained by the duplex-mechanism account of auditory distraction [199]. According to this account, irrelevant sound can impair performance (a) via interference-by-process, if sound and task processing conflict, and/or (b) via attentional capture, when salient or deviant auditory events involuntarily draw attention away from the task. Marsh et al. [295] emphasize that while visuo-spatial serial recall in their study was not affected by verbal changing-state sequences, there is evidence that deviants in an otherwise predictable auditory sequence can disrupt visuo-spatial serial recall [295, 463]. Such susceptibility to deviant effects can be taken as evidence for attention-based disruption. Thus, Marsh et al. [295] suggest that visuo-spatial serial recall is differently affected by changing-state and attention-based auditory deviation effects.

It is possible that the irregular timing between sentences employed by Schlittmeier and Fintor [403] (and also potentially the content of the sentences) evoked attention-based distraction, causing a performance deficit in their way-finding task. In contrast, the regularly changing letters in the present study may have only evoked interference-by-process and thus not affected the visuo-spatial serial recall tasks. Further studies varying the timing and salience of the irrelevant utterances may validate this hypothesis. Especially for employing spatially-changing background signals in the pursuit of more realistic test scenarios, e.g., to simulate open-plan office scenarios (e.g., [500, 383, 128]), investigations specifically targeting aspects of such spatial complexity may be beneficial.

Revisiting the classical models of working memory, the pattern of results in the present study provides differentiated support for the multi-component model by Baddeley and Hitch [25]. Consistent with the multi-component model, verbal changing-state speech at a fixed location (SS) produced a strong ISE in the verbal task but not in the visuo-spatial task, supporting the domain-specific predictions of interference of the multi-component model. However, the data do not support the assumption that spatial variation in auditory input is sufficient to interfere with visuo-spatial serial recall via a shared amodal spatial subsystem. When the irrelevant speech changed spatially, (small) effects could only be found in the verbal task; visuo-spatial recall was not significantly affected. The predicted effect on visuo-spatial recall could not be observed. Thus, the extension proposed by Lehnert and Zimmer [262], which predicts such interference, is not supported under the present conditions. In summary, the visuo-spatial sketchpad appears relatively robust to interference from irrelevant speech, even when that speech conveys spatial order information.

The O-OER model [220] would assume that any form of changing-state characteristic within the irrelevant sequence (verbal, spatial, or otherwise) should disrupt any serial recall task, whether verbal or visuo-spatial, due to process-based inter-

ference. Our data challenge this view: neither verbal nor spatial changing-state speech produced a reliable ISE on visuo-spatial serial recall. This aligns with recent studies (e.g., [295, 265, 153], who also found that changing-state irrelevant speech impaired visual–verbal but not visuo-spatial recall. While these studies did not test a spatially-changing irrelevant sound, our results extend this evidence: in the present design, spatially changing irrelevant speech did not disrupt either visuo-spatial serial recall or verbal serial recall relative to spatially static speech, although the O-OER model predicts disruption of both tasks. Thus, the present findings do not support the O-OER prediction that sequential variation in the distractor, irrespective of whether it is verbal or spatial, should produce domain-general interference in serial recall. An alternative explanation is that the spatial manipulation employed here did not induce sufficiently strong sequential processing to trigger interference-by-process.

3.5 Conclusion

The present study replicated the classic changing-state effect for short-term memory and extended prior work by testing whether spatial variation in irrelevant speech exerts an additional disruptive influence. We found no evidence that spatially varying irrelevant speech impaired visuo-spatial serial recall under the present conditions, and no reliable evidence that it disrupted verbal serial recall beyond the established effect of verbal variation. These findings support a domain-specific account of auditory distraction in serial recall: verbal serial recall is highly vulnerable to variations in verbal order information, whereas visuo-spatial serial recall is comparatively robust to the type of spatial variation in irrelevant speech examined here.

Declarations

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): SPP2236 - Project number 444724862. Sabine Schlittmeier's contribution was supported by a grant from the HEAD-Genuit-Stiftung (HEAD-Genuit Foundation; P-16/10-W). Lukas Vollmer's contribution was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Project number 368482240 / GRK2416.

Competing Interests

The authors report there are no competing interests to declare.

Ethics Approval

The experiment procedure was pre-approved by the Interfaculty Ethics Committee of RWTH Aachen University (FÜEK 85/25) and was conducted in accordance with the rules of conduct stated in the Declaration of Helsinki.

Consent to Participate

All participants gave informed written consent prior to the experiment.

Consent for Publication

Not applicable.

Availability of Data and Materials

The collected experimental data is available in a corresponding data publication: `https://doi.org/10.5281/zenodo.19387019`. The acoustic stimuli were obtained from [295] and can therefore not be shared. None of the experiments was preregistered.

Code Availability

The analysis scripts are available in a corresponding data publication: `https://doi.org/10.5281/zenodo.19387019`. The experimental code was obtained and adapted from [265] and can therefore not be shared.

Authors' Contributions

C.E.: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data curation, Writing - Original Draft, Visualization **L.V.:** Conceptualization, Methodology, Formal Analysis, Data curation, Writing - Original Draft, Visualization **H.D.:** Software, Investigation, Writing - Original Draft, Visualization **S.S.:** Conceptualization, Methodology, Writing - Review & Editing, **J.F.:** Supervision, Writing - Review & Editing, Funding acquisition, Project administration

Acknowledgements

The authors would like to thank Larissa Leist for providing us with the code used in the study by Leist et al. [265], Raoul Bell for sharing the audio stimuli used in the study by Marsh et al. [295], and Maria Xanthopoulou for her contribution to the data collection.

4

Audiovisual angle and voice incongruence

This Chapter was published as an original research article in the journal *PLoS One*:
C. A. Ermert, M. Yadav, J. Ehret, et al. "Audiovisual angle and voice incongruence do not affect audiovisual verbal short-term memory in virtual reality". In: *PLoS One* 20.8 (2025) [127].

4.1 Abstract

VR environments are frequently used in auditory and cognitive research to imitate real-life scenarios. The visual component in VR has the potential to affect how auditory information is processed, especially if incongruences between the visual and auditory information occur. This study investigated how audiovisual incongruences in the VR implemented with a HMD affects verbal short-term memory compared to presentation of the same material over traditional computer monitors. Two experiments were conducted with both these display devices and two types of audiovisual incongruences: angle (Exp 1) and voice (Exp 2) incongruence. To quantify short-term memory, an avVSR task was developed where an ECA was animated to speak a digit sequence, which participants had to remember. The results showed no effect of the display devices on the proportion of correctly recalled digits overall, although subjective evaluations showed a higher sense of presence in the HMD condition. For the extreme conditions of angle incongruence in the computer monitor presentation, the proportion of correctly recalled digits increased marginally, presumably due to raised attention, but the effect size was negligible. Response times were not affected by incongruences in either display device across both experiments. These findings suggest that at least for the conditions studied here, the avVSR task is robust against angle and voice audiovisual incongruences in both HMD and computer monitor displays.

4.2 Introduction

Experiments in auditory and cognitive sciences research are increasingly conducted in VR environments [350, 367]. VR can evoke a higher feeling of presence, the feeling of "being in a scene" [493], compared to computer monitor presentations [161]. Thus, VR is widely employed to create immersive environments that approximate real-life scenarios. Compared to traditional laboratory setups with computer monitors, in most cases, experiments in VR differ in terms of the visual information presented. This offers new possibilities for designing experimental environments and tasks, but also raises the question of how this added visual information affects cognitive performance. Importantly, the effect of VR on participants is not yet fully understood and appears to be highly dependent on the experimental design. Several studies have reported positive effects of VR, such as increased attention [270, 60] and enhanced motivation and engagement [488, 198]. However, studies have also reported disadvantages arising from using VR, since it can cause cybersickness [482] and may evoke a higher mental load [286, 379].

When implementing audiovisual tasks in VR, perceived incongruences between the auditory and visual modalities can arise. This is especially relevant when the visual VR environment is not carefully crafted and aligned with the acoustic VR environment. Such an interplay of auditory and visual information has been a key aspect of cognitive research for multiple years, albeit without VR, and being based instead on computer monitor presentations. These studies have shown that coherent auditory and visual information can be integrated into a single multimodal percept [436, 50, for a review]. Coherence in general is a multidimensional construct, including but not limited to spatial, temporal, and semantic alignment of the auditory and visual stimuli [436]. Spatial audiovisual integration is commonly investigated in terms of the *ventriloquism effect* [84, 454], which refers to the bias in auditory spatial perception toward a synchronous but spatially discrepant visual stimulus. This effect is typically investigated using brief sounds (e.g., clicks) and visual flashes presented with small spatial separations [7]. Further, audiovisual stimuli are perceived as temporally incoherent if their stimulus onset asynchrony, i.e., the temporal offset between the onset of the auditory and visual stimuli, is large enough [50]. If auditory and visual stimuli are aligned in meaning, semantic congruence is elicited [436]. A well-known example of studying semantic incongruence is the *Stroop task* [444], where participants see color words (e.g., "red") printed in a different color (e.g., green), and are asked to name the color of the ink. Another established audiovisual incongruence effect is the *McGurk effect* [306], which is typically studied using lip movements that are presented together with congruent or incongruent syllables (e.g., "ba", "ga").

These studies show that lip movement can alter the auditory perceived syllable, if they share the same features (e.g., "pa" and "ba", but not "pa" and "ta") [306].

While the incongruences listed above have mostly been studied with computer monitor presentations, there is some research investigating audiovisual congruence in VR. For example, the ventriloquism effect has been replicated in VR by Huisman et al. [205]. A large angular audiovisual separation has been shown to impede user experience in terms of presence, immersion, and cybersickness, while small separation angles are typically well-tolerated by participants in VR [235]. Incongruence in lip movement in terms of the McGurk effect is detected more easily in VR than in computer monitor settings [425]. Providing coherent user interaction in VR, e.g., by synchronized audiovisual virtual hand illusions [86] or sound feedback according to the user's movements, e.g., walking [193], has been shown to increase immersion.

Notably, these studies on audiovisual incongruences in VR have primarily focused on perception-based effects, such as detectability, or subjectively perceived presence and immersion. As VR is increasingly being employed in cognitive research, it is crucial to understand whether incongruences impact cognitive functions that rely on higher-level processing of information. While it is logical to hypothesize that audiovisual incongruences beyond a certain range could impede cognitive task performance in VR, to the best of our knowledge, this has not been demonstrated yet in experimental settings.

Moreover, within the context of real-world interaction, a particularly relevant cognitive function where auditory and visual information is integrated includes listening to a talker and trying to remember what has been said. In contrast to controlled laboratory settings, everyday face-to-face interactions typically exhibit high audiovisual congruence: when a person speaks, their voice is accompanied by matching visual cues such as spatial location, lip movement, gaze direction, vocal pitch (e.g., gender), and gestures [452] - all temporally synchronized and spatially aligned. However, in VR, any of these congruences may be compromised, for instance, when the perceived voice does not match the position or gender appearance of the virtual human, the ECA.

Hence, this study aims to investigate the extent to which certain audiovisual incongruences that are relevant to everyday interaction between two interlocutors affect cognitive task performance and subjective assessments across HMD and traditional computer monitor presentations. To examine whether audiovisual incongruences impair memory for spoken content, we implemented an audiovisual version of the verbal serial recall task [207]. This implementation includes measuring verbal short-term memory under systematically controlled audio-visual incongruencies: in the first experiment, the effect of audiovisual angle incongruences was studied, whereas the second experiment studied the effect of voice

incongruence. In both studies, an HMD and a computer monitor were used as display devices to enable a direct comparison.

4.3 Experiment 1: Audiovisual Spatial Incongruence

As mentioned previously, the starting point for this study is a situation common to everyday life: listening to and remembering spoken information by an interlocutor. Traditionally, the ability to rememorize and remember verbal information has been measured using the verbal serial recall task [207]. In this task, participants have to remember and recall a sequence of verbal items, e.g., a sequence of digits or words, for a short period of time. These items are typically presented unimodally, i.e., visually only (as images [102], written [447, 179], or as lip-movements [108]), or auditorily only (as spoken words [404, 447], like in a listening situation). For a discussion of the differences between visual and auditory digit presentations, please refer to Yadav et al. [500].

This task can be extended to the aforementioned scenario of listening to an interlocutor in laboratory settings where a visual representation of the interlocutor - an ECA - speaks the digit sequences in the serial recall task. This involves the digit sequences being presented audiovisually, i.e., as simultaneous lip movements and auditory items. This variation of the classic verbal serial recall is referred to as avVSR in the following.

In selecting the audiovisual incongruence for Experiment 1, our objective was to manipulate co-verbal information conveyed by the ECA, so that the participants would naturally focus on it, while simultaneously ensuring that their ability to successfully complete the task was not directly impeded. Thus, we decided to introduce a spatial incongruence in the audiovisual stimulus presentation. More specifically, an angular offset was introduced between the visual position of the ECA and the direction from which the digit sequences were spoken to create a discrepancy between the auditory and visual scenes. The influence of this *audiovisual angle incongruence* on avVSR performance was evaluated for both HMD and computer monitor presentations as depicted in Fig 4.1.

4.3.1 Method

Participants

A total of 25 adults were recruited for the experiment via mailing lists and posters in the authors' institutes between 27th of April and 13th of May 2022. Participants were required to be native German speakers (as the entire experiment was conducted in German), have normal hearing, and (corrected-to-)normal vision. Normal hearing below 25 dB SPL [498] between 250 Hz and 14 kHz was tested

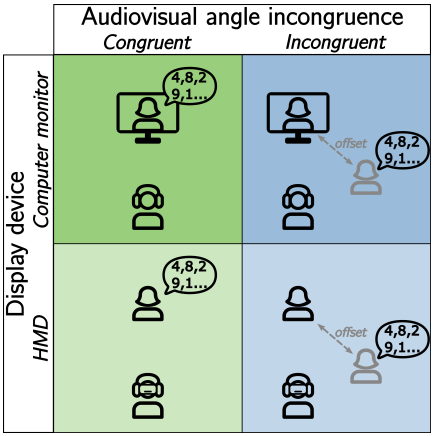


Figure 4.1: Factors in Experiment 1. The between-subject experimental design of the avVSR task with three factors: display device (computer monitor vs. HMD), the audiovisual angle incongruence (congruent vs. incongruent), and the serial position (1 - 8) of the target digit sequence. In the congruent case, the ECA and the virtual sound source are co-located; in the incongruent case, there is an angular offset between the ECA and the virtual sound source.

with an AURITEC Ear 3.0 audiometer and Sennheiser HDA300 headphones using a pulsed pure-tone ascending audiometry. Two participants failed the audiometry and had to be excluded. $N = 23$ participants (16 male, 6 female, 1 non-binary, aged between 19 and 36 years, $M = 25.74$, $SD = 4.03$) passed the audiometry. Normal or corrected-to-normal vision was validated with a Snellen chart up to (20/30) [435]. All participants gave written informed consent before the experiment began. Participants received a 10 € gift voucher for a bookstore. The experiment procedure was pre-approved by the Ethics Committee at the Medical Faculty of RWTH Aachen University (EK396-19). The study was conducted in accordance to the rules of conduct stated in the Declaration of Helsinki.

Cognitive Task

Each experimental trial of the avVSR task was structured as follows (see Fig 4.2): the trial started with a visual *countdown* consisting of three rectangles decreasing in size at a rate of one rectangle per 500 ms, rendered on a plane which covered the ECA entirely. Afterwards, in the *stimulus presentation phase*, participants saw the ECA speak a sequence of eight digits at a rate of one digit per second (digit word duration: 600 ms, interstimulus interval: 400 ms). In this phase, the ECA spoke the digits by moving its lips in accordance with the sound. The digit

sequences were created from digits between one and nine in such a way that no digit was repeated and no more than two consecutive steps, up or down, occurred (e.g., "1-2-3-4" or "9-8-7-6" would not be allowed). After the stimulus presentation, participants had to wait for a *retention interval* of three seconds before they were asked to reproduce the order of the digits by clicking on the corresponding fields in a matrix, which was rendered as blocks in front of the ECA (see Fig 4.2). During this *recall phase*, the ECA was partially visible and animated with an idle movement. Only the eight digits present in the sequence were displayed in the matrix. The order of the digits within the matrix was changed randomly for each trial to avoid the use of visuo-spatial recall techniques [181]. Once participants clicked on a digit, it disappeared. Corrections were not possible. The dependent variables were the proportion of correctly recalled digits and the response times.

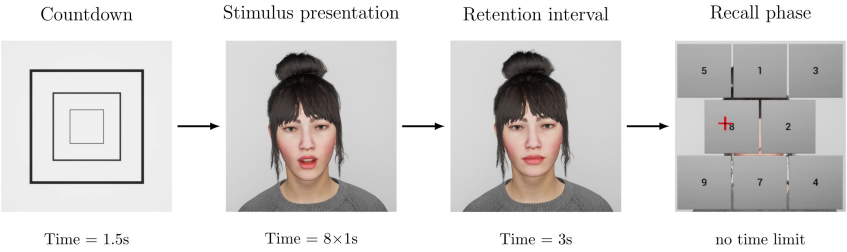


Figure 4.2: One trial of the avVSR task. Graphical depiction of the trial phases *Countdown*, *Stimulus presentation*, *Retention interval*, and *Recall phase* over time. In the *Recall phase*, the cursor is visible as a red cross.

Apparatus and Materials

The avVSR task was configured for both HMD and computer monitor presentations and implemented in Unreal Engine (Epic Games, v4.27) using the following plugins (see Section 4.6 for code repositories): the StudyFramework plugin [113] (v4.26), which handles the data logging and experiment procedure, the RWTH VR Toolkit plugin [157] (v4.27) for basic VR interaction, and the Character plugin (v4.27) for ECA animation. In the HMD condition, the scene was presented using an HTC Vive Pro Eye (dual AMOLED screens with $3.5''$ diagonal display, resolution 1440×1600 pixels per eye) and one HTC Vive controller. In the computer monitor condition, a Fujitsu B24T-7 ($24''$ diagonal display, resolution 1920×1080 pixels) monitor and a wireless computer mouse (Cherry MW2310) were used. To ensure consistency between the presentation modalities, the actual room in which the experiment took place (VR laboratory [353]) was replicated

in VR so that the participants were virtually "in the same room" regardless of the display device used (see Fig 4.3).

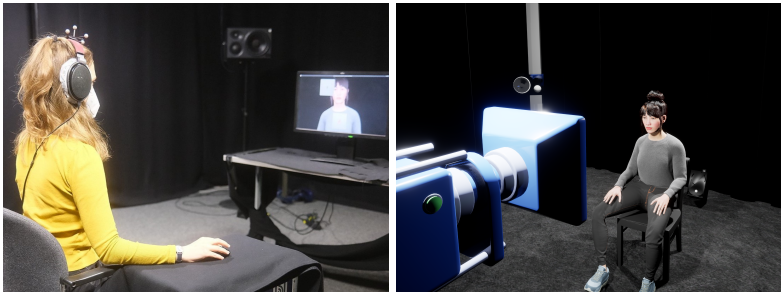


Figure 4.3: Listening experiment setup. The experiment was done with both (left) computer monitor and (right) HMD presentation. *Left*: computer monitor presentation in which a participant is seen wearing headphones with Optitrack tracking markers. *Right*: representation of the VR scene created in Unreal Engine. When wearing the HMD, the point of view of the blue camera represents the participants' point of view.

For the visual stimulus, a female ECA (see Figs 4.2, 4.3) was created with MetaHuman Creator (Epic Games, v0.5.2). Using Oculus Lip Sync (Meta, v20), lip animation was generated from the audio files such that the ECA moved its lips according to the digit sequence heard. In the HMD condition, the ECA sat on a virtual chair at a distance of $d = 2.5\,\mathrm{m}$ from the participant, which is a comfortable inter-person distance to an unknown person [173, 46]. In the computer monitor condition, only the upper part of the body was visible on the computer monitor. This image section and the distance of the computer monitor from the participants were chosen so that the size and position of the ECA were consistent between the two display devices. Participants sat on a chair, which was adjusted so that the participants (measured from the top of the head) and the ECA were approximately at the same height. The two main visual differences between display devices were that in the HMD condition the participants could not see themselves and the ECA's lower body was not visible in the computer monitor condition.

For the acoustic stimulus, German digit words (600 ms in duration) were taken from the database by Oberem and Fels [339] (voice *female b*). The original publication did not include a recording of the digit five, but was made available by the authors. No loudspeakers were used to avoid visual fixation to possible source positions. Instead, virtual sound sources for a binaural headphone-based reproduction were auralized with Virtual Acoustics [209] (v2020a) in a reflection-

free (anechoic) environment using dynamic scene rendering. Communication between Unreal Engine and Virtual Acoustics was established via the Virtual Acoustics Plugin (v4.26) (see Section 4.6). This was achieved by tracking the head movements of the participants and virtually shifting the sound source position accordingly. Head-tracking was done in two different ways depending on the display device: for the HMD, the built-in tracking system was used. This was not possible in the computer monitor presentation. Instead, an OptiTrack system with Motive (OptiTrack, v2.1.1) tracked head movements. To provide plausible localization cues, the HRTF of the Institute for Hearing Technology and Acoustics (IHTA) artificial head [408] was employed. As shown by Oberem et al. [341], this generic HRTF provides sufficiently accurate localization in dynamic scenes for most listeners.

The auditory stimuli were played back using Sennheiser HD650 open-back headphones with a Behringer ADA8200 sound card for both display devices. All auditory stimuli were calibrated to 60 dB(A) as the power sum of both ears using the IHTA artificial head [408]. Headphone equalization was performed individually for each participant following Masiero and Fels [298].

Audiovisual Angle Incongruence

An angle incongruence was introduced to the audiovisual stimulus (audio: spoken digit words, visual: lip movement of ECA) by implementing an angular offset between the auditory source position and the location of the ECA. The position of the ECA as the visual source was kept constant at an azimuth angle of $\varphi = 0°$, straight ahead of the participant (see Fig 4.4). The idea behind the choice of the angular offsets was to include angles small enough to allow the audiovisual stimulus to be perceived as a unit and angles large enough so that an audiovisual angle incongruence is expected to be detectable. Since conversational partners are approximately at the same height in seated settings, we limited our selection to angles on the horizontal plane. Studies on the ventriloquism effect report various threshold angles up to which spatially separated visual and auditory stimuli are integrated. It is suggested that angular offset between auditory and visual source positions should not exceed 15° for audiovisual integration to occur [434, 84], but integration effects up to 30° have been shown on the horizontal plane [212]. In their study on audiovisual congruence in VR, Kim and Lee [235] found that participants generally tolerate angular offsets of up to 30° on the horizontal plane and detect angle offsets of 15° only with a probability of 18.8%. It should be noted that these angular limits are usually determined with simplified audiovisual stimuli such as noise bursts and flashing lights. Thresholds for realistic sounds, such as speech, have hardly been investigated [205].

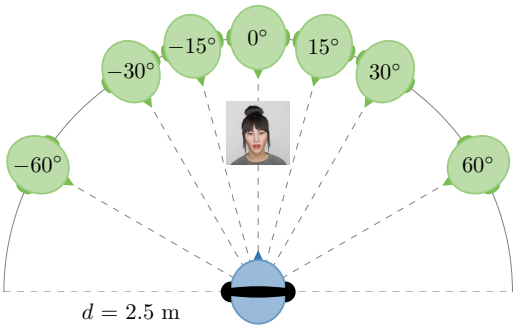


Figure 4.4: Audiovisual angle incongruence in Experiment 1. The female ECA is displayed at an azimuth angle of 0° on the horizontal plane at a distance of $d = 2.5\,\text{m}$ from the listener (blue). The possible virtual sound source positions (green) are at the azimuth angles $0°, \pm 15°, \pm 30°$, and $\pm 60°$ at the distance d.

Since headphone reproduction with a generic HRTF is used here, a localization offset compared to a headphone reproduction with an individual HRTF would be expected. However, Oberem et al. [341] showed that this localization offset is non-significant on the horizontal plane with dynamic reproduction with head movements for the given generic HRTF with speech stimuli.

Based on the above, four angular offsets were chosen for the audiovisual angle incongruence. As a baseline condition, $\Delta\varphi = 0°$ was included, where the position of the auditory target and visual target are aligned. Furthermore, $\Delta\varphi = \pm 15°$, $\pm 30°$, and $\pm 60°$ were included with decreasing probability of integration expected for increasing angular differences.

Questionnaires

To link the behavioral data from the avVSR in the two display devices with subjective impressions and to quantify presence, two distinct questionnaires were used: one questionnaire that had to be completed for each display device and a second general questionnaire. Both questionnaires were implemented using SoSciSurvey in German [263].

In the first questionnaire, the perceived presence during the HMD presentation compared to the computer monitor presentation was evaluated with a German translation [45] of the SUS [462] on a 7-point Likert scale with the corresponding anchors. Additionally, participants were asked about the appearance of the ECA (following Ehret at al. [112]): (**Q__Nat**: *How natural did the interlocutor seem to you?*, **Q__Spe**: *How natural did the way the interlocutor spoke seem to you?*,

Q__Sen: *To which degree did the interlocutor appear to be a sentient being?*). Furthermore, the modality participants focused on (**Q__Vis**: *How much did you focus visually on the interlocutor?*) and the controller handling (**Q__Con**: *How intuitive was the handling of the controller?*) were examined.

At the end of the experiment, participants completed a second questionnaire. This questionnaire focused on the task itself (**Q__Dif**: *How difficult was the task for you?*) and on the audiovisual incongruence (**Q__Inc**: *Did you notice any audiovisual incongruence?* [Asked on a yes/no basis], **Q__Imp**: *If you noticed the audiovisual incongruence, did it impact your performance?*, **Q__Dom**: *If you noticed the audiovisual incongruence, did you shift your attention towards one domain (audio, visual) as a result?*). All these ratings, except for **Q__Inc**, were made on a 7-point Likert scale between 1 (minimum, e.g., "Not at all") and 7 (maximum, e.g., "Strongly"). A comment field was provided for additional remarks. All questions and their German translation with corresponding anchors can be found in S4.

Procedure

The experiment was conducted in individual sessions at the IHTA, RWTH Aachen University. The audiometric and visual screening was performed in a sound isolated hearing booth [353]. After passing the screening, the main experiment started in the VR laboratory [353]. The display device was varied across participants in two blocks, meaning that participants had to complete all trials on one display device before moving to the next one. The starting display device was counterbalanced between participants. At the beginning of the experiment, written instructions explaining the task were given. The participants were instructed to attend to the digit sequences and remember them, without being informed as to whether they were to prioritize the auditory or visual information. Participants were further asked not to vocalize the digits or use their fingers as an aid. The instructions were followed by a training block consisting of eight trials, two for each of the four audiovisual offsets angles ($|\Delta\varphi| = 0°, 15°, 30°, 60°$) in counterbalanced order. The training was followed by twelve trials for each of the four angles, yielding a total of $12 \times 4 = 48$ trials per display device in the main experiment, divided into three blocks of 16 trials each. Each offset angle was presented twice in a block, once from the left and once from the right side; the offset angle of $|\Delta\varphi| = 0°$ was presented twice from the front. The order of angles was counterbalanced across participants to avoid order effects. Between blocks, participants could take a break of no fixed length. After each display device block, participants filled out the first questionnaire. Participants completed the second questionnaire at the end of the experiment. The reason for presenting the second

questionnaire at the end rather than between the display device blocks was to avoid participants' performance being affected by inadvertently disclosing the research questions of the listening experiment. The entire experiment, including screening and questionnaires, took approximately 90 minutes.

Statistical Analysis

All statistical analyses were conducted in the software R (version 4.2.2). The R package *tidyverse* [490] was used for data processing and plots. For the avVSR task, there were two dependent variables (i.e., fixed effects): the *Accuracy* and the average response times *RT_mean*. *RT_mean* is defined as the average time between two button clicks in the recall phase or, for the first digit, the time between the start of the recall phase and the first button click. *Accuracy* is the proportion of digits recalled at the correct serial position. The independent variables included the serial position, the angle (see Fig 4.4), and the display device. Separate Bayesian generalized mixed-effects models were created for each dependent variable using the R package *brms* (version 2.18) [79]. In each model, the repeated-measures experimental design was incorporated as independently varying intercepts (i.e., random effects) for the participants. Since *Accuracy* includes proportion data with many values close or equal to 1, a zero-one-inflated beta distribution with a logit link function was used as the family for the models with this dependent variable. A gamma distribution with a log link function was used as the family for *RT_mean* as the dependent variable since such distributions are commonly used for time data.

The modeling process for the avVSR task included starting with an intercept-only baseline model, followed by including an independent variable or an interaction between the independent variables in subsequent models. The prior distributions per independent variable and their interactions were sampled 16000 times, using four independent chains of 5000 samples each and discarding the first 1000 warm-up samples. To avoid dependencies per chain, No-U-Turn Sampling was employed. Weakly-informative prior distributions (in the logit or log scale, depending on the dependent variable) were used, which for the fixed effects included normal distributions each with a mean (M) of 0 and a standard deviation (SD) of 2. The latter is relatively large given the scales, and hence conservative. For the random effects, heavily tailed Student-t distributions with $M = 0$ and $SD = 2$ each were used. The performance across models was compared using the leave-one-out cross-validation criteria, based on the difference in values of the expected log point wise predictive density and the standard error of the expected log point wise predictive density of the models [429]. Posterior predictive checks were performed for all the models.

Pairwise comparisons across independent variable levels were conducted using the R package *emmeans* [267]. These comparisons are summarized using the median and 95% CIs calculated using the highest density interval of the PPD. The Bayesian 95% CI provides the interval within which 95% of the PPD lies. This interval considers the probability range or the uncertainty/belief regarding the parameter value (e.g., mean), contingent on both the present sample and the prior information. This presents a more direct statement regarding the parameter than the 95% confidence interval used in frequentist statistics. The latter generally means that upon repeated random sampling and calculating the 95% confidence interval per sample, the true parameter value would be contained within approximately 95% of those intervals. This so-called long-run probability takes all values being equally likely a priori and parameter values are then contingent upon the present sample.

To determine whether an effect (comparison across independent variable levels) exists, we consider the probability of direction (PD) of the PPD, which provides the probability of an effect going in either the positive or negative direction. PD further provides a link to frequentist statistics in that the two-sided p-values of .05 and .01 approximately correspond to PD of 97.5% and 99.5%, respectively. PD, however, is not an index of significance. In order to consider whether an effect is meaningful/significant, we use the proportion of the 95% CI inside the ROPE. The latter is the range signifying an effect of negligible magnitude and is conceptually similar to the null hypothesis in frequentist statistics. The range for the ROPE was specified as $\pm 0.1 \times SD_{IV}$ [246]. PD and ROPE were calculated using the R package *bayestestR* [288].

For each of the questionnaire items, separate Bayesian mixed-effects models were created with the display device as the only fixed effect and independently varying intercepts (random effects) for the participants. For all questions except for Q_Inc, the ratings (the dependent variable) were modelled using the cumulative family with a probit link, since the ratings were on an ordinal scale. For Q_Inc, the rating was modelled using the Bernoulli family with a logit link, since the ratings were binary (yes/no). A weakly informative prior with a mean of 0 and a standard deviation of 5 was used for the dependent variables. Other modeling criteria were the same as used for the avVSR data. The differences in ratings between the display devices are summarized using the standard deviation and 95% CI (calculated using the highest density interval of the PPD) for all questions except Q_Inc, for which the change in log-odds of the response and the associated 95% CI is used.

4.3.2 Results

Audiovisual verbal serial recall (avVSR)

The full result data is given in S1, plots of the data can be found in S3. Based on the leave-one-out criteria with *Accuracy* as the dependent variable, the baseline (intercept-only) model was the worst model, while the performance across models with the various independent variables and their interactions were not substantially different. Hence, the model with all the independent variables and an interaction term between the audiovisual offset angle and display device was selected as the final model since it encompasses the experimental design most comprehensively. The interaction term with serial position was not considered useful for the research intend and hence not included. The pairwise comparison for the final model is presented in Table 4.1. The median differences across the

Table 4.1: Summary of pairwise comparisons between the audiovisual Angle offsets (i.e., between the auditory source and the visual representation of the ECA) and the display Device combinations with *Accuracy* as the dependent variable.

Comparison		Median	95% CI	PD	% in ROPE
Angle	**Device**				
$0° - 15°$	Monitor	-0.01	[-0.04, 0.01]	83.88%	72.15%
$0° - 30°$	Monitor	-0.02	[-0.04, 0.01]	91.64%	56.24%
$0° - 60°$	Monitor	-0.03	[-0.06, -0.01]	99.69%	11.01%
$15° - 30°$	Monitor	$-5.13e^{-3}$	[-0.03, 0.02]	65.78%	89.24%
$15° - 60°$	Monitor	-0.02	[-0.05, 0.00]	95.71%	45.07%
$30° - 60°$	Monitor	-0.02	[-0.04, 0.01]	90.73%	61.86%
$0° - 15°$	HMD	$9.44e^{-3}$	[-0.02, 0.03]	76.80%	81.20%
$0° - 30°$	HMD	$2.77e^{-3}$	[-0.02, 0.03]	50.84%	93.01%
$0° - 60°$	HMD	$-7.33e^{-3}$	[-0.03, 0.02]	72.30%	86.04%
$15° - 30°$	HMD	$-9.14e^{-3}$	[-0.03, 0.02]	77.12%	82.05%
$15° - 60°$	HMD	-0.02	[-0.04, 0.01]	91.25%	61.02%
$30° - 60°$	HMD	$-7.67e^{-3}$	[-0.03, 0.02]	72.99%	86.01%

CI = Bayesian credible interval, PD = probability of direction, ROPE = region of practical equivalence

comparisons were generally quite low with PD less than 97.5% except for the difference between the 0° and 60° angles with the computer monitor display device. The latter indicates that there is a higher *Accuracy* in the 60° compared to the 0° condition, suggesting an effect. However, the high percentage of the PPD within the ROPE suggests that this effect is likely of negligible practical significance (i.e.,

negligible effect size), as it falls within the range which is generally considered too small to be meaningful.

For *RT_mean* as the dependent variable, similar to *Accuracy*, the final model included all the independent variables and an interaction term between the audio-visual offset angle and display device. However, none of the pairwise comparisons were statistically meaningful with PPDs with low PD values and high % in ROPE (see Table 1 in S3), and are not considered further.

Questionnaire

The full questionnaire data is given in S5. As seen in Fig 4.5, for all the questions from the SUS questionnaire, the differences in SD between the HMD and computer monitor display conditions were statistically robust, indicating that the participants reported a higher perceived presence in the HMD condition. This is depicted by the 95% CI of the SD difference between display conditions not crossing the zero line, suggesting a meaningful effect.

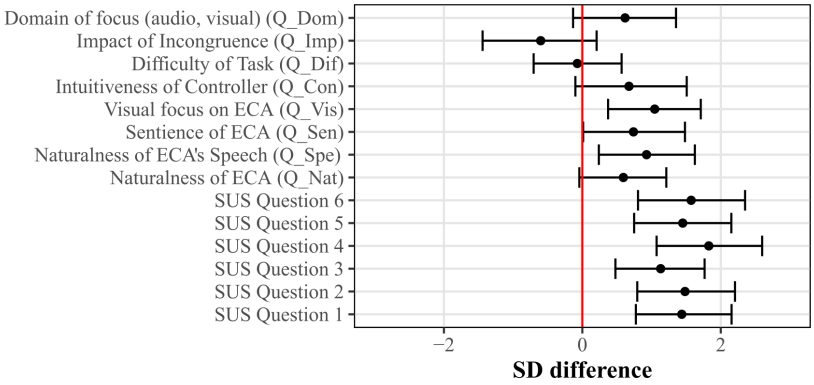


Figure 4.5: Questionnaire results of Experiment 1. The SD difference between the HMD and the Monitor display conditions is shown for each question on the y-axis. Error bars indicate the 95% credible intervals. All questions were rated on a scale of 1 to 7, except for Q_Inc, which was asked on a yes/no basis and is thus not displayed here.

For the other questions, the 95% CI crosses the zero line, indicating that the SD differences of the ratings between display devices were not statistically robust. This holds true for the naturalness of the ECA (Q_Nat), the extent to which the ECA was perceived as a sentient being (Q_Sen), the intuitiveness of the controller handling (Q_Con), the perceived difficulty of the task (Q_Dif),

and the perceived impact of the incongruence on the participants' performance (Q_Imp).

Four of the 23 participants reported not noticing the audiovisual angle incongruence (Q_Inc). The change in the log odds of the perceived audiovisual incongruence (Q_Inc) between the monitor and HMD condition (-1.14) was not statistically robust with the 95% CI ([-4.63, 1.74]) crossing zero. This indicates that the perceived noticeability of the incongruence did not differ between the display devices.

Statistically robust SD differences between ratings of the display devices could be detected for the naturalness of speech (Q_Spe), with a higher rating of naturalness in the HMD condition, and in the visual focus (Q_Vis), where participants reported that they focused significantly more on the ECA in the HMD condition.

4.3.3 Discussion

The participants' avVSR performance was evaluated in terms of *Accuracy* and *RT_mean* for both HMD and computer monitor presentations. No statistically robust influence of the display device on either metric could be found, which is in accordance with the participants' self-assessment that the task difficulty was similar for both display devices. This suggests that the avVSR task can be successfully conducted in VR, at least in conditions close to the given setup. However, while the SUS questionnaire data indicated a significantly higher perceived presence in the HMD condition, this increased presence did not translate into changes in cognitive task performance. This finding highlights that higher perceived presence in VR does not necessarily correspond to enhanced performance in cognitive tasks.

Regarding the impact of audiovisual angle incongruence, a marginal increase in recall performance was observed for an offset angle of 60° in the computer monitor presentation, albeit with a negligible effect size. This is quite interesting as a contrary effect, i.e., a decrease in performance, could have been expected. However, it has been shown before that more demanding tasks can evoke increased performance [294], presumably because participants put in more effort. A similar effect may have occured here, i.e., audiovisual angle incongruence might have had an alerting effect, leading to higher concentration and, thus, improved performance. However, we can only speculate as to why this effect occurred in the computer monitor, but not in the HMD condition. As indicated in the questionnaire (Q_Vis), participants focussed more on the visual signal, the ECA, in the HMD than in the computer monitor condition. It is possible that the visual stimulus dominated over the auditory signal and thus the audiovisual an-

gle incongruence could be ignored better. Furthermore, HMDs influence sound source localization [204] and depth perception [243]. This could have impacted the perception of the audiovisual angle incongruence as well. Since the effect size was rather small, further investigations are necessary to validate it.

For all other combinations of display device and audiovisual angle incongruence, no effect of incongruence on recall performance and response time was detected. This is consistent with participants' ratings, which indicated that they felt only marginally influenced by the audiovisual angle incongruence. One possible explanation for the lack of interference is that participants could largely ignore the incongruence, perhaps because the task-relevant information was primarily auditory, with visual information being redundant.

It is worth noting that the audiovisual incongruence could also be interpreted as a visual distractor. There are only very few studies on the impact of visual distractors on verbal serial recall and visual signals have not been shown to impact recall performance. Liebl et al. [271] investigated relatively rapid changes in workspace lighting, namely the effects of a traveling beam on a projector screen in the participants' field of vision, and found no effect on the visual-verbal serial recall performance. In an experiment with flashing color stripes during stimulus presentation, Lange [255] found no reduction in performance for visual-verbal serial recall. To the best of the authors' knowledge, auditory-verbal serial recall has not been investigated with visual distractors yet. However, since the ineffectiveness of visual distractors on verbal serial recall is argued to be due to the separation of auditory-verbal and visual-spatial short-term memory according to the model of Baddeley and Hitch [25], it is reasonable to assume that the audiovisual incongruence did not act as a performance-relevant distractor in the present study with auditory reproduction either. In summary, our results suggest that participants are able to ignore audiovisual angle incongruence in avVSR tasks, at least in experimental settings resembling the current.

The ratings regarding the general appearance of the ECA did not differ between display devices in a statistically robust way. Similar effects regarding ECAs can be found in literature. McKeown et al. [307] conducted a study where participants rated social interactions presented either on an HMD or a computer monitor. These ratings were not influenced by the display device. In a study on human-ECA interactions, Zojaji et al. [504] found no influence on the perceived persuasiveness and offensiveness between display devices. This suggests that the perception of an ECA may not generally be influenced by the display device. Interestingly, in the present study, the ECA's speech was perceived as more natural in the HMD condition compared to the computer monitor condition, although the exact same material (and in the case of the speech even audio

reproduction) was used. This may be linked to the higher presence overall and higher visual focus reported in the HMD condition.

The experiment had several limitations for both display devices. Firstly, with the employed animation technique (computer-generated lip movement, no blinking, and only idle movement), the subjective evaluation of the naturalness of the ECA was generally low. Based on the general comments, the participants found the ECA's eye movement, or rather the lack thereof, particularly unnatural, which provided an opportunity for improvement. Secondly, regarding the choice of incongruence, the behavioral results indicate that participants were able to ignore it. According to the questionnaire, a total of 13% of participants did not notice the audiovisual angle incongruence at all. It is possible that the selected angle differences were insufficiently large, and that only the 60° condition entered the range of disruptive incongruences. We derived the selection of angles from ventriloquism experiments, in which participants focus on the spatial characteristics of the stimuli since their task is to localize them. In contrast, the perception of spatial direction was incidental in our study, as subjects were instructed to focus on the content, specifically the to-be-remembered digits. This could have shifted the threshold for detection, and/or the need to attend to incongruences. Also, participants were likely experienced in integrating audiovisual angle incongruences. For example, when watching a movie, no directional discrepancy between audio and video is perceived actively, even though loudspeakers are usually placed beside the screen. It remains unclear if the participants were able to ignore the incongruence due to the small angle range and the familiarity of the situation - or if incongruent audiovisual stimuli can be ignored in general if they are not task-relevant.

4.4 Experiment 2: Audiovisual Semantic Incongruence

A limitation of Experiment 1 was that the spatial incongruence went unnoticed by many participants, potentially confounding the findings on audiovisual incongruences in VR. Moreover, the chosen angle incongruence made it challenging to distinguish between the transition from congruent to incongruent cases, as the angle is a continuous variable with no clear threshold for noticeability. To address this issue, a clearly noticeable audiovisual incongruence was considered in Experiment 2. One potential approach to ensure perceptibility is a pre-test with each participant to determine individual angular thresholds of detectibility, similar to the study by Kim and Lee [235]. However, even with this refined approach, we could not be sure that the participants would interpret the spatial incongruence as such. As discussed previously, most people naturally integrate spatial incongruent audiovisual stimuli, especially in entertainment applications. Therefore,

we wanted to use a binary and more easily distinguishable incongruence which allows for clearly comparing congruent with incongruent cases.

For this purpose, in Experiment 2, a voice swap, i.e., two ECAs speaking with each others' voices, was introduced. To allow a clear distinction between the voices, a swap between a male and female ECA was chosen. A voice swap can be considered as a special case of semantic incongruence because it disrupts the expected auditory-visual correspondence, leading to a mismatch between the perceived voice and the visual identity or context. This *audiovisual voice incongruence* was investigated in both HMD and computer monitor presentations. An overview over the experimental design can be found in Fig 4.6. Compared to Experiment 1, the position of the ECA was not changed in this experiment, and hence no spatial incongruence was included.

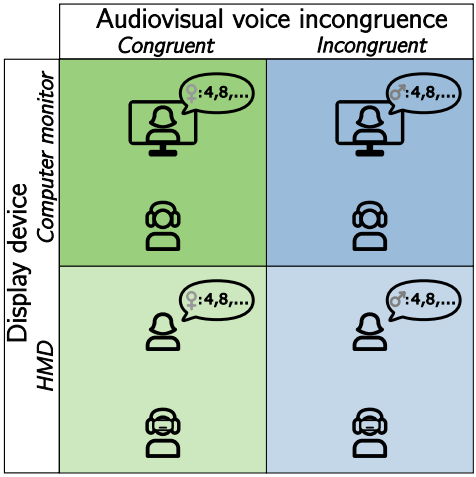


Figure 4.6: Factors in Experiment 2. The between-subject experimental design of the avVSR task with three factors: display device (computer monitor vs. HMD), the audiovisual voice incongruence (congruent vs. incongruent), and the serial position (1 – 8) of the target digit sequence. In the congruent case, the ECA speaks with their own voice; in the incongruent case the female ECA speaks with the male ECAs' voice and vice versa (here shown for the female ECA).

4.4.1 Method

Participants

The recruitment procedure, screening, ethics approval, consent form, and compensation were the same as in Experiment 1. Participants were recruited between

30th of November and 23rd of December 2022. $N = 26$ adults (16 male, 10 female) aged between 21 and 39 years ($M = 28.08$, $SD = 4.39$) participated in Experiment 2. It was not an exclusion criterion if participants had already participated in the first study.

Cognitive Task

The paradigm for the cognitive task was the same as in Experiment 1: an avVSR task in which digit sequences were played back auditorily and an ECA was animated to speak the sentences (see Fig 4.2).

Audiovisual Voice Incongruence

A voice swap was implemented as the audiovisual incongruence. At the beginning of the experiment, two ECAs one female and one male (see Fig 4.7), were introduced with their respective voices. In the experiment, they unexpectedly spoke with the voice of the other ECA, thus, swapping voices for an audiovisual incongruence (male-female, female-male). Voice swaps have been used in cognitive research before. Walker-Andres et al. [481] found that infants as young as six months prefer to look at audiovisual stimuli with congruent gender compared to incongruent gender. In an experiment of just noticeable onset asynchronies of audiovisual stimuli, Vataksi and Spence [468] found a higher sensitivity for gender-incongruent stimuli. They suggested that a incongruence between the auditory and visual stimuli in terms of gender leads to them not being perceived as a unit. A study by Szerszen [449] indicated that a gender incongruence in the voice can lead to an uncanny valley effect. In this experiment, the position of both the auditory and visual sources was kept constant at an azimuth angle of $\varphi = 0°$ on the horizontal plane to avoid the additional effect of spatial incongruence.



Figure 4.7: ECAs used in Experiment 2. *Left:* The female ECA was the same as in Experiment 1. *Right:* A male ECA was created for Experiment 2.

Apparatus and Materials

The same equipment and software as in Experiment 1 were used. A second, male, ECA was created. Since participants reported that the gaze of the ECA seemed unnatural in Experiment 1, an idle eye movement was added for both ECAs using the VHGazeComponent of the Character Plugin in Speech Mode (for more information see [114]). The male voice stimuli were taken from [339] (voice *male a*).

Questionnaires

The questionnaires remained unchanged from Experiment 1.

Procedure

The procedure was mainly the same as in Experiment 1. The training block consisted of six trials (three with the female ECA, three with the male ECA) with matching voices. This way, participants could familiarize themselves with the task and the voice matches. In the main experiment, twelve trials were presented for all four possible combinations of the ECAs and their voices, resulting in $12 \times 4 = 48$ trials per display device, divided into four blocks of twelve trials each. The order of the visual ECA and voice combinations was counterbalanced across participants in each block.

Statistical Analysis

The procedure for the statistical analysis was identical to that in Experiment 1 except for two changes. First, audiovisual voice incongruence was introduced as a binary independent variable. The other two independent variables were the same as in Experiment 1 (display device and serial position), with the dependent variables also being the same (*Accuracy* and *RT_mean*). Secondly, the priors for display device and serial position were set based on the respective PPD from models in Experiment 1.

4.4.2 Results

Audiovisual verbal serial recall (avVSR)

The full result data of Experiment 2 is given in S2, plots of the data can be found in S3. Similar to Experiment 1, the leave-one-out criteria was used to determine the final models. For *Accuracy* as the dependent variable, the model with all the IVs and an interaction term between the audiovisual voice incongruence and display device was selected as the final model since it encompasses the experimental design most comprehensively. The pairwise comparisons for this model are presented

in Table 4.2. While the PD for the Monitor condition was greater than 97.5% (equivalent to a p-value of less than .05), suggesting better performance for incongruent compared to congruent trials, neither of the comparisons here was statistically meaningful with the % of the PPD within ROPE being very high.

Table 4.2: Summary of pairwise comparisons between the voice incongruence (Match) and display Device combinations with *Accuracy* as the dependent variable.

Comparison		Median	95% CI	PD	% in ROPE
Match	Device				
Incong.−Cong.	Monitor	0.02	[0.00, 0.04]	97.71%	58.72%
Incong.−Cong.	HMD	0.01	[0.00, 0.03]	94.80%	72.24%

CI = Bayesian credible interval, PD = probability of direction, ROPE = Region of practical equivalence, Incong. = Incongruent, Cong. = Congruent

For *RT_mean* as the DV, the final model also included all the independent variables and an interaction term between audiovisual voice incongruence and display device. However, as in Experiment 1 and for *Accuracy* as the dependent variable in this experiment, none of the pairwise comparisons were statistically meaningful (see Table 2 in S3), and are not considered further.

Questionnaire

All questionnaire data is given in S5. The results of the questionnaire analysis are displayed in Fig 4.8. As in Experiment 1, the SD differences between the ratings on the SUS questionnaire for the two display conditions were statistically robust (95% CI does not include zero), with the participants reporting higher presence in the HMD vs. the computer monitor condition. The ECAs' naturalness (Q_Nat) was rated higher in the HMD condition.

For the remaining questions, the SD differences between ratings for the conditions were not significantly robust. Only one participant reported not noticing the voice incongruence (Q_Inc). However, the change in the log odds of the perceivability of the voice incongruence (Q_Inc) between the monitor and HMD condition (-0.01) was not statistically robust with the 95% CI ([-4.01, 3.99]) crossing zero. This suggests that the perception of incongruence did not vary across the devices.

Finally, the questionnaire results from Experiment 1 were compared with the results from Experiment 2 to see if the changes in the experimental design (changing the type of audiovisual incongruence, adding eye movements to the ECA animation) had any effect. To that end, the subjective impact of the two types of

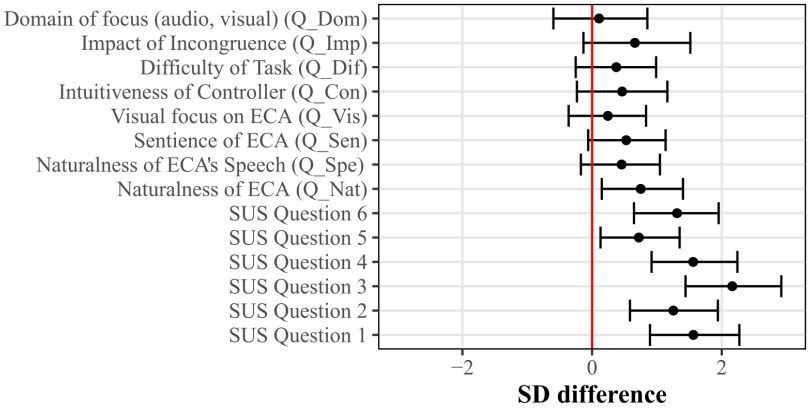


Figure 4.8: Questionnaire results of Experiment 2. The SD difference between the HMD and the Monitor display conditions is shown for each question on the y-axis. Error bars indicate the 95% credible intervals. All questions were rated on a scale of 1 to 7, excluding *Q_Inc*, which was asked on a yes/no basis and is thus not displayed here.

audiovisual incongruences (Q_Imp) and the three questions related to the ECA (Q_Nat, Q_Spe, Q_Sen) were compared between Experiment 1 and Experiment 2 using the same analysis procedure as for the individual questionnaires with the experiment as the dependent variable. The results are displayed in Fig 4.9. Indeed, statistically relevant differences in terms of perceived naturalness (Q_Nat) and sentience (Q_Sen) were found between Experiments 1 and 2, with higher ratings in Experiment 2. Differences in the naturalness of speech (Q_Spe) and the perceived impact of the incongruence (Q_Imp) were not statistically robust.

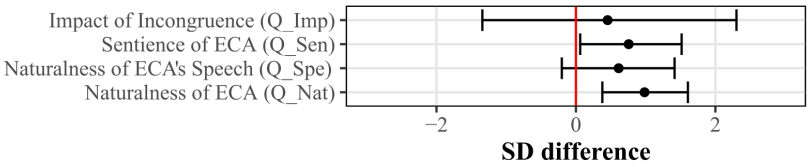


Figure 4.9: Standard deviation (SD) difference of questionnaire data between Experiment 1 and 2. Results are shown for the questions Q_Nat, Q_Sen, Q_Spe, and Q_Imp (on the y-axis). Error bars indicate the 95% credible intervals. All questions (see Section *Questionnaire*) were rated on a scale of 1 to 7.

4.4.3 Discussion

In Experiment 2, task performance in terms of *Accuracy* and *RT_mean* in the serial recall were examined both for HMD and computer monitor presentations. As in Experiment 1, no meaningful influence of the display device on either performance metrics could be found. This again suggests that the avVSR task can be successfully conducted in VR, at least in conditions close to the given setup.

In this experiment, we used a more obvious and binary audiovisual distractor by introducing an audiovisual voice incongruence as a type of semantic incongruence. This reduced the number of participants who failed to notice the incongruence from 13% in Experiment 1 to 4% in Experiment 2. Even though more people noticed the incongruence, there was no effect on the task performance between the two display conditions, which was the common finding for both the experiments. In addition, the ratings in Q_Imp regarding the perceived impact of the incongruence did not differ between the experiments.

In Experiment 1, where audiovisual angle incongruence was investigated, a small increase in recall performance could be found for the audiovisual offset angle of 60° in the computer monitor presentation. We hypothesized this to be due to the alerting effect of the incongruence. In the current experiment employing the audiovisual voice incongruence, again, the *Accuracy* was slightly increased in trials with incongruent voice and the computer monitor as display device. However, this finding was not statistically robust. Even though the type of incongruence was fundamentally different between the experiments, it is reasonable to argue that *Accuracy* is not affected by these incongruences in a statistically relevant way. In other words, the cognitive performance in the avVSR task seems robust against the two incongruences that were tested. While our findings contribute to the understanding of the complex mechanisms underlying the processing of audiovisual stimuli, further research is needed to comprehensively explore the intricate interactions and neural pathways involved.

As in Experiment 1, increased presence could be detected in Experiment 2 for the HMD presentation in the questionnaires. In Experiment 1, the naturalness of speech of the ECA was rated higher in the HMD condition and the ratings of the general naturalness of the ECA was not affected by the display device. This was the other way around in Experiment 2, where the ratings of the naturalness of speech were not affected by the display device, but the general naturalness of the ECA was rated higher in the HMD condition, even though the same lip animation algorithm and speech material were used. This indicates that while the display device does influence the perception of an ECA, the specific nature and nuances of this influence remain unclear. The improved animation of the

ECA led to better ratings with regard to naturalness and the degree to which the ECA is perceived as a sentient being compared to Experiment 1. Adding even more animations to the ECA, like gestures and plausible mimics, could further increase perceived naturalness. It should be noted that participants' ratings were based on their perception of the scene while performing the memory task, which may not be generalizable to other contexts. Still, since differences in the display device became apparent only in the questionnaire and not in the performance data, our study highlights the need for a multidimensional evaluation framework when assessing VR environments.

4.5 Limitations

It is important to note that both the experiments were conducted under relatively simple conditions - specifically, a minimalistic and static laboratory scene featuring only the conversing ECAs as an additional element. Coherent audiovisual cues are likely to play a more critical role when focusing on a target source becomes more difficult in more complex and busier scenes. To gain further insights into how incongruences affect participants, future research should consider increasing the complexity of the scenes by introducing acoustic distractors (such as background noise) and visual distractors (like additional embodied conversational agents or moving scene objects) that bear increasing resemblance to real-life situations. In a busier scene with multiple distractions, participants may find it increasingly challenging to extract and attend to the voice of an ECA if it does not match its visual representation (cp. Experiment 2) or if it is perceived from a different direction (cp. Experiment 1). Another limitation was that no a-priori power calculations were conducted, and the sample size may be considered relatively small. The Bayesian analysis used here can be robust against smaller sample sizes when the priors are informative and the model is specified correctly [308]. The priors used here were, however, weakly/mildly informative since effect sizes from previous studies (not compatible enough with the current experimental design) could not be incorporated within informative priors. Hence, arguably larger sample sizes may be necessary in future studies that can at least benefit from the findings here.

4.6 Conclusion

Two experiments were conducted to investigate the influence of audiovisual spatial (angle) and semantical (voice) incongruence, simulated over two display devices (HMD and computer monitor), on avVSR performance and subjective ratings. With regard to audiovisual incongruences, no clear effect on the proportion of cor-

rectly recalled items (*Accuracy*) or response times (*RT_mean*) could be detected. Some tendencies of increased performance were observed within the computer monitor presentation for the more extreme angle incongruences. The only difference between display devices could be found in the SUS questionnaire, which consistently revealed higher presence when using an HMD compared to a computer monitor in both experiments. Ratings of the naturalness and sentience of the ECAs did not change with the display device.

So far, research on audiovisual incongruence has mostly focused on the perception level, whereas this paper investigated its relevance for verbal short-term memory, which has not been studied previously. Generally speaking, participants were able to ignore the types of audiovisual incongruences presented in the avVSR task. Further research is needed to explore the relation between perceptibility and task impact in more complex scenes, which could provide further insights into the role of audiovisual congruence in cognitive processing and VR design.

Supporting Information[1]

S1 File. **Data of Experiment 1**. *Accuracy* and *RT_mean* for each condition and serial position in Experiment 1 (audiovisual angle incongruence). (CSV)
S2 File. **Data of Experiment 2**. *Accuracy* and *RT_mean* for each condition and serial position in Experiment 2 (audiovisual voice incongruence). (CSV)
S3 File. **Data plots and analysis of RT_mean Experiment 1 and 2**. Plots of *Accuracy* and *RT_mean* as well as results of the statistical analysis of *RT_mean* for Experiment 1 and 2. (PDF)
S4 File. **Questionnaire**. Questions asked in the questionnaires (except for SUS questions) with corresponding anchors. (TXT)
S5 File. **Questionnaire Data Experiment 1 and 2**. Ratings for all questions of the questionnaire for Experiments 1 and 2. (CSV)

Acknowledgments

The authors would like to thank Jamilla Balint, Carolin Breuer, Karin Loh, Lukas Vollmer, and Julia Seitz for the fruitful discussions, and Alissa Wenzel for assisting with the experiment conduction.

[1] To access the supplementary material, please refer to the original publication ([127]).

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation[2]): SPP2236 - Project number 444724862. Manuj Yadav was supported by the DFG: Project number 503914237. Sabine Schlittmeier's contribution to fundraising was supported by a grant from the HEAD-Genuit-Stiftung (Head-Genuit Foundation[3]; P-16/10-W). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

[2] www.dfg.de

[3] www.head-genuit-stiftung.de

5

Heard-text recall and listening effort under irrelevant speech and pseudo-speech in virtual reality

This Chapter was published as an original research article in the journal *Acta Acustica*:
C. A. Ermert, S. J. Schlittmeier, et al. "Heard-Text Recall and Listening Effort under Irrelevant Speech and Pseudo-Speech in Virtual Reality". In: *Acta Acustica* 10 (2026) [125].

5.1 Abstract

Introduction: Verbal communication depends on a listener's ability to accurately comprehend and recall information conveyed in a conversation. The heard-text recall (HTR) paradigm can be used in a dual-task design to assess both memory performance and listening effort. The HTR paradigm uses running speech to simulate a conversation between two talkers. Thereby, it allows for talker visualization in virtual reality (VR), providing co-verbal visual cues like lip-movements, turn-taking cues, and gaze behavior. While the HTR in a dual-task design has been investigated under pink noise, the impact of more realistic irrelevant stimuli, such as speech, that provide temporal fluctuations and meaning compared to noise, remains unexplored.

Methods: In this study ($N = 24$), the HTR task as primary task was administered in an immersive VR environment under three noise conditions: silence, pseudo-speech, and speech. Participants performed a vibrotactile secondary task to quantify listening effort in a dual-task design.

Results: The results indicate an effect of irrelevant speech on memory and speech comprehension as well as secondary task performance, with a stronger impact of speech relative to pseudo-speech.

Discussion: The study validates the sensitivity of the HTR in a dual-task design to background speech stimuli and highlights the relevance of linguistic interference-by-process for listening effort, speech comprehension, and memory.

5.2 Introduction

Listening to and remembering speech is a demanding task, especially in listening scenarios with background noise. Background noise can give rise to listening effort, even if the target speech may still be perfectly intelligible. Listening effort generally refers to the cognitive resources allocated to understanding speech, especially under adverse conditions [150], and is defined as the "deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task" (p.10) [361]. Importantly, compared to, e.g., speech reception thresholds, listening effort can be measured at suprathreshold listening conditions - which is a lot more common for everyday listening scenarios [497], benefiting ecological validity [232].

Listening effort can be modulated by a multitude of factors [266, for a review], e.g., the intelligibility of the target signal [503, 486, 245], room acoustic reverberation [382], target location changes [239], the listeners' age [103, 415], or hearing impairment and hearing device algorithms [443, 494, 365, 194, 9, 247, 345]. But also different types of background noise impose qualitatively distinct challenges: from an acoustic perspective, a distinction is made between energetic masking (e.g., white noise, speech-shaped noise), where the target speech signal is obscured due to an energy overlap in relevant frequency bands [74], and information masking (e.g., speech, multi-talker babble), which occurs when listeners fail to successfully segregate target and masker streams or to allocate attention effectively to the target [423, 93, 233]. Informational masking, therefore, refers to perceptual and attentional challenges that go beyond spectrotemporal overlap. From a cognitive-psychological perspective, meaningful background speech as an informational masker poses yet another source of disruption: because it is processed up to the semantic level, it can interfere directly with comprehension and memory of the target material – a mechanism described as interference-by-process [203, 293].

In listening effort experiments, meaningful background speech almost invariably elicits larger listening effort than signals without linguistic meaning: this has been reported for irrelevant one-talker speech [240, 346, 487, 469, 470, 440, 241], two-talker babble [103, 141], as well as three- to four-talker babble [344, 459, 252]. When several speech maskers are compared directly, research suggests that, after controlling for the signal-to-noise ratio, effort increases as the number of competing voices decreases [213, 487]. The effect is likely two-fold: fewer

voices might preserve more prosodic and semantic information and therefore induce stronger informational masking, but also allow for in-the-dip listening [233], which represents an energetic masking benefit. Even when speech is manipulated to be unintelligible, e.g., through time-reversal, it tends to evoke more effort than amplitude-modulated or stationary speech-shaped noise, underscoring the role of temporal and spectro-temporal variability [73]. Amplitude-modulated speech-shaped noise occupies an intermediate position: several studies observed greater listening effort for modulated than stationary speech-shaped noise, but less than for intelligible speech [241, 309, 244, 192]. As a general trend, listening effort seems to be modulated by the informational content and the time-variance of the background speech, e.g., allowing for catching parts of the target sentence in energy dips of the background speech [257, 139, 393, 381].

There are multiple ways to quantify listening effort [303, 356]; it can be measured subjectively, e.g., via questionnaires [418, 337, 503], through physiological measures, e.g., via pupillometry [503, 247] and neural imaging [40, 106], or behaviorally, e.g., via dual-task paradigms [105, 499, 365]. Dual-task paradigms are based on the theory that the total cognitive processing resources are limited in capacity [65, 230], leading to performance limitations when two tasks demand the same processing capacities [150, 248]. Participants perform a primary listening task while at the same time attending to a secondary, concurrent task (e.g., reaction-time measurements or categorization tasks). Due to the limitation of cognitive resources, increased effort is evidenced by compromised secondary task performance as speech processing demands in the primary task increase [162, 150, 248]. To assess this decline in secondary task performance, both tasks are administered in isolation in single-tasking (i.e., only one task at a time; the primary task alone and the secondary task alone) and together as a dual-task (i.e., both tasks at the same time; the primary task and the secondary task together). Performance metrics are assessed for both tasks (primary task, secondary task) in both conditions (single-tasking, dual-tasking).

Dual-task designs with running speech are rare in the literature; most paradigms use sentence-, word-, or syllable-recognition tasks [150], which does not reflect speech processing in realistic listening situations. The recently developed HTR by Schlittmeier et al. [406, 135], in which participants listen to short, coherent texts and are asked content-related questions afterwards, assesses text comprehension and verbal memory. When implemented in a dual-task design in combination with a secondary task, listening effort in running speech can be assessed in addition. The paradigm can also be embedded into VR environments [114, 115, 399], where ECAs are animated to speak the HTR texts, effectively conveying co-verbal visual cues like lip movements, turn-taking cues, and gaze behavior. Because the primary HTR task explicitly requires semantic comprehension and

memory, it should be especially susceptible to meaningful background speech that induces interference-by-process. In contrast, a non-linguistic secondary task (e.g., vibrotactile pattern discrimination) serving as the indicator of listening effort is not directly vulnerable to linguistic interference; decrements in performance are expected to reflect competition between limited cognitive resources [448] rather than semantic processing per se. Differences between single- and dual-tasking further reflect the listening effort exhibited in the primary listening task. In the present experiment, the HTR served as the primary task within a dual-task paradigm, combined with a non-linguistic vibrotactile pattern-discrimination secondary task.

Mohanathasan et al. [320] investigated the influence of stationary pink noise with a challenging and a moderate signal-to-noise ratio on the primary HTR task and a secondary vibrotactile task. They demonstrated an effect of background noise on the primary HTR task only in the challenging signal-to-noise ratio and a sensitivity of the secondary vibrotactile task response times to both signal-to-noise ratios, while the performance in terms of error rates was unaffected. While studies employing a more realistic background noise (classroom noise) using the same dual-task exist [115, 399], these studies did not explicitly focus on the effect of noise variations. Thus, how different maskers (e.g., intelligible vs. unintelligible speech maskers) influence both performance and effort within the specified dual-task paradigm is still insufficiently explored. In the present study, building upon the VR implementation by Ehret et al. [114], we test whether maskers differing in semantic content and meaningfulness but matched in spectrotemporal structure (intelligible speech vs. pseudo-speech) [123], differentially affect primary HTR and secondary vibrotactile task performance.

We aim to disentangle effects of energetic masking (expected for both speech and pseudo-speech) from additional linguistic interference (expected only for intelligible speech) in the primary HTR task, and to test whether such additional interference indirectly taxes resources and thereby impairs secondary task performance in dual-tasking. Based on these considerations, we hypothesized that primary HTR performance would be best in silence, reduced under pseudo-speech due to energetic masking, and most strongly impaired under intelligible speech due to concurrent linguistic interference-by-process and increased informational masking.

For the dual-task condition, we further expected the pattern of performance decrements in the primary HTR task to be mirrored in the secondary task performance (i.e., decreased accuracy and longer response times) as an indicator of listening effort. As more cognitive resources are drawn upon to maintain text comprehension and verbal memory in the primary HTR task under energetic and concurrent linguistic processing, fewer cognitive resources remain available for

the secondary vibrotactile task. Even though the vibrotactile task is non-auditory and thus does not require semantic processing, background noise might influence the performance even in single-tasking [320] due to noise consuming mental resources and increasing stress [448]. To ensure that the performance decrements of the vibrotactile secondary task in dual-tasking can be attributed to increased listening effort in the primary task instead of general noise effects [448], and to obtain baseline performance, both tasks are also administered in single-tasking.

5.3 Method

5.3.1 Participants

We conducted a Bayesian design analysis for the present experiment design. Assuming a moderate main effect of the *masker condition*, a strong main effect of *number of tasks* [320, 135], and a small interaction, we performed a Bayesian assurance analysis in R (version 4.4.1) [373]. More details regarding the models can be found in the Data Analysis Section. The minimal sample size required to achieve at least 80% assurance was $N = 21$, to accommodate potential data loss and for a fully balanced design, we recruited $N = 24$ participants. Participants were adults (9 female) and aged 22-33 years ($M = 25.2$, $SD = 2.6$). They had German as their first language, as the HTR task is only available in German at this point. Participants had to pass an audiometry (below 25 dBHL between 125Hz and 8kHz; pulsed pure-tone) and a Snellen Test (20/30) [435] to verify their visual acuity. They received a 10-euro voucher for a local bookstore as compensation for their time.

5.3.2 Dual-Task Paradigm

The dual-task paradigm consisted of a primary HTR and a secondary vibrotactile task. In the primary HTR [406] task, participants listened to and saw two ECAs narrate family stories in German consisting of ten sentences each. After the presentation of each text, nine content-related questions were displayed sequentially on a virtual television (see Figure 5.1), and participants had to answer them vocally. An instructor listened to the participants' input and categorized the answer as correct or incorrect based on the ideal solution provided in the HTR database by Schlittmeier et al. [406], detailing all the texts, questions, and answers. By evaluating the number of correct responses, speech comprehension and memory could be examined [406].

The two agents took turns between sentences while narrating the stories. This turn-taking behavior was implemented according to the suggestions by Ermert et al. [129]: successive sentences that form a unit of meaning were assigned to



Figure 5.1: Virtual reality setup with two ECAs narrating the HTR stories, and the television displaying instructions and the HTR questions whenever appropriate.

one ECA before yielding the turn. Thus, the number of sentences spoken by each agent was pseudo-random, but the speaking time per text was approximately the same. Agents never interrupted each other. The texts had an average presentation duration of $M = 43.4\,\text{s}$, $SD = 1.9\,\text{s}$. The audio stimuli for the target sentences of the HTR were taken from the audio-visual speech and text database (AuViST) [129], which provides audio recordings and face tracking data for all sentences of the HTR.

The secondary task assessing listening effort was vibrotactile [320]. Participants held one HTC Vive controller (High Tech Computer Corporation, Taiwan) in each hand. The controllers vibrated in four distinct patterns: short-short, long-long, short-long, and long-short. If the second vibration was a repetition of the first vibration (i.e., short-short, long-long), participants had to click a designated button on the left controller. Otherwise, they had to click a designated button on the right controller [114].

Both tasks were administered together (dual-task) and also in isolation (single-task) to obtain a baseline performance.

5.3.3 Masker Conditions

Three masker conditions were examined: silence (no masking), pseudo-speech (primarily energetic masking), and speech (energetic and informational masking). In the silence condition, only the target talkers were audible. In the pseudo-speech and speech conditions, a male and a female distractor talker spoke simultaneously and without breaks, in addition to the target talkers.

The stimuli in both the background speech conditions were obtained from female [480] and male [479] voice recordings of the German matrix sentences test [478, 476, 477]. In the pseudo-speech condition, the matrix sentences were cut into syllables, one-syllable words were excluded, and the remaining syllables were rearranged in random order, ensuring that no meaningful words were formed. Crossfading of 10 ms was applied to smooth transitions (cf. [128]). In the speech condition, the original matrix sentences were played back in random order.

While other energetic maskers could have been possible, e.g., steady-state speech-shaped noise, time-modulated speech-shaped noise, foreign language recordings, or time-reversed target speech, none of them would have retained the same frequency spectrum of the speech condition while also preserving the temporal variations. Thus, differences in energetic masking compared to the speech condition could be possible [233]. Further, we aimed for a speech-like signal to increase realism of the background sounds compared to the steady-state noise signals used in the study by Mohanathasan et al. [320]. Therefore, the pseudo-speech signal described here was deemed a suitable solution. The stimuli were calibrated using an HMS III artificial head (HEAD acoustics GmbH, Germany) to 60 dB(A) for the target and 57 dB(A) for the background.

5.3.4 Audio-Visual Virtual Reality Setup

The experiment was conducted in a living room VR environment, which was presented via an HTC Vive Pro Eye head-mounted display (High Tech Computer Corporation, Taiwan) and created in Unreal Engine 5.3 (Epic Games Inc., United States of America). In the VR environment, participants saw two ECAs [83], one male and one female MetaHuman (Epic Games Inc., United States of America), narrating the HTR stories as a conversation. The ECAs were animated with gestures and gazing based on Ehret et al. [114] with synchronized face-tracked lip-movement from the audio-visual speech and text database [129]. All agent movement was implemented using the Character Plugin [472]. The study flow was controlled using the StudyFramework Plugin for Unreal Engine [113]. Further, the RWTH VR Toolkit Plugin was used to implement the participant interaction [157]. A video of an exemplary HTR text is provided in the publication by Ehret et al. [114] (supplementary video, condition "$\mathrm{T_{full}}$" at $t = 3:23$ min).

The ECAs, as target talkers, were located at a distance of 1.35 m and a horizontal offset of $\pm 45^\circ$ from the frontal direction. Invisible but audible distractor sound sources for pseudo-speech and speech conditions were located at $\pm 90^\circ$ from the frontal direction, i.e., to the left and right (see Figure 5.2). The gender of the target and background was matched, i.e., the distractor sound source at -90°

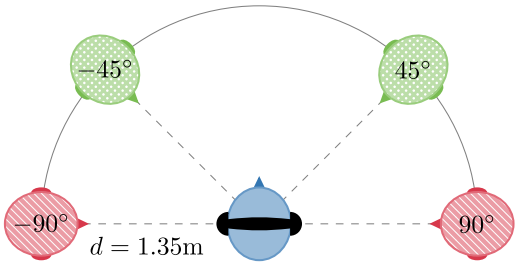


Figure 5.2: Auditory virtual setup with two virtual target sources ( ) and two invisible distractor sources ( ).

played irrelevant speech in a male voice, and the sound source at +90° in a female voice.

The auditory scene was created in Virtual Acoustics v2022a [209] in combination with the Virtual Acoustics plugin for Unreal Engine [471] and reproduced via Sennheiser HD650 headphones (Sennheiser electronic SE & Co. KG, Germany) using a generic HRTF [408] and individual headphone equalization [298]. Using head tracking via the head-mounted display, dynamic scene rendering was employed.

5.3.5 Procedure

The experiment was performed in a sound-isolated booth. Participants sat in a chair. After passing the audiometry and vision test, participants were introduced to the head-mounted display and the virtual environment. Then, the training started. First, the secondary vibrotactile task was administered in silence. Participants completed at least 20 trials but could repeat the task as often as desired until they felt secure. Feedback was given automatically on the virtual television screen. Afterwards, the primary HTR task was trained in single tasking: one HTR text was administered without the secondary vibrotactile task. Participants answered the questions displayed on the virtual television screen. They practiced one more HTR text in the dual-tasking condition together with the secondary vibrotactile task. No background noise was played during training. Then, the main experiment started. No feedback was given in the main experiment. All participants performed both tasks under all combinations of the variables *number of tasks* (dual-task, single-task) and *masker condition* (silence, pseudo-speech, speech) in a within-subject design. The experiment consisted of three blocks: one per masker condition. The order of the noise blocks was counterbalanced using balanced Latin squares. At the beginning of each block, participants could

listen to the noise to familiarize themselves with it. In each block, four HTR texts in dual-tasking (together with the secondary vibrotactile task), one HTR text in single-tasking, and 40 trials of the vibrotactile task in single-tasking were administered. The order of these tasks was balanced within the noise blocks. Texts 1–17 [406, 129] were used in random order. No texts were repeated. Participants could take breaks of no fixed length between noise blocks. At the end of the experiment, participants filled out a questionnaire regarding the perceived ecological validity of the experiment. As this aspect is not the focus of the present research question, no further details about the questionnaire are reported here. The experiment took around 90 minutes, of which around 60 minutes were spent immersed in the VR environment.

5.3.6 Data Analysis

The statistical analysis was carried out in R (version 4.4.1) [373]. The three outcome measures were the binary *performance* in the primary HTR task (1 = answer correct, 0 = answer incorrect), the binary *performance* in the secondary vibrotactile pattern categorization task (1 = answer correct, 0 = answer incorrect or missed), and the *response time* in the secondary vibrotactile task (cf. [320]). The independent variables were *number of tasks* (dual-task, single-task) and *masker condition* (silence, pseudo-speech, speech). Although participants were encouraged to prioritize the primary HTR task over the secondary vibrotactile task, some neglected the secondary task to such an extent that it raises concerns about the validity of the corresponding data. Therefore, a miss rate was calculated for each participant, and outliers were identified using the *is_outlier()* function from the *rstatix* package (version 0.7.2) [231] and excluded from the analysis. This procedure resulted in the exclusion of three participants.

Separate Bayesian generalized mixed-effects models were fitted for each of the three outcome measures using the R package *brms* (version 2.18) [79]. The best-fitting model was determined in a step-wise procedure, starting with an intercept-only baseline model. In subsequent models, the random intercepts *participant* and *question* (combination of text and question number for the primary HTR task) and the independent variables *number of tasks* and *masker condition*, as well as their interaction, were incorporated. The prior distributions were sampled 16000 times, employing No-U-Turn Sampling and using four separate chains of 5000 samples each, with the initial 1000 warm-up samples discarded for each independent variable and their interactions. A Bernoulli distribution on the logit scale was used for the *performance* and a Gamma distribution on the log scale for the *response time*. Priors were weakly informative, i.e., normal distributions each with a mean $M = 0$ and a standard deviation $SD = 2$ for the independent

variables and heavily tailed Student-t distributions with $M = 0$ and $SD = 2$ for the random intercepts. Using the leave-one-out cross-validation criterion, performance across models was compared based on the differences in elpd and the standard error of the elpd of the models [429]. For all three independent variables, *performance*, *question* (for the primary HTR task), *number of tasks*, and *masker condition* improved the model fit. Models including the interaction between *number of tasks* and *masker condition* did not differ in a meaningful way from additive models for all three variables ($elpd_{diff} < 2$). Thus, the interaction model was chosen to comprehensively encompass the experimental design.

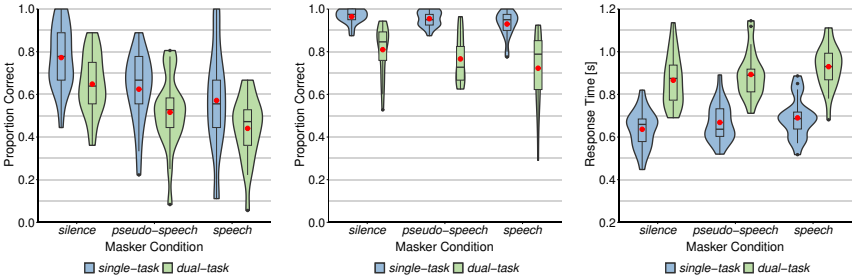
Meaningful main effects or interactions detected in these models were explored further with post-hoc pairwise comparisons of estimated marginal means, conducted using the emmeans package [267] (v 1.10.4), and further summarized using the median and 95% CIs calculated using the HDI of the PPD. The PD of the PPD and the proportion of the 95% CI inside the ROPE (defined as $\pm 0.1 \times SD$ [246]) were used to quantify effects. The PD of the PPD indicates whether an effect exists, with a PD roughly of 97.5% roughly corresponding to two-sided p-values of .05. As PD is not a clear index of a meaningful effect, the proportion of the 95% CI inside the ROPE was employed to signify an effect of negligible magnitude.

5.4 Results

The accumulated performances for the primary HTR task, secondary vibrotactile task, and response times in the secondary vibrotactile task are displayed in Figure 5.3. The pairwise comparisons for the final models are presented in Table 5.1. The pairwise comparisons for *number of tasks* indicated a meaningful effect for all three masker conditions and all three outcome measures with PDs $> 99\%$ and 0% in ROPE.

Regarding the comparisons of *masker condition* in the primary HTR task, PDs were generally quite high ($> 97.5\%$), indicating the existence of an effect. This was supported by the low percentage within ROPE across conditions. An exception was the comparison of pseudo-speech and speech in the *single-task* condition with a PD of 78.08%.

In the secondary vibrotactile task, PDs for all comparisons of *masker condition* were high, and the percentage of PPD within ROPE was low across conditions; except for the comparisons between speech and pseudo-speech in the *single-task* condition with a PD of 84.80% and a high percentage within ROPE, and the comparison between pseudo-speech and speech, in which the 95% CI encompasses the null. Thus, no clear evidence for an effect between these conditions could be gathered.



(a) Primary HTR task

(b) Secondary Vibrotactile Task

(c) Secondary Vibrotactile Task Response Times

Figure 5.3: Results of the (a) primary HTR task, (b) secondary vibrotactile task, and (c) response times in the secondary vibrotactile task. Proportion correct (for the primary and secondary vibrotactile task) and accumulated response times in seconds (for the secondary vibrotactile task) are plotted as a function of the *masker condition* (silence, pseudo-speech, speech) and the *number of tasks* (*single-task*, *dual-task*), averaged per participant. The boxes represent the interquartile ranges, while the median is indicated with a horizontal line and the mean with a red circle.

Lastly, most PD of the response times in the secondary vibrotactile task were over 95.7% combined with low percentages within ROPE, signifying an effect. Only the comparison between pseudo-speech and speech in the *single-task* evoked a PD below 97.5%, and the CI encompasses the null, indicating the absence of an effect.

5.5 Discussion

As evident in other studies employing the HTR in a dual-task design [135, 399, 320], performance in dual-tasking was worse than in single-tasking in all three outcome measures (primary HTR task performance, secondary vibrotactile task performance, secondary vibrotactile task response time), reflecting a dual-task cost.

Our hypothesis that primary HTR performance would be best in silence, reduced under pseudo-speech, and most strongly impaired under intelligible speech in both single- and dual-tasking, was supported by the results: a high dependency on the masker condition could be detected in the primary HTR task. Short-term memory of verbal information, which the primary HTR task depends on, has been demonstrated to be impacted more strongly by meaningful than foreign speech or pseudo-speech that does not convey meaning [500, 128]. Similar results can

Table 5.1: Summary of pairwise comparisons of the combinations of *masker condition* and *number of tasks* in the primary HTR and the secondary vibrotactile task. Meaningful effects are indicated in **bold.**

	Masker Condition	Number of Tasks	Median	95% CI	PD	% in ROPE
Primary Task	silence – pseudo-speech	single-task	0.18	[0.08, 0.28]	**99.99%**	**0%**
	silence – speech		0.22	[0.11, 0.32]	**100%**	**0%**
	pseudo-speech – speech		0.04	[-0.06, 0.15]	78.07%	14.09%
	silence – pseudo-speech	dual-task	0.17	[0.11, 0.22]	**100%**	**0%**
	silence – speech		0.25	[0.19, 0.30]	**100%**	**0%**
	pseudo-speech – speech		0.07	[0.02, 0.13]	**99.56%**	**0%**
	silence	single-task – dual-task	0.13	[0.05, 0.20]	**99.92%**	**0%**
	pseudo-speech		0.12	[0.03, 0.21]	**99.55%**	**0%**
	speech		0.15	[0.06, 0.24]	**99.98%**	**0%**
Secondary Task	silence – pseudo-speech	single-task	8.76e-03	[-0.01, 0.03]	84.80%	39.59%
	silence – speech		0.03	[0.01, 0.05]	**99.96%**	**0%**
	pseudo-speech – speech		0.02	[0.00, 0.05]	98.47%	3.91%
	silence – pseudo-speech	dual-task	0.04	[0.01, 0.08]	**99.64%**	**0%**
	silence – speech		0.09	[0.05, 0.13]	**100%**	**0%**
	pseudo-speech – speech		0.04	[0.01, 0.08]	**99.22%**	**0%**
	silence	single-task – dual-task	0.15	[0.11, 0.19]	**100%**	**0%**
	pseudo-speech		0.18	[0.14, 0.23]	**100%**	**0%**
	speech		0.20	[0.16, 0.25]	**100%**	**0%**
Secondary Task Response Time	silence – pseudo-speech	single-task	-31.97	[-54.57, -11.05]	**99.83%**	**0%**
	silence – speech		-51.92	[-74.50, -30.41]	**100%**	**0%**
	pseudo-speech – speech		-19.95	[-41.43, 3.39]	96.08%	11.63%
	silence – pseudo-speech	dual-task	-32.85	[-57.57, -5.70]	**99.29%**	**0.41%**
	silence – speech		-65.21	[-92.51, -37.85]	**100%**	**0%**
	pseudo-speech – speech		-32.39	[-60.84, -4.87]	**99.01%**	**1.45%**
	silence	single-task – dual-task	-225.75	[-250.90, -200.42]	**100%**	**0%**
	pseudo-speech		-226.68	[-252.50, -200.16]	**100%**	**0%**
	speech		-238.89	[-267.10, -211.45]	**100%**	**0%**

be found for speech comprehension [213, 150]. The comparison between pseudo-speech and speech was not meaningful only in single-tasking. It is possible that the difference between these two markers was not as pronounced in single-tasking as in dual-tasking, as more cognitive resources were available for processing. Further, the limited number of trials in the single-task condition compared to the dual-task condition might not have been sensitive enough to account for the difference between pseudo-speech and speech. The current design was chosen to minimize variance within the dual-task condition, as these conditions were the central focus of our research hypotheses, and in accordance with previous HTR studies [135, 399, 320]. For future studies, a more balanced number of trials should be aimed for.

We further hypothesized that if more cognitive resources are drawn upon to maintain primary task performance, the performance in the secondary vibrotactile task would be affected in dual-tasking. The results generally suggested that listening in the speech condition elicited the worst secondary task performance and the highest response times, while performance in silence was best. This is supported by literature, where listening effort has been shown to be higher if the target speech is degraded by speech than by non-speech or pseudo-speech stimuli [470, 103, 150].

It should be noted, however, that this difference is rather subtle, as seen in the median in Figure 5.3. Further, performance in single-tasking was comparable to dual-tasking, hinting towards a general noise effect within the vibrotactile secondary task, potentially due to the maskers drawing upon mental resources [448]. This is further supported by the results of Mohanathasan et al. [320], who also found a sensitivity of the secondary task towards noise in single-tasking. Listening effort should be primarily apparent in the differences of the masker conditions between single-tasking and dual-tasking. There were some comparisons that did not yield meaningfulness in single-tasking, but were significant in dual-tasking, indicating increased listening effort. For example, we could not detect a meaningful difference between silence and pseudo-speech, as well as between pseudo-speech and speech, in the secondary vibrotactile task performance in single-tasking, contrary to dual-tasking, while the response times mirrored the pattern of the primary HTR task. In dual-tasking, all three comparisons were meaningful in both behavioral metrics. This suggests that the added cost of processing the meaningful speech signal compared to the pseudo-speech signal leads to more errors and longer response times in dual-tasking and thus increased listening effort.

These results should, however, be taken with a grain of salt. Generally, rather few mistakes were made in the secondary task in single-tasking in general, as visible in Figure 5.3b, indicating a ceiling effect in the secondary task performance. This is comparable to participants' performance in Mohanathasan et al.[320]: albeit in a non-VR scenario, they report average performances of close to one in single-tasking. Schiller et al. [399] report a similar high secondary vibrotactile task performance in single-tasking in VR. Further, as discussed before, the number of single-tasking trials in the primary task was limited.

Listening-effort effects may be more pronounced when the primary task is strongly prioritized. In canonical dual-task paradigms, participants are instructed to maintain performance on the primary listening task across conditions, so that any decrements are expressed in the secondary task [150]. We did not observe this pattern: performance on the primary HTR task was affected by both the masker condition and the number of concurrent tasks. This likely attenuated listening-effort effects in the secondary task. We can only hypothesize why this was the case. One possibility is a timing difference between the tasks: information in the primary HTR task had to be retained until the end of the text (with no predefined memory target), whereas the vibrotactile task required immediate responses. This may have made it easier to prioritize the secondary task over the primary HTR task, with participants only realizing insufficient prioritization when answering the content-related questions after each trial. Notably, three participants exhibited exceptionally high proportions of missed responses in the

vibrotactile task, suggesting that some may have abandoned the secondary task in most trials. The choice of a suitable secondary vibrotactile task is an intricate pursuit, balancing factors such as task difficulty, timing, and modality [150]. While the secondary vibrotactile task is advantageous in the sense that it is decoupled from the audio-visual VR scene (cf. [320]), an improved balance between the difficulty of the primary and secondary vibrotactile task could be aimed for.

The HTR task together with a vibrotactile secondary task has been examined by Mohanathasan et al. [320] under pink stationary noise at two noise conditions: a moderate and challenging signal-to-noise ratio. The primary HTR task was sensitive only to the challenging signal-to-noise ratio, and the response times in the secondary vibrotactile task were sensitive to both signal-to-noise ratios. The secondary vibrotactile task performance in terms of errors was, in contrast, unaffected. It is possible that the stationary noise conditions in Mohanathasan et al. [320] were not strong enough to elicit clear differences in all three outcome measures. In the present study, with the pseudo-speech and speech as more challenging auditory markers, all three outcome measures were affected. Speech comprehension and memory of running speech were worse under intelligible speech than under pseudo-speech and best in silence. While differences in vibrotactile secondary task performance, both with respect to the accuracy and response times between masker conditions could be shown, these indicate listening effort only to a limited extent and rather show that also in the secondary task, noise consumes cognitive resources and triggers compensatory efforts [448]. Still, small differences between masker effects when comparing single-tasking and dual-tasking indicate that listening effort rises for speech compared to pseudo-speech.

5.6 Limitations

While an effort was made to provide a realistic listening scenario with spatialized audio and co-verbal visual cues, there are still some design factors that provide room for improvement when aiming for complex, realistic scenarios. As has been shown in literature, the processing of audio-visual speech conveyed by virtual agents is different than that of real-talkers [337, 104, 185], and the naturalness of the animation of the agents could still be improved [114]. The background talkers were not visualized. This was done to investigate the effect of background speech in isolation from possible confounding effects of visual distractors. This approach may have presented certain plausibility and ecological validity challenges, potentially limiting the transferability of the results to realistic scenarios. For example, it might have increased task demands due to the added burden of resolving sensory mismatches [420]. However, Ehret et al. [115] investigated the importance of embodied background noise sources with a dual-task paradigm

using the HTR as primary task and a vibrotactile dual-task and found that listening effort and memory were not influenced by the visualization or absence of embodied distractors.

Furthermore, room acoustic effects were not included to avoid distinct signal-to-noise ratios for each target source due to differential effects of room reflections, e.g., by varying distance to the walls or room modes. As reveberation can be, however, a relevant modulator of listening effort [382, 363, 362, 372, 415], they should be included in further studies.

As discussed, some comparisons might not have yielded meaningfulness due to the limited number of datapoints (i.e., only nine questions in the HTR in single-tasking). This decision was made to reduce variance in the dual-task condition while balancing time- and resource-demands of the participants. More trials, or a more balanced distribution of HTR texts between single- and dual-tasking, are recommended for future studies.

5.7 Conclusion

The HTR paradigm by Schlittmeier et al. [406] in which participants listen to short stories and have to remember the content, measuring speech comprehension, memory, and - when combined with a secondary task - listening effort, was examined under three types of background noise: silence, pseudo-speech without meaning, and intelligible meaningful speech. A plausible scenario was created by presenting an audio-visual VR scene, in which ECAs spoke the HTR sentences. Memory of running speech in the primary HTR task was best in silence and worse under meaningful speech, replicating established findings from paradigms using simpler stimuli. An increased listening effort for speech over pseudo-speech is indicated in the vibrotactile secondary task. These findings validate the HTR dual-task in VR within the broader scope of listening effort literature and highlight the relevance of linguistic competition in speech background noises for memory and listening effort. To further enhance the realism of the experimental environment, room acoustics, embodied distractor talkers, and more diverse sound environments should be included.

Acknowledgements

The authors would like to thank Chinthusa Mohanathasan and Jonathan Ehret for their valuable project collaboration.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): SPP2236 - Project number 444724862: Listening to and remembering conversations between two talkers: cognitive research using embodied conversational agents in audiovisual virtual environments. Sabine Schlittmeier's contribution was supported by a grant from the HEAD-Genuit-Stiftung (Head-Genuit Foundation; P-16/10-W).

Conflict of interest

The authors declare that there is no conflict of interest.

Data availability

The experimental result data are available from the corresponding author upon reasonable request. All HTR material is openly available (Schlittmeier et al. [406] and Ermert et al. [129]). The plugins used to implement the VR experiment are open-source: the RWTH VR Toolkit for basic VR interaction [157], the Character Plugin for animation of the MetaHumans [472], the Study Framework Plugin for controlling the listening experiment and logging the behavioral data [113]), and the Virtual Acoustics Plugin for communication with the Virtual Acoustics server [471]. Virtual Acoustics is an open-source software [209].

For the experiment reported in this article, we confirm that we have reported all measures, all experimental conditions, all data exclusions (for the secondary response times), and the procedures used to determine our sample sizes.

Author contributions

C.E., S.S., and J.F. conceptualized the experiment. C.E. and S.S. developed the methodology. C.E. and A.B. developed the experiment software. C.E. curated the data and analyzed the results. C.E. wrote the manuscript; A.B., S.S., T.K., and J.F. reviewed and edited the manuscript. J.F., S.F., and T.K. provided resources and acquired funding. J.F. supervised the experiment and was responsible for project administration.

Ethical approval

The study was pre-approved by the ethics committee of the medical faculty of RWTH Aachen University (EK396-19) and conducted in accordance with the Declaration of Helsinki.

Informed consent

Participants gave informed written consent prior to participating in the experiment.

6

Listening Effort In Complex Audiovisual Scenes Under Varying Simulated Room Acoustic Conditions

This Chapter was accepted for publication as an original research article in the journal *Trends in Hearing* as part of the collection *ISAAR 2025*:
C. A. Ermert, K. Loh, et al. "Listening Effort In Complex Audiovisual Scenes Under Varying Simulated Room Acoustic Conditions". In: *Trends in Hearing* 30 (2026) [124].

Copyright ©2026 by The Author(s). Reprinted by Permission of Sage Publications. (see Appendix A)

6.1 Abstract

Reverberation can degrade speech intelligibility and comprehension, increase cognitive load, and impair memory. Many studies examining reverberation effects use audio-only conditions and stationary background noise. Thus, it remains unclear to what extent findings generalise beyond laboratory setups. In the present study, we examine the influence of reverberation on memory and behavioural listening effort in a complex, audiovisual virtual reality (VR) environment. We implemented a VR open-plan office with realistic room acoustics and time-varying background noise, including animated virtual agents. Three simulated room-acoustic conditions were compared: a free-field ($T_{30} = 0\,\mathrm{s}$), an acoustically optimised office ($T_{30} = 0.5\,\mathrm{s}$), and an untreated office ($T_{30} = 1.0\,\mathrm{s}$). Under these conditions, $N = 30$ participants performed a dual-task. The primary Heard-Text Recall (HTR) task measured speech comprehension and memory. Participants listened to short stories narrated by two speakers and answered content-related questions. The vibrotactile secondary task measured listening effort. In single-tasking, memory for spoken content was significantly better in free-field than in both offices, with no significant difference between offices. In dual-tasking, vibrotactile

response times indicated that listening effort increased in the presence of reverberation. Across single- and dual-task conditions, vibrotactile error rates decreased in the presence of reverberation, possibly due to elevated concentration in more difficult conditions. These findings suggest that room acoustics, compared to free-field environments, influence memory and listening effort in complex audiovisual VR and highlight the importance of incorporating reverberations when designing audiovisual environments for listening research.

6.2 Introduction

The acoustic characteristics of our environment can fundamentally influence our listening experience. Everyday listening often takes place in complex environments, where room acoustics ([68, 13, 14, 42, 5, 363, 197]), but also time-varying background noise ([484, 62]), and the presence of visual cues ([436, 114, 116]) can affect the cognitive processing of sounds. However, most experimental studies on the effects of reverberation have been conducted in simplified, audio-only laboratory setups, so it remains unclear to what extent their findings can be translated to everyday situations. This study aims to investigate the effects of reverberation on speech processing, taking into account the aforementioned aspects of complex acoustic environments.

Room acoustics, more specifically the reverberation of an environment, are commonly quantified using the RT. T_{60} is the time it takes for sound energy to decay by 60dB. In practice, this decay is often not fully observable above noise, which is why the RT is often estimated from a shorter decay range. In this manuscript, we report RT using T_{30}, i.e., the decay slope over 30dB, which is a common metric in literature to quantify room acoustic conditions (e.g., [254, 363, 197, 5]).

The influence of RT variations on cognitive processing has been examined in multiple audio-only studies. For example, it has been shown that longer RTs impair recall of spoken material, leading to lower memory scores and higher error rates compared to short RTs ([276, 215, 500]). The strongest effects occur in the primacy portion of the to-be-recalled word lists, suggesting that reverberation disrupts early encoding processes ([276]). Participants also make significantly more intrusion errors, i.e., recalling non-list items, in long RT conditions compared to short RT conditions ([276]). Simulated reverberation has further been shown to reduce memory and comprehension of spoken lectures ([237]). Furthermore, adults have been demonstrated to exhibit a general decline in working memory performance during speech exposure under both short and long RTs, reflecting a higher perceived workload ([182]). In summary, these findings indicate that

increased RTs increase cognitive load, thereby limiting resources available for memory encoding and recall ([363]).

Not only working memory, but also listening effort as an indicator of sound processing has been examined under varying RTs in audio-only studies. Listening effort is defined as the mental resources required to attend to and understand speech ([150]). It has gained popularity as a measure of auditory performance because it can be assessed in adverse, but suprathreshold conditions in which speech remains largely intelligible. Compared to, for example, speech reception threshold tests, which are commonly conducted at very low SNRs uncommon for real-life scenarios ([497]), such suprathreshold paradigms allow test conditions to be closer to everyday listening situations. Employing such conditions can potentially increase the ecological validity of auditory assessments ([232]).

According to the ELU framework ([391]), increased RTs are predicted to elevate listening effort by degrading the speech signal, which requires additional cognitive resources for successful decoding ([363]). Psychophysical and subjective measures of listening effort support this assumption. Under severe reverberation ($T_{30} \approx 2\,\mathrm{s}$) combined with background noise, delayed semantic processing and greater mental effort could be shown in electrophysiological measurements ([254]). Subjective ratings similarly indicate increased perceived effort with longer RTs ([398, 276]). Distortions of the temporal speech envelope further necessitate stronger cortical tracking, reflecting increased attentional engagement and effort ([131]). In summary, listening effort is sensitive to RT variations in physiological and subjective measures.

In addition to physiological and subjective measures, listening effort is frequently quantified behaviourally via dual-task paradigms based on the assumption of limited cognitive capacity ([65, 230]). In such paradigms, two tasks are administered together and in isolation. The primary task focuses on listening, while performance or response times of the secondary task serve as indicators of listening effort by computing performance differences between the single and the dual-task conditions of the secondary task ([150]).

In contrast to the aforementioned physiological and subjective findings, behavioural measures show mixed effects with regard to RT variations. In a study by Picou et al. [363], RT was varied from low ($T_{30} < 0.1\,\mathrm{s}$), to moderate ($T_{30} \approx 0.83\,\mathrm{s}$) conditions. Response times on a secondary task did not differ significantly across RT levels, even when SNR was fixed and word recognition performance declined significantly. Similarly, increasing RT did not alter secondary-task response times. Seitz et al. [415] examined behavioural listening effort under comparably short RTs: in free-field ($T_{30} = 0\,\mathrm{s}$) and moderate reverberations ($T_{30} = 0.63\,\mathrm{s}$) no effect of RT on dual-task performance of adults was found. The absence of RT effects on dual-task performance also persists for longer RTs: comparing subjective and

behavioural listening effort under RTs ranging from 0.4 s to 1.2 s, Peng et al. [358] found a sensitivity of subjective listening effort towards RT, but no differences in behavioural listening effort as examined with a speech comprehension dual-task. In contrast, Huang et al. [197] reported that high (T_{30} = 1.22 s) versus moderate (T_{30} = 0.47 s) RTs did affect both behavioural and subjective listening effort. These mixed results suggest that the influence of RT on behavioural listening effort depends on several interacting factors.

One potential factor could be that RT differences were too small. Picou et al. [363] hypothesised that the relatively narrow RT range used in their study (below 0.85 s) may have been insufficient to elicit measurable effort differences. Another reason could be small source-receiver distances. If the source-receiver distance lies within the critical distance of the room ([395]), the direct sound components dominate over the reflected sound, effectively reducing room acoustic effects ([197]). Picou et al. [363] note that this may have diminished reverberation effects in their study, whereas Huang et al. [197] deliberately placed listeners outside the critical distance and observed clear RT effects in behavioural listening effort.

A further limitation of many previous studies is the use of simplified acoustic scenes without spatial sounds, using stationary background sounds, and/or short isolated target words. Spatial sound reproduction is important, as reverberations can affect the localization of target and background sources ([13]). To assess this, however, the scene must be complex enough to provide localization cues for each source individually (e.g., via room acoustic simulations) instead of using artificial reverb. The nature of the background sounds is important as reverberations can affect speech intelligibility, which, in the case of background speech, can result in a release from informational masking ([74, 233]), but may also affect target speech intelligibility ([138]). To assess those effects, it is necessary to employ realistic, time-varying multi-talker speech instead of synthetic or static background noise, nonsense speech, or multi-talker babble where intelligibility is not given. This would in turn also enable the listeners to exploit temporal "dips" or gaps in fluctuating maskers where the target is more intelligible ([393, 233]), an ability also reduced by reverberations ([90]). Lastly, target utterances should be long enough to include multiple reflections as occurs with extended speech material, in contrast to brief isolated target words often employed in simplified study paradigms (e.g., [415]).

An additional important limitation is that most of the work above was conducted in audio-only laboratory environments. Yet, listening in everyday life is often multimodal ([436]). For example, co-verbal cues like lip movements have been shown to significantly improve speech intelligibility compared to audio-only conditions ([445, 184, 418]). This can also be observed for virtual agents ([104, 337,

95]), albeit not to the same degree as with real human faces ([104]). In contrast, co-verbal cues can also increase cognitive load through multimodal processing ([71]). Neglecting these aspects of realism potentially creates a gap between the real life and the study scenario. This gap may limit the applicability of findings to naturalistic listening scenarios.

With the development of VR technology, conducting studies in controlled yet complex audiovisual environments has become possible. Such scenarios can provide co-verbal cues, e.g., by simulating scenarios with fully animated virtual agents with synchronised lip-, head- and body-movement. Consequently, VR has become a popular tool in listening experiments to increase the complexity and fidelity of experimental conditions (e.g., [328, 105, 60]). It has been used to quantify cognitive performance ([189, 71]), listening effort ([399, 115, 125]), and also room-acoustic effects on performance and perception. For example, Doggett et al. [109] examined cognitive performance in a virtual open-plan office and varied the acoustic treatment of the walls ($T_{30} = 0.9\,\mathrm{s}$ vs. $T_{30} = 0.5\,\mathrm{s}$) and background noise. They could demonstrate that the performance decrement caused by the noise could be ameliorated by the acoustic treatment. Comparing three room acoustic conditions ($T_{30} = 0\,\mathrm{s}$, $T_{30} \approx 0.5\,\mathrm{s}$, $T_{30} \approx 1.6\,\mathrm{s}$), Ahrens et al. [5] found that participants located a target talker in multi-talker VR environments both faster and more accurately in short compared to long RTs. The effect of sound isolation between adjacent rooms on working memory has been examined by Muhammad et al. [328] and the results were comparable to those of laboratory settings. Incorporating RT in VR has been further shown to improve immersion (e.g., [467]) and increase perceived room size (e.g., [343]). VR has also been used to demonstrate RT effects on listening effort: [133] found higher subjective listening effort in a reverberant virtual room ($T_{60} = 1.44\,\mathrm{s}$) compared to free-field using a questionnaire, although background speech was nonsensical and only embodied as virtual loudspeakers instead of animated talkers.

To the best of our knowledge, no study has examined behavioural listening effort across different RTs in a visually rich VR environment with embodied background agents and realistic, time-varying multi-talker noise. However, as elaborated above, these aspects of complex real-life scenarios might influence speech processing. Traditional studies may not provide a complete picture of reverberation effects. Consequently, the effect of room-acoustic reverberation on behavioural listening effort in multisensory VR is still not comprehensively understood. This study aims to address this research gap.

6.2.1 Present Study

The present study investigates whether room-acoustic variations within a virtual office affect memory and behavioural listening effort in a complex audiovisual VR environment. To this end, we implemented a complex open-plan office in VR with animated target talkers and background agents, head-tracked binaural rendering, and time-varying quasi-random background noise. We examined three room-acoustic conditions: a free-field anechoic baseline ($T_{30} = 0\,\mathrm{s}$), an untreated open-plan office ($T_{30} = 1.0\,\mathrm{s}$), and the same office with acoustic treatment ($T_{30} = 0.5\,\mathrm{s}$) designed to meet relevant standards for open-plan offices ([210, 12]). The two reverberant conditions were derived from the same room geometry: the untreated office used material-dependent absorption and scattering coefficients and yielded $T_{30} \approx 1.0\,\mathrm{s}$, while the treated office included acoustic treatments like a carpet, an acoustic ceiling, and hanging absorbers to reduce T_{30} in the mid-frequency range. These treatments were not visualised.

Previous work suggests that small RT differences may not always yield measurable behavioural effort effects, especially when listeners are within the critical distance ([363]). In our setup, source–receiver distances were deliberately chosen to be outside the critical distance so that reverberant energy dominates the target signal. Outside of the critical distance, Huang et al. [197] reported robust behavioural listening-effort effects for RT differences similar to the ones employed here. We therefore expected that even moderate RTs typical of office environments could influence listening effort in our complex audiovisual VR setting, without resorting to unrealistic RT conditions; our goal was to create realistically "good" and "bad" open-plan offices rather than arbitrarily extreme RTs.

Listening effort was measured using a dual-task paradigm in VR. As the primary task, we employed the HTR paradigm by Schlittmeier et al. [406], in which participants listen to short family stories and subsequently answer content-related questions, providing measures of text comprehension and verbal memory. Employing virtual agents, these stories were presented audiovisually in VR as a conversation between a female and a male talker (e.g., [115, 399, 125]). As the secondary task, a vibrotactile pattern categorization task was used (cf. [320, 399, 125, 114]). Both tasks were performed in single-task conditions (each task in isolation, serving as a baseline) and in dual-task conditions (both tasks simultaneously) in each acoustic environment. Listening effort is hereby conceptualised as the cognitive resources specifically required to attend to and process speech, i.e., as an auditory task-related subcomponent of overall cognitive load ([150, 303]). The absolute level of cognitive load in a complex, multisensory VR environment is certainly influenced by visual complexity and task demands. However, the audio-visual scenario and secondary task are held constant across all acoustic

conditions in the present study. Such influences are therefore assumed to be constant across conditions. Listening effort is consequently understood as the change in secondary-task performance (both in terms of proportion of correct responses and response times) between single- and dual-tasking under different room-acoustic conditions, rather than as an absolute performance level.

In our study, we addressed the following research question: do room-acoustic differences between a free-field condition and a simulated treated and untreated open-plan office affect memory and behavioural listening effort in a complex audiovisual VR scenario? Based on the literature reviewed above, we formulated the following hypotheses:

- H1 Participant performance in both the primary listening task and secondary vibrotactile task will be better under single-tasking and worse under dual-tasking (i.e., lower memory performance, lower categorisation accuracy, and longer response times) as evident in previous studies using the specified dual-task paradigm (e.g., [115, 399, 125, 320]).
- H2 In the primary memory task, evaluating text comprehension and verbal memory, the participants' performance will decrease with increasing RT, in line with literature on the effect of RT variations on working memory performance ([276, 215, 500]). Ideally, this effect only surfaces in single-tasking and performance remains constant in dual-tasking, given that the primary task is prioritised correctly by participants.
- H3 In the vibrotactile secondary task, the influence of the acoustic condition on single-tasking will be negligible, because the vibrotactile task does not rely on the processing of auditory stimuli. In dual-tasking, however, we expect reverberant conditions to induce higher listening effort than in the free-field, in line with ELU predictions ([391, 363]) and prior evidence for reverberation-related increases in behavioural effort ([197]).

6.3 Methods

6.3.1 Participants

A power analysis was carried out in the software R (v4.4.1, [374]) with the simR package (v1.0.7, [166]). For the room conditions, we assumed small (between the two RT conditions) to moderate (between the freefield and the two RT conditions) effects, and high dual-task costs for the number of concurrent tasks. On the basis of a generalised linear mixed model (see Section Data Analysis), approximately $N = 28$ participants were required for a power of 80% in the interaction effect. A total of $N = 30$ adults (15 female, 15 male; age: $18 - 33$ years, $M = 25.1$, $SD = 3.1$) were recruited for the experiment, providing a buffer for potential data



(a) Target talkers.



(b) Virtual computer screen.

Figure 6.1: Experiment environment. 6.1a: target talkers narrating HTR texts in virtual open-plan office. The virtual agents were animated with lip movement, gestures, and head movement. Background agents at the desks and in the kitchen are visible. 6.1b: virtual computer screen which displays the HTR questions after the presentation of the text. Participants respond to questions vocally, and an instructor marks the answer as correct or incorrect.

loss and ensuring that the sample size was a multiple of the number of conditions, facilitating condition balancing. All participants were native German speakers. Hearing thresholds were assessed using pulsed pure-tone audiometry (AURITEC Ear 3.0 audiometer, Sennheiser HDA 300 headphones), ensuring normal hearing sensitivity (< 20 dBHL between 125 Hz and 8 kHz) in both ears. No participants had to be excluded. Participants additionally reported normal or corrected-to-normal vision. Written informed consent was obtained before testing. Participants received 25 € as compensation for successful participation.

6.3.2 Dual-Task Paradigm

Cognitive performance was assessed using a dual-task paradigm consisting of the HTR task ([406]) as the primary task and a vibrotactile secondary task ([320, 144, 162]). The HTR task required comprehension and recall of running speech, while the secondary task served as an indicator of listening effort.

During the HTR task, one female and one male speaker, embodied as virtual agents, narrated family stories in German (see Figure 6.1a). They took turns according to the turn-taking pattern suggested in the AUdioVIsual Speech and Text (AuViST) database ([129]) for the HTR. One agent spoke successive sentences that form a unit of meaning and then yielded the turn. Agents never interrupted each other. The number of sentences spoken by each agent was pseudo-random, and the speaking time per text was approximately the same. Texts differed in duration, but were approximately $M = 43.4$ s ($SD = 1.9$ s) in length. After each story, participants answered nine questions about specific details such as

names, family relations, and other content-related information. The questions were presented visually on a virtual computer screen in the open-plan office environment (see Figure 6.1b). Participants answered the questions vocally, and an instructor marked the answer as correct or incorrect based on the solutions provided in the HTR database by Schlittmeier et al. [406]. Audio recordings (male and female voices) with synchronised lip movements and face-tracking data were taken from the AuViST database ([129]) and augmented with gestures, gaze behaviour, and breathing as described by Ehret et al. [114] (publication includes a supplementary video; condition "$\mathrm{T_{full}}$" at $t = 3 : 23\,\mathrm{min}$ was used for the present publication). In the vibrotactile secondary task, two handheld controllers vibrated in pairs of impulses, each either short (0.2 s) or long (0.6 s) ([320, 399, 115]). If both impulses were of equal length, participants should press the right trigger; if they differed, the left trigger. Responses were categorised as correct, incorrect, or missed. Response times were recorded as well. Both the secondary task performance and secondary task response times are interpreted as indicators of behavioural listening effort ([320, 399, 125]).

Both tasks were performed separately (single-task as baseline) and concurrently (dual-task). In the single-task condition, each trial contained 25 vibration patterns. In the dual-task condition, the number of vibration patterns per trial varied according to the length of the story and the proportion of short and long impulses, resulting in $12 - 15$ patterns per trial ($M = 13.5$, $SD = 0.81$).

6.3.3 Audiovisual Environment

Visual Model

An open-plan office ($l \times w \times h = 16.5\,\mathrm{m} \times 8\,\mathrm{m} \times 3.5\,\mathrm{m} = 462\,\mathrm{m}^3$) was created in Unreal Engine 5.3 (Epic Games) using an open-plan office asset package for Unreal Engine ([1]). The participant was seated at a desk facing two virtual target speakers narrating the HTR texts (see Figure 6.1a).

Four additional animated virtual background agents were positioned at nearby desks and one in the adjacent kitchen area, all facing away from the participant to eliminate the need for gaze and lip-synchronization management (see Figure 6.2).

All virtual agents, including the target talkers, were generated with the MetaHuman Creator (Epic Games). The following plugins were used to implement the experiment: the Character Plugin ([472]) for agent animations, the StudyFramework Plugin ([113]) for controlling the experiment flow, and the RWTH VR Toolkit Plugin ([157]) for participant interaction.

The same visual model was used for all RT conditions. To avoid bias toward specific acoustic expectations, no visual cues of acoustic treatment (e.g., absorbers at the ceiling) were included. Instead, subtle variations in wall and ceiling color

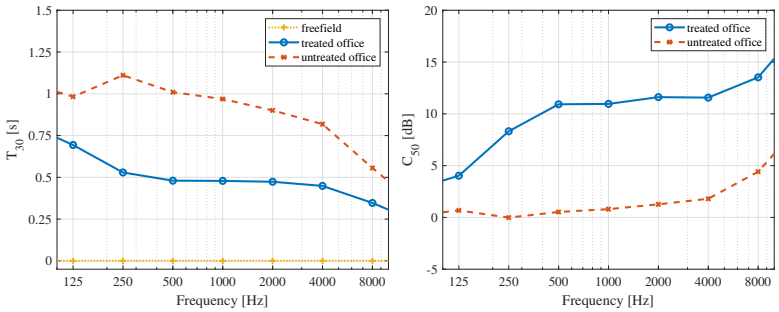


Figure 6.2: Top-down model of the open-plan office. The kitchen area (left) is separated from the office area (right) with a shelf. The position of the participant is depicted in blue, the two target talkers in green, and the five background agents in red.

distinguished the acoustic conditions. The colors were matched in warmth and saturation to avoid visual contrast while maintaining perceptual differentiation between conditions: RGB(153, 115, 115) (see Figure 6.2), RGB(153, 133, 115) (see Figure 6.1b), and RGB(145, 115, 153) (see Figure 6.1a). The assignment of acoustic conditions and colors was balanced between participants. The visual model was presented with an HTC VIVE Pro 2 HMD.

Acoustic Models

Two room acoustic models with the same dimensions as the visual model were created in SketchUp Make 2016 (Trimble Inc) using the RAVEN plugin ([410]), corresponding to an acoustically treated and an untreated version of the visual model. A third free-field condition simulated direct sound only. Material absorption and scattering coefficients were derived from the Benchmark for Room Acoustical Simulation (BRAS) database ([15]) and [107]. Because diffraction is not supported in RAVEN, human bodies were modelled as equivalent absorption areas at their seated locations. The untreated model replicated the office geometry without further modifications. The treated model included additional absorption through a carpeted floor, acoustic ceiling, and hanging absorbers. The amount of absorptive material was adjusted until T_{30} was below 0.6 s between $250 - 2000$ Hz, in line with workplace standards for room acoustics ([210, 12]). RTs were computed from simulated room impulse responses and averaged across all source–receiver pairs. Mid-frequency RTs were $T_{30} = 0.0$ s for the anechoic free-field condition, $T_{30} = 0.5$ s for the treated office, and $T_{30} = 1.0$ s for the untreated office. The room acoustic specifications are displayed in Figure 6.3



(a) Reverberation time (RT) T_{30} **(b)** Clarity C_{50}

Figure 6.3: Reverberation time (RT) (T_{30}) and clarity (C_{50}) for the three acoustic conditions $T_{30} = 0\,\text{s}$ (free-field), $T_{30} = 0.5\,\text{s}$ (treated office), and $T_{30} = 1.0\,\text{s}$ (untreated office). C_{50} for the free-field is not displayed as it theoretically equals + inf.

with regard to the RT T_{30} and the clarity C_{50} ([51]). It was ensured that the distance between the participant and the target talkers exceeded the critical distance ([197]).

Speech transmission index (STI) estimates were calculated based on the room impulse responses according to Larm and Hongisto [256] and ranged from $0.78 - 0.85$ (background sources) and 0.92 (targets) in the treated office ($T_{30} = 0.5\,\text{s}$) and from $0.61 - 0.67$ (background sources) and 0.74 (targets) in the untreated office ($T_{30} = 1.0\,\text{s}$). In the free-field condition ($T_{30} = 0\,\text{s}$), STI theoretically equalled 1.0. These values provide relative indicators of intelligibility differences between conditions, as variations in SNR were not explicitly modelled (see Section Acoustic Calibration). The acoustic scene was presented with Sennheiser HD650 headphones.

Audiovisual Stimuli Design and Rendering

Complex audiovisual scenes were created to provide a realistic office background activity. Four sound categories were used: human sounds (e.g., speech, non-speech), appliances (e.g., kitchen, computer, phone), traffic (cars), and static background (e.g., fan noise). Most sounds were newly recorded to ensure comparable acoustic characteristics (i.e., by using the same measurement equipment and setup) and to provide agent-specific voices. Recordings were made as mono signals in a sound-isolated booth at 44.1 kHz/16-bit with a Zoom H2n, capturing two keyboards, two computer mice, general office task noises (e.g., writing), plus two female and two male German speakers producing telephone conversations, short utterances, and non-speech vocalizations. Additional appliance noises (e.g.,

fridge, computers) were recorded in quiet environments outside of the booth. Recording were done at a distance of 0.8 m to reduce room acoustic influences while minimizing near-field effects, such as strong sibilants and other articulation noises. The resulting mono signals were post-processed to reduce noise effects and influences of the measurement setup. Even though a sound-isolated booth was chosen for the recordings and quiet conditions were ensured for the recordings outside of the booth, a low noise floor was still present in the recordings. Since this noise floor would add up when playing multiple stimuli simultaneously in the sound sequences, noise reduction was applied using Audacity (v3.7.3, Muse Group & contributors). Further, room resonances present in the recordings were attenuated with bell filters in Ableton Live (v11.3.42). Edited files were convolved with the source-position specific room-impulse responses for the treated and untreated office, which were obtained through RAVEN simulations from the two room acoustic models as specified in the previous Section (Acoustic Models).

The scenes were reproduced using VA ([209]) and the VA plugin for Unreal Engine 5.3 ([471]) with a HRTF of the KEMAR (GRAS Sound & Vibration, Denmark) artificial head ([53]). Head movements were enabled via Unreal Engine and VA and the built-in tracking system of the HMD. Visual animations within the scene matched background events (e.g., phone reach/hang-up, nodding, coughing/yawning); traffic lacked visuals (which was deemed acceptable, as participants did not sit close to virtual windows), while static backgrounds used object animations. To obtain a time-varying quasi-random noise, but to counteract high variances in the noise and to achieve equal difficulty between participants, the background noise was pre-rendered. Using the Complex Noise Scene Generator (CNSG) by Breuer et al. [62], sequences of the background noise stimuli, "background sequences", were pre-generated in which the noise categories appeared in a balanced manner. No more than two background agents spoke at the same time. The coordination between auditory and visual stimuli was realised using a custom simulation scheduler class in Unreal Engine ([115]). All background sequences were presented to all participants in counterbalanced order.

Acoustic Calibration

Calibration of time-variant noise in the presence of reverberation is not a straightforward task. Generally speaking, reverberations increase sound pressure level through the added contribution of the reflected sound. Moreover, this effect is frequency dependent and might affect different sound sources differently. Thus, controlling the SNR would eliminate an important contribution of the room acoustics. Therefore, the following approach was chosen: target speech during

HTR was calibrated to 60 dB(A) at the listener in the treated office ($T_{30} = 0.5$ s) by adjusting source sound power in VA. For comparability, the same sound power was then used in the untreated office ($T_{30} = 1.0$ s), as a source in real-life would also produce the approximately same sound power in both conditions. Due to the increased contribution of reflections compared to the treated office, this yielded 64.5 dB(A) at the listener. This difference was intentionally retained to comprehensively capture the effects of room acoustics. For the background agents, the same procedure was applied.

However, using the same sound power in the free-field would have led to significantly reduced sound pressure levels due to the absence of reflections. To enable comparability with the other conditions, the sound power of each source in the free-field model was adjusted to match the corresponding level in the treated office. Since the free-field condition is an artificial condition itself, the level of sources is oftentimes adjusted freely, and this procedure was deemed acceptable. Mean background sequence levels were 56.9 dB(A) for free-field ($T_{30} = 0$ s), 57.3 dB(A) for the treated office ($T_{30} = 0.5$ s), and 63.5 dB(A) for the untreated office ($T_{30} = 1.0$ s). The resulting time averaged SNRs were thus approximately 3 dB(A) in the free-field and treated office, and approximately 1.5 dB(A) in the untreated office.

Levels were derived from recordings with a calibrated artificial head microphone as A-weighted binaural power sums using the ITA Toolbox ([41]) in MATLAB (v2023b, Mathworks).

6.3.4 Procedure

The experiment was conducted in a hearing booth as it offers stable acoustic conditions through low background noise and sound isolation. After reading the study description and providing informed consent, participants were introduced to the HMD and controllers. They were instructed to prioritise the primary task while performing both as accurately as possible. The session began with three practice trials: the secondary vibrotactile task alone, repeated until participants felt comfortable with the task, one trial of the primary HTR task alone, and one trial of both tasks together in the dual-task condition in the free-field condition without background noise. Subsequently, the main experiment started. The main experiment consisted of three blocks, one per acoustic condition (free-field: $T_{30} = 0$ s, treated office: $T_{30} = 0.5$ s, untreated office: $T_{30} = 1.0$ s). At the start of each block, participants entered a new virtual "room" (defined by a change of wall color) and familiarised themselves with the corresponding acoustic condition by listening to excerpts of background noise. In each block, participants completed four dual-task trials, one HTR single-task trial, and 25 vibration pat-

terns in the secondary task single-task condition. To counteract potential fatigue effects, participants were encouraged to take self-paced breaks between blocks during which they could take off the HMD. The order of acoustic conditions was counterbalanced using a balanced Latin square; the order of tasks within each condition was also balanced. In total, participants performed 17 primary task trials, including practice. Seventeen of 34 available HTR texts were used, selected for medium and consistent difficulty ([323]). After each story, questions were displayed visually on a virtual screen once the audio and noise had stopped. Answers were given verbally without time limits. Using an "all-or-nothing"-approach, partially correct responses were marked incorrect. For the 21 trials including background noise, 21 background sequences were used. Balanced Latin squares ensured quasi-randomised presentation of texts, background sounds, and wall color assignments across conditions. After the experimental blocks, participants completed a post-experiment questionnaire regarding the perceived ecological validity. The questionnaire did not serve the primary research hypothesis of this paper and is thus not further discussed.

Including breaks, the total duration of the experiment was approximately 90 min, of which approximately 60 min were spent immersed in the VR environment.

6.3.5 Data Analysis

The analysis was carried out in R (v4.4.1, [374]). Generalized linear mixed models (GLMMs) were fitted using the *glmmTMB* function of the *glmmTMB* package (v1.1.10, [69]) for the primary HTR task performance (binary; 1 = question correct, 0 = question incorrect), as a measure of text comprehension and memory, as well as the secondary task performance (binary; 1 = pattern categorised correctly, 0 = wrong categorization) and the response time in the secondary vibrotactile task (in milliseconds), as a measure of listening effort.

A significance level of $\alpha = .05$ was applied for each analysis, and training trials were removed. Individual models were fitted for all three outcome variables. The initial models comprised the fixed-effect factorial predictors *number of tasks* (single-task, dual-task) and *acoustic condition* (free-field: $T_{30} = 0\,\mathrm{s}$, treated office: $T_{30} = 0.5\,\mathrm{s}$, untreated office: $T_{30} = 1.0\,\mathrm{s}$), and random intercepts for the *participant* (to account for variations between participants) and the *background sequence* (to account for differences in the background noise scenarios). Since varying text difficulty within the HTR has been demonstrated ([323]) and further differences in difficulty between the HTR questions can be expected, the primary task analyses further comprised *HTR question*, a unique identifier for each question within the HTR, as a random intercept. Balancing the maximum

model approach ([28, 300, 70]) and experiment design justification, random by-participant slopes for *acoustic condition*, *number of tasks*, and their interaction were initially included in the models, but led to non-convergence and singularities. The step-by-step simplification process ([28]) led to a final model which retained fixed-effect predictors and interactions as well as random intercepts for all three outcome measures, but by-participant random slopes only for the *number of tasks*, reflecting inter-participant differences in dual-task cost.

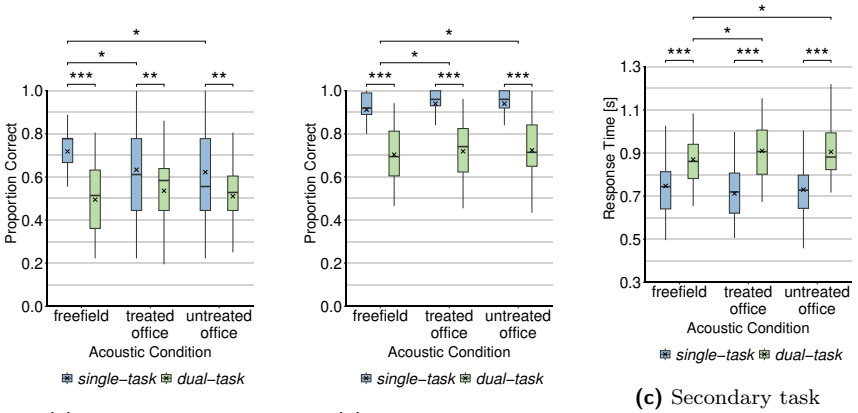
In the final models, the package *DHARMa* (v0.4.6, [178]) was used to check for under- or overdispersion and other violations of distributional assumptions based on simulation-based residuals. Subsequently, the contribution of the fixed effects *number of tasks* and *acoustic condition* and their interactions were evaluated via likelihood-ratio tests using the *mixed* function of the *afex* package (v1.5-0, [428]). For categorical predictors with more than two levels and for significant interactions, these omnibus tests only indicate whether any effect is present, but do not identify which specific levels differ. Significant main effects or interactions detected in these tests were therefore further explored with post-hoc pairwise comparisons of estimated marginal means, conducted using the *emmeans* package (v1.10.4, [267]), which allows contrasts between all pairs of levels on the response scale, rather than only against a single reference level. To control for multiple comparisons, p-values were adjusted via the Tukey method. As a goodness-of-fit measure, marginal and conditional R^2 were computed with the *r2_nakagawa* function of the *performance* package (v0.15.2, [279]). They describe the proportion of variability in the model outcome that can be explained by the fixed factors (marginal R^2) and the fixed and random factors (conditional R^2) in the model [332]. Details of the individual models are provided in the following Section.

6.4 Results

The primary task performance, secondary task performance, and secondary task response times are displayed in Figure 6.4.

6.4.1 Primary Task Performance

The outcome variable in the primary task was the task performance, specified as a binary variable, where 1 represents correct responses and 0 represents incorrect responses. A GLMM with a binomial distribution and a logit link function was fitted. Please refer to the previous Section for details on the software and packages used and the specific tests performed. A model summary is given in Table 6.1. The marginal and conditional R^2 were 0.026 and 0.309, respectively.



(a) Primary task **(b)** Secondary task **(c)** Secondary task response times

Figure 6.4: Performance in the (a) primary task, (b) secondary task, and (c) response times in the secondary task. Proportion correct for the primary task and secondary task as an average of correct and wrong responses per condition and participant, and response times in seconds for the secondary task are plotted as a function of the *acoustic condition* (free-field: $T_{30} = 0\,\text{s}$, treated office: $T_{30} = 0.5\,\text{s}$, untreated office: $T_{30} = 1.0\,\text{s}$) and the *number of tasks* (single-task, dual-task). The boxes represent the interquartile ranges, while the median is indicated with a horizontal line and the mean with a cross (×). ***$p < .001$, **$p < .01$, *$p < .05$.

Table 6.1: Generalised linear mixed model for primary task performance. Comparisons between the levels of the fixed-effects *acoustic condition* (free-field: $T_{30} = 0\,\text{s}$, treated office: $T_{30} = 0.5\,\text{s}$, untreated office: $T_{30} = 1.0\,\text{s}$) and *number of tasks* (single-task, dual-task). Significant p-values below .05 are highlighted in bold font.

	Estimate	SE	z	p
(Intercept: single-task – free-field)	1.230	0.217	5.669	**< .001**
single-task – treated office	−0.530	0.215	−2.461	**.014**
single-task – untreated office	−0.586	0.211	−2.779	**.005**
dual-task – free-field	−1.259	0.197	−6.402	**< .001**
dual-task – treated office	−0.731	0.238	3.076	**.002**
dual-task – untreated office	−0.647	0.235	2.757	**.006**

Number of observations: 4023; groups: participant = 30, background sequence = 21, HTR question = 153. Model equation: performance ∼ acoustic condition * number of tasks + (1 + number of tasks | participant) + (1 | background sequence) + (1 | HTR question); family = binomial, link function = logit.

Likelihood-ratio tests on the fixed effects indicated a significant main effect of *number of tasks* ($\chi^2(1) = 24.22$, $p < .001$), and a significant interaction between *acoustic condition* and *number of tasks* ($\chi^2(2) = 11.46$, $p = .003$). The main effect of *acoustic condition* did not reach significance ($\chi^2(2) = 5.16$, $p = .076$).

Post-hoc tests indicated that in single-tasking, the free-field condition yielded significantly higher performance than the treated office ($OR = 1.70$, $z = 2.46$, $p = .037$) and the untreated office ($OR = 1.80$, $z = 2.78$, $p = .015$). Performance in the treated and the untreated office did not differ significantly ($OR = 1.06$, $z = 0.28$, $p = .959$). No differences between *acoustic condition* could be found in dual-tasking (free-field vs. treated office: $OR = 0.82$, $z = -2.00$, $p = .11$; free-field vs. untreated office: $OR = 0.94$, $z = -0.60$, $p = .82$; treated office vs. untreated office: $OR = 1.15$, $z = 1.38$, $p = .350$). Single-tasking performance was significantly better than dual-tasking in free-field ($OR = 3.52$, $z = 6.40$, $p < .001$), in the treated office ($OR = 1.70$, $z = 2.81$, $p = .005$), and the untreated office ($OR = 1.84$, $z = 3.29$, $p = .001$).

6.4.2 Secondary Task Performance

The outcome variable in the secondary task was the task performance, specified as a binary variable, where 1 represents correct vibrotactile pattern categorization, and 0 represents incorrect categorization. A GLMM (see Section Data Analysis) with a binomial distribution and a logit link function was fitted. Please refer to Section 6.3.5 for details on the software and packages used and the specific tests performed. A model summary is given in Table 6.2. The marginal and conditional R^2 were 0.172 and 0.305, respectively.

Likelihood-ratio tests on the fixed effects indicated a significant main effect of *number of tasks* ($\chi^2(1) = 47.98$, $p < .001$), and a significant main effect of the *acoustic condition* ($\chi^2(2) = 8.45$, $p = .015$). The interaction between both effects was not significant ($\chi^2(2) = 3.52$, $p = .172$), but was retained in the model as it encompasses the experimental design better.

Post-hoc tests indicated that the free-field condition yielded significantly worse performance than the treated office ($OR = 0.76$, $z = -2.44$, $p = .039$) and the untreated office ($OR = 0.76$, $z = -2.511$, $p = .032$) as a main effect. Performance in the treated office and the untreated office did not differ significantly ($OR = 0.99$, $z = -0.06$, $p = .998$). Single-tasking performance was significantly better than dual-tasking ($OR = 6.80$, $z = 10.20$, $p < .001$).

6.4.3 Secondary Task Response Times

As a third outcome measure, the response times of the secondary task were evaluated. Herein, the time from vibration pattern offset to button press was

Table 6.2: Generalised linear mixed model for secondary task performance. Comparisons between the levels of the fixed-effects *acoustic condition* (free-field: $T_{30} = 0\,\text{s}$, treated office: $T_{30} = 0.5\,\text{s}$, untreated office: $T_{30} = 1.0\,\text{s}$) and *number of tasks* (single-task, dual-task). Significant p-values below .05 are highlighted in bold font.

	Estimate	SE	z	p
(Intercept: single-task – free-field)	2.649	0.218	12.146	$<$ **.001**
single-task – treated office	−1.675	0.218	−7.689	$<$ **.001**
single-task – untreated office	0.461	0.212	2.178	**.030**
dual-task – free-field	0.464	0.212	2.187	**.029**
dual-task – treated office	−0.367	0.228	−1.614	.107
dual-task – untreated office	−0.358	0.228	−1.570	.116

Number of observations: 7062; groups: participant = 30, background sequence = 21. Model equation: performance ∼ acoustic condition * number of tasks + (1 + number of tasks | participant) + (1 | background sequence); family = binomial, link function = logit.

measured in milliseconds. Following Mohanathasan et al. [320], response times below 250 ms and above $M+2SD$ were excluded ([489, 38]). A GLMM (see Section Data Analysis) with a Gamma distribution and a log link function was fitted. The latter outperformed Gaussian or lognormal distribution families. Please refer to Section 6.3.5 for details on the software and packages used and the specific tests performed. A model summary is given in Table 6.3. The marginal and conditional R^2 were 0.109 and 0.263, respectively.

Likelihood-ratio tests on the fixed effects indicated a significant main effect of *number of tasks* ($\chi^2(1) = 29.69$, $p < .001$), and a significant main effect of the interaction between *acoustic condition* and *number of tasks* ($\chi^2(2) = 10.31$, $p = .006$), but the main effect of *acoustic condition* was not significant ($\chi^2(2) = 1.91$, $p = .384$).

As evaluated via post-hoc tests, response times did not differ significantly in single-tasking for all *acoustic conditions* (free-field vs. treated office: $OR = 1.04$, $z = 2.27$, $p = .061$; free-field vs. untreated office: $OR = 1.01$, $z = -0.68$, $p = .775$; treated office vs. untreated office: $OR = 0.97$, $z = -1.62$, $p = .238$). In dual-tasking, significant differences were found between free-field and the treated office ($OR = 0.97$, $z = -2.543$, $p = .030$), as well as between free-field and the untreated office ($OR = 0.97$, $z = -2.67$, $p = .021$), but not between the treated office and the untreated office ($OR = 1.00$, $z = -0.11$, $p = .994$). Single-tasking performance was significantly better than dual-tasking in free-field ($OR = 0.81$, $z = -8.06$, $p < .001$), the treated office ($OR = 0.76$, $z = -10.66$, $p < .001$), and the untreated office ($OR = 0.78$, $z = -9.75$, $p < .001$).

Table 6.3: Generalised linear mixed model for secondary task response time. Comparisons between the levels of the fixed-effects *acoustic condition* (free-field: $T_{30} = 0\,\text{s}$, treated office: $T_{30} = 0.5\,\text{s}$, untreated office: $T_{30} = 1.0\,\text{s}$) and *number of tasks* (single-task, dual-task). Significant p-values below .05 are highlighted in bold font.

	Estimate	SE	z	p
(Intercept: single-task – free-field)	6.606	0.032	208.424	$<$ **.001**
single-task – treated office	−0.038	0.017	−2.267	**.023**
single-task – untreated office	−0.011	0.017	−0.680	.496
dual-task – free-field	0.207	0.026	8.063	$<$ **.001**
dual-task – treated office	0.069	0.021	3.307	$<$ **.001**
dual-task – untreated office	0.034	0.021	2.096	**.036**

Number of observations: 6167; groups: participant = 30, background sequence = 21. Model equation: performance ∼ acoustic condition * number of tasks + (1 + number of tasks | participant) + (1 | background sequence); family = Gamma, link function = log.

6.5 Discussion

This study aimed to evaluate the influence of reverberation on behavioural listening effort in a complex audiovisual VR environment. To this extent, three research hypotheses were formulated (see Section Present Study), which will be discussed in the following.

First, it was hypothesised in **H1** that performance in terms of accuracy in the primary text comprehension and memory task and in terms of accuracy and response times in the vibrotactile secondary task would be negatively impacted by dual-tasking compared to single-tasking. This was supported by the data; in all three outcome measures, a significant main effect of the number of tasks was found. This is in alignment with previous studies employing the HTR as the primary task in a dual-task paradigm ([399, 115, 320, 125]) and mirrors a general trend in dual-task paradigms ([415, 363]).

Secondly, in **H2**, we expected performance in the primary memory and text comprehension task to decrease with longer RTs. To some extent, this could be validated in single-tasking: performance was significantly better in free-field than in both simulated rooms, with no difference between the treated office and the untreated office. Situating these findings in literature is challenging: as Mohanathasan et al. [321] demonstrated, comparability of the HTR with other memory tasks (e.g., the operation span task) is limited. The HTR primary task uses short stories of $M = 43.39\,\text{s}$ in duration, and the "memory target" is not necessarily known at onset, which places different demands on attention and

working memory than span tasks in which the "memory target" is specified explicitly. Further, most of these studies use short utterances as target words. Still, the present findings partially align with prior studies on memory and speech comprehension which demonstrate that reverberation impairs memory recall and speech comprehension (e.g., [236, 237, 276]). Furthermore, the absence of a clear gradient between the treated to the untreated office is compatible with work reporting small or no effects of moderate RT differences on word recognition ([363]). Regarding the second part of H2, in dual-tasking, the performance in the primary task should prevail, given that the primary task is prioritised by the participants ([150]). In the present study in dual-tasking, the primary task did not differ across acoustic conditions, suggesting that participants prioritised the HTR equally across conditions and allowed the costs of reverberation to emerge primarily in the secondary task. This prioritization pattern is consistent with resource-limited accounts of cognitive processing ([65, 230]).

Lastly, in **H3**, we hypothesised that the secondary task performance and response times in dual-tasking as an indicator of listening effort would be affected by the acoustic condition. Considering secondary task response times, the data supports no behavioural decrement in listening effort between the two reverberant conditions, even though the source-receiver distance exceeded the critical distance. Still, response times increased in both reverberant rooms relative to free-field, while single-task response times did not vary with the acoustic condition. This pattern indicates that additional processing resources were recruited in reverberant conditions when participants simultaneously prioritised primary-task performance. This aligns with the mixed picture painted in the literature; some studies reported no effect of RT differences on dual-task performance (e.g., [363, 358]), while others found a significant effect (e.g., [197]). Similarly, subjective listening effort ratings have shown both presence and absence of reverberation effects (e.g., [398, 276, 197, 351]. Comparisons of free-field versus reverberant environments also yield inconsistent outcomes. For example, Seitz et al. [415] found no effect on dual-task performance, while Kyong et al. [254] observed physiological brain response alterations attributed to speech degradation in reverberation. It should be noted, however, that Seitz et al. [415] attribute the missing finding to the inappropriate difficulty of the secondary task used and Kyong et al. [254]'s use of very long RTs ($\approx 2\,\mathrm{s}$) might have exaggerated effects outside typical real-world conditions. While our response time data supports a sensitivity of behavioural listening effort towards the presence of reverberation compared to free-field in naturalistic audiovisual listening scenarios, further research is necessary.

With regard to secondary task accuracy, performance increased in the treated and untreated office compared to the free-field across single and dual-tasking. This should, however, not be interpreted as a decrease in listening effort, as

listening effort is mainly reflected in performance differences between single- and dual-tasking. Rather, the more difficult listening scenarios could have caused the participants to concentrate more. In dual-tasking, together with slower response times, the findings could also indicate a speed–accuracy trade-off. Differential findings between the secondary task accuracy and response times, as in the present study, have been reported in several HTR studies. For example, Schiller et al. [399] and Mohanathasan et al. [321] both reported a greater sensitivity of response times than the accuracy to experimental variations. The present data, therefore, reinforce that dual-task response times are a sensitive behavioural index of listening effort.

In summary, a decrease in memory and text comprehension performance and an increase in listening effort were observed in the presence of reverberation compared to free-field conditions in the present study. No differences were found between the two acoustic conditions with different levels of reverberation. When interpreting the findings within the ELU framework ([391]), contrasting free-field to reverberant conditions increases temporal-envelope smearing and reduces the fidelity of phonological cues, which yields more frequent mismatches with stored phonological representations and potentially triggers a shift from rapid, implicit binding to effortful, working-memory dependent processing. Our data suggests that once this explicit route is engaged, further lengthening RT from $T_{30} = 0.5\,\mathrm{s}$ to $T_{30} = 1.0\,\mathrm{s}$ produces little additional loss in accuracy for running speech.

It could also be assumed that the differences between the free-field and the two reverberant conditions were not caused by the reverberation itself, but by the differences in SNR (free-field and treated office $\approx$ 3dB(A), untreated office $\approx$ 1.5dB(A)). However, the SNR would not predict a performance decrement between the free-field and the two reverberant conditions, but between the untreated office and the free-field or treated office, respectively. This supports the ELU prediction that even if the SNR is controlled between environments with varying reverberation, listening effort is still expected to increase with higher reverberation because the resulting increased listening difficulty may increase reliance on explicit auditory processing ([363, 361]). It also contrasts the finding by Picou et al. [363], who found no differences in listening effort for controlled SNRs. Evidently, the interplay between reverberations, the nature of the target and background noise, and their interaction is complex and not fully modelled by the ELU ([363]), necessitating more complex models (e.g., [34]).

The mixed findings in the literature and the present study suggest that RT alone may be an unreliable predictor of listening effort. Other room acoustic parameters like clarity or speech transmission index should be included in the analysis (e.g., [382, 351]), ultimately leading to a multidimensional framework of listening effort in rooms. Single-value metrics fail to capture the influences of

spatial arrangement and the nature of the background sources within the environment (e.g., [197]). The results further motivate a broader discussion about what constitutes a valid measure of listening effort. Evidence is mixed not only across measurement domains (subjective ratings, behavioural performance, and physiological indices) but also within each domain. This heterogeneity suggests that these measures capture related yet non-identical facets of listening effort, underscoring the construct's multidimensionality (e.g., [150, 140, 196]). The sensitivity of dual-task paradigms, for example, highly depends on the choice and evaluation of the secondary task ([196]), to avoid floor or ceiling effects (e.g., [415]). Our findings suggest that dual-task response times are a sensitive behavioural indicator of listening effort, but this should be cross-validated in comparison with other measures of listening effort, such as questionnaires and physiological indicators.

6.6 Limitations

Several limitations of the present study should be acknowledged. First, although we explicitly chose HTR texts with comparable difficulty according to Mohanathasan et al. [323] and modelled variation in HTR text and question difficulty via random intercepts in the primary task model, the variability in item difficulty remained substantial and accounted for a large proportion of the total variance, potentially reducing statistical power. Similarly, the presentation of a complex time-varying audiovisual VR environment might have resulted in increased variance. While moving towards more complex audiovisual paradigms and scenes may enhance the ecological validity of the test results ([232]) it can, in turn, reduce control and the generalisability of the findings to other background sound scenarios, rooms, and paradigms. However, in our view this step is needed to investigate listening effort and memory under conditions that more closely resemble the real-life. Furthermore, by employing statistical evaluation which can account for these variations, as in the present study, such effects can be reduced.

Second, the choice and implementation of the secondary task lead to constraints. The vibrotactile task is continuous and time-critical, which may have made prioritization of the primary task more challenging. Further, there was no contextual representation of the secondary task within the VR scene, which might have broken the VR illusion. Alternative implementations, such as secondary tasks more tightly coupled to the primary task and the context of the scene (e.g., measuring primary task response times or assessing responses to specific trigger words, as proposed by Buchholz et al. [78]) could further reduce temporal competition between tasks, and should be explored in future work.

Third, several aspects of the audio rendering could be improved, such as applying individual headphone equalisation and individual HRTFs. Appliance noises were further recorded outside the sound-isolated booth for practical reasons and thus contain a small amount of real-room reverberation. However, post-processing was applied and the same processed signals were used in all three acoustic conditions, ensuring that potential influences of these remaining reverberations were constant across acoustic conditions.

Fourth, it is possible that the (potentially) very small differences between the two reverberant conditions lay below the sensitivity of the study design. For example, the sample size of $N = 30$ might have been too small. However, it was selected based on an a-priori power analysis. Increasing the sample size might increase detectability of results in a follow-up study, potentially also combining different methods of listening effort assessment (e.g., [363, 415]). Further, both RTs were moderate, designed to reflect realistic scenarios. Longer RTs, e.g., in larger rooms or with more reflective materials, could lead to different results ([363]). Additionally, the background noise in the presented open-plan office was moderate. In more noisy environments, like a bar or restaurant, the influences of room acoustics might become more apparent - or be masked by the noise floor entirely. Evaluating such scenarios is also particularly relevant as they pose a challenge for people with hearing impairment ([497]).

Finally, while VR was used to provide conditions more close to real-life while retaining experimental control, there is potential for further realism. For example, all virtual humans in the environment remained at their static positions and interaction possibilities with the scene were limited. Further, while hearing booths with headphones and HMDs are standard hardware for VR studies, they might have contributed to a perceived incongruence between the VR environment and the real world. Consequently, in the context of listening experiments, VR should be regarded as a means to approximate aspects of everyday listening (e.g., spatialised, audiovisual multi-talker scenes with time-varying noise and co-verbal cues), rather than as a guarantee of realism (e.g., [31]). Still, a step towards higher realism was made in the experimental design.

6.7 Conclusion

A central aim of this study was to examine whether realistic room-acoustic differences between an anechoic free-field and a treated and untreated open-plan office affect memory and behavioural listening effort in a complex audiovisual VR environment. To this end, we combined an audiovisual virtual office scenario with animated target talkers and background agents, head-tracked spatialised audio, and time-varying background noise with the HTR dual-task paradigm.

The results show sensitivity towards room acoustic reverberation in contrast to an anechoic free-field conditions. Text comprehension and memory were only affected in single-tasking, with higher performance in free-field than in reverberant conditions. This translated into an increase of listening effort, indicated by increased secondary task response times in dual-tasking. Interestingly, participants were able to achieve better performance in the vibrotactile task in reverberant conditions both in single- and dual-tasking, which can be attributed to an increased concentration afforded to complete the task in more challenging conditions. Comparing the treated office ($T_{30} = 0.5\,\mathrm{s}$) and the untreated office ($T_{30} = 1.0\,\mathrm{s}$) did not yield differences; potential reasons for this were discussed.

Generally, the fact that a difference between free-field and reverberant conditions became apparent in text comprehension, memory performance, and listening effort indicates a sensitivity of auditory processing within reverberant compared to anechoic acoustic environments in complex scenarios. In summary, our study highlights the importance of considering room acoustics when conducting listening experiments in audiovisual environments.

Declarations

Author Contributions

The concept of this study was conceived by C.E.
The study was designed by C.E., K.L., K.B., and J.F.
K.B. and C.E. implemented the experiment, supported by K.K. and A.B.
K.B. conducted the study and collected the data.
C.E. designed the analysis plan and carried out statistical analyses with the help of K.B.
C.E. interpreted the results with the support of K.L., K.B., and J.F.
C.E. created the plots and graphics with the assistance of K.B.
C.E. and K.L. took the lead in writing the manuscript.
K.B., K.K., A.B., T.K., and J.F. reviewed the manuscript.
Throughout the project, J.F. provided supervision and all necessary resources.

Ethical Considerations

Ethical approval for this study was granted by the Medical Ethics Committee at RWTH Aachen University with the protocol number EK 23-192.

Consent to Participate

Informed written consent was obtained from all participants.

Consent for Publication

Not applicable.

Data Availability

The datasets collected and analyzed during the current study are available from the corresponding author, C.E., on reasonable request. All HTR material is openly available ([406] and [129]). The plugins used to implement the VR experiment are open-source: the RWTH VR Toolkit for basic VR interaction [157], the Character Plugin for animation of the MetaHumans [472], the Study Framework Plugin for controlling the listening experiment and logging the data [113]), and the Virtual Acoustics Plugin for communication with the Virtual Acoustics server [471]. Virtual Acoustics is an open-source software [209]. We confirm that we have reported all measures, all experimental conditions, all data exclusions (for the secondary response times), and the procedures used to determine our sample sizes.

Acknowledgements

The authors thank all participants for their time.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): SPP2236 - Project number 444724862.

Declaration of conflicting interests

The authors declare no conflict of interest.

7

Discussion and Conclusion

The overarching aim of this thesis was to determine how systematic increments in audiovisual complexity affect listening performance in terms of memory performance and listening effort. To this end, seven experiments distributed across five publications were conducted, in which aspects of acoustic, visual, and task-related complexity were manipulated in a controlled manner. These aspects included target and background source position changes, differences of VR display devices, audiovisual incongruences, characteristics of the background sounds, room acoustic variations, and implementation of different working memory and listening effort paradigms.

In this chapter, the findings are integrated into a broader conceptual framework of audiovisual complexity, as introduced in Section 1.2.6. In Section 7.1 the experiments conducted in the scope of this thesis are contextualised within the framework of complex acoustic environments by Weisser et al. [484]. Subsequently, the role of visual information in acoustic studies is discussed (see Section 7.2). Building on this, a framework of audiovisual complexity is proposed that is tailored to scene design for listening research (see Section 7.3). In Section 7.4, a novel distinction between descriptive and generative approaches to complexity is introduced and discussed, with an emphasis on how these can inform the development of listening experiments. Finally, the findings are summarised in Section 7.5 and an outlook for future research directions is given.

7.1 Acoustic Complexity Framework

With increasing realism, it becomes increasingly difficult to classify and describe acoustic and audiovisual scenarios. However, this is, on the one hand, crucial to ensure that scenarios of sufficient quality are used and, on the other hand, to compare and discuss findings from different studies in a meaningful way.

Table 7.1: Overview of the experiments presented in this thesis within the framework of complex acoustic environments by Weisser et al. [484]. A plus (+) indicates that the characteristic (column) is part of the research question of the respective experiment (row). A zero (0) indicates that the characteristic was incorporated, but not explicitly evaluated. A minus (-) indicates that the characteristic was not included in the respective study scenario.

		Source			Environment		Source-environment interactions			Receiver
		1 Multiple sources	**2** Source diversity	**3** Source-source interaction	**4** Reverberation	**5** Non-uniform medium	**6** Amplification	**7** Adaptation	**8** Other modalities	**9** Receiver's task
Chap.	**Experiment**									
2	**Spatial Target Source Alternations**									
	Experiment 1	+	+	+	-	-	-	-	-	-
	Experiment 2	+	-	+	-	-	-	-	-	-
3	**Spatial Background Source Alternations**	+	-	-	-	-	-	-	-	-
4	**Audiovisual Incongruence**									
	Experiment 1	**0**	-	-	-	-	-	-	+	-
	Experiment 2	**0**	-	-	-	-	-	-	+	-
5	**Heard-text recall Irrelevant Speech**	**0**	+	**0**	-	-	-	-	**0**	**0**
6	**Heard-text recall Room Acoustics**	**0**	**0**	**0**	+	-	-	-	**0**	**0**

The framework of complex acoustic environments by Weisser et al. [484] (see Section 1.2.6) can serve as a starting point for such endeavours. In the present thesis, it serves as a reference for a systematic overview and motivation of the presented experiments. This section contextualises and discusses the experiments within that framework.

The framework by Weisser et al. [484] categorises complexity into four types (source, environment, source–environment interactions, and receiver's task) with a total of nine characteristics. An overview over the experiments presented in this thesis in the context of the complexity framework is provided in Table 7.1. The table illustrates that six of the nine complexity characteristics were either explicitly addressed as experimental variables or are implicitly incorporated into the scene design. In the following, the role of the factors for the experiments within this thesis is discussed. Only loudspeaker amplification, source–environment adaptations (such as the Lombard effect), and non-uniform propagation media are not part of the present work and will therefore not be reported in detail.

7.1.1 Source Complexity: Multiple Sources

The first complexity characteristic, "multiple acoustic sources distributed in space" [484, p. 4], indirectly encompasses spatial reproduction, spatial configuration, and the presence of multiple sound sources. All of these three factors will be discussed in the following.

First, spatial sound reproduction was used in all seven experiments presented here. Its contribution was not specifically evaluated, as previous work has already demonstrated the benefit of reproducing accurate spatial cues (e.g., [341], but see [130]).

Secondly, the spatial configuration of target and background source locations, and by extension, changes thereof may introduce an additional layer of complexity: in the literature, some studies employ a single target talker at a fixed position (e.g., in front), whereas in others the target talker position changes. In the former, participants can maintain a stable spatial focus and suppress spatially distributed background sounds. In the latter, they must dynamically refocus their attention in space and thus also their spatial suppression (e.g., [60, 57, 278]). Investigating spatial configurations of target and background sounds and changes thereof is therefore crucial for evaluating listening performance.

Thus, the influence of spatial configurations was examined in Chapter 2 [128] (Experiment 1), which extended classic verbal serial recall by introducing spatially separated target talkers. The influence of this spatial separation, however, became apparent only at high rates of location switches that are uncommon in naturalistic conversations (cf. Chapter 2 [128], Experiment 2). For this reason, switch rates were not manipulated in Chapters 5 [125] and 6 [124] which employed the HTR; although these studies also employed spatially separated target talkers, the conversations within the HTR exhibited switch rates well below those that caused detrimental effects in Chapter 2 [128]. Furthermore, previous work has shown that spatial separation and co-location of talkers within the HTR do not affect performance [135]. Spatial background changes were examined in Chapter 3 [126], which showed that verbal serial recall is robust against such variations in spatial configuration of the background. These results suggest that while spatial information per se might be beneficial, changes in spatial configuration might play a subordinate role.

Third, with respect to multiple sound sources, a distinction is made here between target and background source diversity. Notably, in their framework, Weisser et al. [484] do not specify further whether this includes target sources, background sources, or both. This may be explained by the fact that the framework was developed for describing recordings of real-life scenes. In such scenes, the distinction between target and background talkers is often less clear-cut than in experimental designs, where usually a well-defined distinction between the target and background is made (but see, e.g., [5]). Thus, it is important to detail whether multiple target and/or multiple background sources were provided. In the present thesis, multiple target talkers were simulated in Chapters 2 [128] and evaluated with regard to their spatial configuration (see also Chapters 5 [125] and 6 [124]). Multiple background sound sources were further present in Chapters 2 [128] (Experiment 1), 5 [125], and 6 [124], and were evaluated with regard to source diversity, which contributes to complexity characteristic 2.

7.1.2 Source Complexity: Source Diversity

Characteristic 2 of the framework by Weisser et al. [484] concerns acoustic source diversity. Again, this can be subdivided into target and background source diversity. Targets in this thesis were always controlled speech stimuli. The isolated digit words in Chapter 2 [128] were replaced by the HTR sentences (e.g. [129]) in Chapters 5 [125] and 6 [124], which provide more contextual cues than digit words, increasing complexity. However, there is still room for improvement in the diversity of the stimuli. For example, systematic variations within target speech (e.g., gender, accent, speech rate) were not examined, even though Cherry [85] already noted these as relevant aspects contributing to auditory scene analysis. Further, the HTR speech material is based on recordings where talkers read out the HTR material instead of speaking freely. This does not constitute a natural conversation, neither in prosody nor in conversational flow. Still, the HTR naturally includes more aspects of speech variety than the serial recall, especially in the longer sentences of the HTR, representing an increase in target source diversity. However, conversational speech corpora with realistic prosody exist; e.g., Miles et al. [316] generated a corpus which contains recordings of free instead of read conversational speech. These could be employed in future studies as an even more complex alternative to the HTR.

Background source diversity was a key focus in Chapter 2 [128] (Experiment 1) and Chapter 5 [125], which directly contrasted silence with either semantically meaningful speech or pseudo-speech without meaning. Both paradigms employed in these studies (serial recall and HTR dual-task) proved sensitive to the influence of semantics, highlighting the relevance of competing intelligible speech in complex acoustic environments. An example of a complex acoustic environment incorporating diverse speech and non-speech stimuli was provided in Chapter 6 [124], which utilised a time-varying background scene.

7.1.3 Source Complexity: Source-Source Interactions

In characteristic 3, Weisser et al. [484] highlight the relevance of source–source interactions, i.e., communication between sources, noting that "the sound produced by several sources (e.g., talkers) is not the same as each of them alone" [484, p. 4]. Within this thesis, source-source interactions were considered by talker interactions and turn-taking.

Target talker interactions were the explicit research focus of Experiments 1 and 2 in Chapter 2 [128], which examined turn-alternation rates and found that they only affected memory performance at unnaturally high rates. Target talker interactions were also implicitly modelled in Chapter 5 [125] and Chap-

ter 6 [124], as the target stimuli within the used HTR paradigm were presented as conversations between two talkers.

Besides in Chapter 6 [124], where distractor talkers reacted to incoming phone calls, interactions between background sources were deliberately omitted. More precisely, in Chapter 2 [128] (Experiment 1) and Chapter 5 [125], background talkers were present but did not interact: two background talkers were auralised as speaking constantly rather than alternating. This was done to eliminate potential effects of spatial and temporal variations in the background signal caused by, e.g., one talker pausing to listening to the other. While such a situation is not necessarily unrealistic — both background talkers could, for instance, be engaged in monologues during phone calls instead of talking to another, especially since in both experiments no visualisation suggested otherwise — it is likely that the influence of spatial target talker switches interacts with spatial background talker switches. Conversational dynamics of this kind may be crucial in realistic multi-talker situations and provide motivation for future studies.

7.1.4 Environmental Complexity

Environmental complexity was examined with regard to characteristic 4, which encompasses reflections, reverberation, scattering, diffraction, and diffusion [484]. These aspects were the key interest of the study presented in Chapter 6 [124], which assessed memory and listening effort under varying room acoustic conditions. An anechoic free-field condition was compared to two room acoustic simulations of an open-plan office: one version with acoustic treatment and one without. The results did not reveal differences between the two simulated offices, but performance did reliably differ between anechoic and reverberant conditions. This highlights the relevance of incorporating room acoustic simulations when studying complex acoustic environments.

7.1.5 Visual Complexity

Cues of other modalities (characteristic 8) were examined in the form of visual information. Chapters 4 [127], 5 [125], and 6 [124] utilised ECAs to embody target talkers and, in Chapter 6 [124], also to embody background talkers.

Chapter 4 [127] directly evaluated the effect of visual cues on working memory performance. There, a particular focus was placed on the influence of the display device (computer monitor vs. HMD) and on audiovisual incongruences (spatial vs. voice incongruence). HMD presentation influenced the subjective perception of the scene, as the ECAs were rated as more natural when presented via HMD compared to the computer screen. Moreover, improved ECA animation, for example through gaze behaviour, also increased participants' subjective evaluations.

Behavioural results, however, showed a robust insensitivity to the display device and incongruence, suggesting that the role of visual cues strongly depends on the dependent variable. Visual cues that do not directly support the listener's task may not inform performance and, depending on task load, may even go unnoticed [426]. The study further employed no background sounds. While there is a huge body of literature showing that in adverse listening scenarios, audiovisual stimuli outperform auditory stimuli (see also [122]), Chapter 4 [127] suggests that this is not the case in silent conditions. Further evidence in this regard is also provided by Mohanathasan et al. [322] who reported that in silence, animated ECAs can increase listening effort, potentially due to the additional cognitive cost of integrating auditory and visual cues.

Taken together, these findings paint a mixed picture about the influence of co-verbal visual cues on memory performance. Predictions regarding the influence of audiovisual stimuli on performance should thus always be based on a multifactorial assessment of the scene, incorporating aspects such as the background sounds and the specific task demands.

7.1.6 Receiver's Task Complexity

According to Weisser et al. [484], the experimental paradigm, i.e., the receiver's task (characteristic 9), is a significant contributor to complexity. Two tasks were employed in this thesis: the serial recall (Chapters 2, 3, 4) and the HTR in a dual-task design (Chapters 5, 6). Thus, there was a progression in complexity not only with respect to scene design but also with respect to the task. None of the studies directly contrasted both tasks, but some indirect comparisons are possible. Experiment 1 in Chapter 2 [128] and the experiment presented in Chapter 5 [125] used comparable background stimuli (silence, pseudo-speech, intelligible speech). Both the HTR and the serial recall showed similar error rates in silence and were similarly influenced by the background stimuli (cf. Figures 2.2 and 5.3a), which supports comparability between tasks.

One disadvantage of the HTR compared to the serial recall is the limited number of HTR texts (34), which effectively constrains the number of experimental conditions. Further, a HTR trial is more time-consuming than a serial recall trial. If participants register for multiple HTR studies, they may also remember the texts and questions, which is less of a risk for the random digit sequences in the serial recall. Difficulty across HTR texts has also been shown to differ significantly [323], while it is more balanced within the serial recall. This restricts flexibility when utilising the HTR compared to the serial recall. Still, since the HTR is based on running speech instead of isolated digits, it can be assumed that it challenges speech processing in a way that is more similar to real-life

behaviour than serial recall. In comparison to established paradigms based on matrix sentences, the HTR provides prosody, a more diverse vocabulary, and a higher variation in sentence structure. Thus, in contrast to the classic serial recall task, the HTR is a valuable paradigm when investigating listening performance and aiming for increased complexity. Together with a secondary task, it further enables the examination of listening effort.

Evaluating the secondary task utilised in combination with the HTR task raised several issues in this thesis. In both studies that employed the HTR in a dual-task design (Chapters 5 [125] and 6 [124]), a vibrotactile secondary task as suggested by Mohanathasan et al. [320] was used. In these studies, low error rates in the secondary task relative to the primary HTR suggest suboptimal prioritisation. A possible explanation for this is that the time-critical nature of the vibrotactile task may have taken precedence over the primary task, which required retaining information over a longer period. Participants may only have realised at the time of answering the HTR questions after the text presentation that they had not attended to the story in sufficient detail and that a different prioritisation between the primary and secondary tasks would have been advantageous.

An alternative secondary task potentially reducing such problems was proposed by Buchholz et al. [77], albeit for a different primary task. In this secondary task, listeners respond to specific trigger words occurring in the target text, for example weather-related vocabulary, names, or places. In this way, listeners perform two tasks without being forced to split their attention between listening and an unrelated modality. A similar approach would be possible for the HTR, for example by instructing listeners to respond to names, numbers, or relationship words in the story. A comparative study directly contrasting these secondary tasks would therefore be advisable to assess their relative suitability and applicability for measuring listening effort in HTR-based paradigms.

In summary, the increments in complexity within the characteristics of Weisser et al. [484] paint a mixed picture with regard to their influence on listening performance in audiovisual scenarios. Many acoustic variations, such as changes in stimulus content or reverberation time, showed clear effects, whereas others, such as changes in spatial source configuration, did not. Further, no clear prediction can currently be made regarding visual complexity; the influence of visual cues seems to depend strongly on the scene design and the evaluation method. This opens up a broader discussion about the role of visual cues in listening related research.

7.2 The Role of Visual Cues in Listening Research

As listed above, one major contributor to complexity examined in this thesis was the addition of visual cues. The related findings, however, painted a mixed picture, highlighting a need for a more in-depth examination of visual cues. This section therefore discusses the role of visual cues by introducing and highlighting two contrasting accounts: visual cues as a subcomponent of auditory information and visual cues as an independent contributor to complexity.

7.2.1 Visual Cues as a Subcomponent of Auditory Information

Weisser et al. [485] conceptualised audiovisual interaction as a subcomponent of acoustics within their framework for complex acoustic environments. This secondary nature is understandable given their goal of providing an acoustically oriented description of scenes. When focusing on classic listening performance metrics, visual cues (e.g., through visual VR) have been shown beneficial when they are directly relevant to the auditory stimulus, e.g., through visual speech cues [445], or by providing semantic context [63]. If in contrast, the visual information is not directly necessary for achieving a specific goal, it might not affect performance. For example in Chapter 4 [127], no effect of variations in visual information was found; neither the display device (HMD vs. computer monitor) nor an explicit audiovisual incongruence (spatial and voice mismatch) had a significant impact on performance. As argued before, one plausible explanation is that the visual information was not directly task-relevant in that study: no background sound was present, and the task could be accomplished based on the auditory information alone. The relevance of task-dependency is supported by the literature, where the benefit of lip-movement in auditory tasks has been shown to depend strongly on the background sounds; participants rely more on lip-reading if they have difficulties listening to the target [165]. From this perspective, visual cues modulate complexity only insofar as they change the effective auditory task demands. Thus, it is understandable that some accounts, such as the framework by Weisser [484], argue that visual information can be evaluated solely in terms of its interaction with the auditory environment, which in turn directly informs the relevance and necessity of visual cues.

7.2.2 Visual Cues in Relation to other Factors of Complexity

However, when moving from purely acoustic to audiovisual environments, it becomes less clear whether treating the visual modality as a subcomponent of auditory information is sufficient. As discussed in Section 1.6, visual cues can be interpreted as a critical part of a holistic scene impression, which in turn influences

cognition and perception well beyond classic performance metrics. For example, visual information has been shown to influence the sense of presence, engagement, and motivation, convey information about the context and the environment, and communicate emotions (see Chapter 1).

Considering this, it may be more informative to ask not only *whether* visual information improves or worsens task performance, but also why visual information sometimes *fails* to influence performance, despite being noticed and potentially altering the overall experience of the scene. This shifts the focus from the interaction between the auditory and visual information alone to the interaction between the audiovisual scene and the task. In other words, many tasks may not be sensitive enough to capture aspects of a scene, such as visual information, which still may be relevant to naturalistic listening.

Instead of focussing on performance, a shift should be made towards experience based metrics. An example for a metric following such an approach is the construct of listening effort (see Section 1.7.2), which postulates that even if similar performance is achieved in two conditions, the effort elicited to achieve this performance (as part of the listening experience) may differ. Another example for a less restricted assessment method is moving to observation-based methods, e.g., neural imaging techniques. Several studies, e.g., immerse multiple participants at once in virtual acoustic environments and analyse their communication behaviour [33, 32, 313, 315, 284], e.g., through interpersonal distance, heart-rate variability, voice level, or similar. Such approaches typically require extensive instrumentation (e.g., motion tracking, eye tracking, and physiological or neural sensors), but they can provide insights into subtle variations that might be missed by more traditional paradigms. They also inherently support interactivity, which many authors consider highly relevant for ecological validity, but which is still rarely implemented (e.g. [350, 475]).

Tightly controlled setups might conceal effects, but also the contrary is possible: highly artificial tasks and scenarios can create situations in which effects are detectable but of limited real-world relevance especially if the specific situation almost never occurs in everyday life. An example for this is found in Experiment 2 of Chapter 2 [128] where a decrement in serial recall performance was found for alternating target sources only if this alternation happened at rates uncommon for naturalistic conversations. This means that decrements by spatial changes have only limited real-world applicability. For example, in two studies investigating spatial position changes with the HTR instead of the serial recall, leading to slower rates due to the higher item duration (sentences in the HTR instead of digit words in the serial recall) no effect of spatial changes on performance were found [135, 319]. This shows that paradigms which mirror the real-life better might result in more generalizable findings and supports the use of meaningful,

connected speech material when aiming for valid findings (e.g. [316]). Similar findings have been made for controlled auditory selective attention tasks [57].

Not only the task or measurement method, but also the participants themselves influence how audiovisual information is processed. People with, e.g., hearing loss tend to rely more on visual cues than normal-hearing listeners, both for lip-reading and for emotion recognition [165, 366]. Hearing loss can also influence turn-taking behaviour [360], which may rely both on auditory and visual cues. In contrast, a restricted field of view may shift processing demands more strongly to the auditory domain. Beyond impairments, personal experiences may also shape how participants interact with components of the audiovisual environment. For example, talkers adapt their communication behaviour if conversing with hearing impaired people [359]. This may translate to interactions with elderly ECA and could hold true also for other social attributes [350]. Manipulations of ECA appearance might therefore allow for studying how people speak and listen to specific target groups and how this in turn affects audiovisual processing.

Considering relationships between participants and ECAs is especially relevant since many studies employ young homogeneous ECAs and recruit mostly young, normal-hearing participants, often from university contexts [187]. Their behaviour in experiments may differ from that of the population. Diversification of ECAs and participant groups should thus be considered. Further, differences in technological affinity and prior experience may be particularly relevant in studies using VR. Besides the appearance of the ECAs, the appearance of the participant might also be relevant. Few studies include an explicit embodiment of the receiver, even though embodiment is likely to influence perception and social cognition [350].

Taken together, this suggests that audiovisual interaction is not merely a subcomponent of acoustic information. Rather, it interacts with several other dimensions of complexity, including the environment, the receiver, and the technical implementation. This motivates the development of a framework of *audiovisual* complexity, as proposed in the following section.

7.3 Audiovisual Complexity Framework

Audiovisual complexity is often increased with the goal of approaching real-world listening situations and thereby achieving higher ecological and/or external validity. However, no systematic framework for audiovisual complexity currently exists (but see [6, 232]). As a consequence, and given the inherently multidimensional nature of complexity, the isolated and combined contributions of individual complexity dimensions to listening performance are still not fully understood.

Building on the framework for complex acoustic environments by Weisser et al. [484], this thesis therefore proposes a framework of *audiovisual complexity*. In contrast to Weisser et al. [484], the visual modality is not treated merely as a subcomponent of acoustics, but as an independent dimension. Based on the literature discussion in the Introduction (Chapter 1) and the findings of this thesis, an effort was made to find visual counterparts to the acoustic characteristics of the framework by Weisser et al. [484]. This includes, e.g., also technical aspects of auditory and visual presentation (see Chapter 4 [127]). Further, the dimension of participant complexity is added, since participant-related factors can shape how auditory and visual information is processed (see Section 7.2.2).

The framework is intended to reflect the multidimensional nature of audiovisual scenes as discussed in Section 7.2, in which visual information affect not only performance in controlled auditory tasks but also, e.g., experience, behaviour, engagement, and social interaction. In its current form, it is solely based on the findings and literature background of the present thesis. A systematical evaluation of the framework, potentially resulting in a meaningful assessment scheme for listening experiments, is advisable but beyond the scope of this thesis.

The proposed framework is summarised in Table 7.2. It lists auditory and visual factors of complexity that were described and, in some cases, explicitly examined in the context of this thesis. First, the structure follows three main types of complexity of the original acoustic framework (source, environment, source–environment interactions) and extends it by additional categories of participant-related and technical aspects.

For each type, auditory and visual characteristics are listed in parallel to emphasise both modality-specific contributions and potential cross-modal interactions. For example, source complexity, spatial distribution, number of sources, source diversity, and source–source interactions are specified separately for auditory and visual cues, reflecting that spatial separation or turn-taking behaviour can be conveyed acoustically (e.g., via SRM and prosody) as well as visually (e.g., via positioning, gaze, and gestures). Similarly, room and environmental characteristics include both classic acoustic properties (reverberation, scattering, diffusion) and visual attributes such as audiovisual room congruence and the fidelity of the environment visualisation.

A central addition compared to the original acoustic framework is the explicit inclusion of participant-related and technical aspects. Participant characteristics comprise not only the task but also factors such as hearing status, language abilities, visual abilities, and the relationship to the acoustic and visual scene (including ECAs). Technical aspects include how the scene is reproduced and interacted with, for instance via the auditory resolution and display device, or

Table 7.2: Suggestion for a framework of audiovisual complexity for listening-related research, based on the framework for auditory complexity by Weisser et al. [484].

Type	Characteristic	Auditory	Visual
Source	Spatial distribution	Spatial relationship between target and background sources, directivity	Spatial congruency, orientation
	Multiple sources	Multiple target and background sources	Embodiment
	Source diversity	Speech and non-speech sounds, different languages, accents, gender, age, isolated items vs. running speech (prosody), emotions	Animations, e.g., lip movements, gestures, facial expressions, diversity of visualisations (e.g., age, gender, appearance)
	Source-source interactions	Turn-taking prosody of target and background sources, speech context, trigger-signal reactions (e.g., phone calls)	Turn-taking behaviour (gaze, gestures, breathing), co-presence and awareness signals
	Modality Interaction	Audiovisual source congruence (spatialisation, animation, gender, age, etc.)	
Environment	Room / environment	Reverberation, reflection, scattering, diffraction, diffusion	Fidelity of environment visualisation, environment animations
	Nonuniform medium	Auralisation according to weather (temperature, humidity, wind)	Environmental animations (e.g., snow, rain, wind, lighting)
	Modality Interaction	Audiovisual room congruence (size, materials, furniture, etc.)	
Source-environment interactions	Loudspeaker amplification	Public address systems, device directivity and transfer characteristics	Corresponding visualisation of devices
	Source-environment adaptation	Lombard speech	Lombard-related animations, simulated awareness of the environment and the participant (e.g., through gaze)
Participant	Participant's task	Explicit task or behaviour observation, interaction possibilities	Availability and necessity of audiovisual information for the task, interaction requirements
	Participant	Hearing status, language abilities, relationship to the acoustic scene and context, noise sensitivity	Visual abilities, affinity for technology, relationship to the visual scene and ECAs, embodiment of the participant, sensitivity for motion sickness
Technical aspects	Reproduction	Auditory resolution, simulation vs. measurement/recording	Display device, simulation vs. video
	Interactivity	Adaptive auralisation, movement possibilities, reactivity of the presented scene to user input and behaviour	Movement possibilities, controllers, reactivity of the presented scene to user input and behaviour, embodiment of participant movements and potentially controllers

via the degree of interactivity and the responsiveness of the scene to user input and behaviour.

Table 7.2 does not claim to be exhaustive and should not be understood as a strict checklist that must be fulfilled in every experiment. Rather, it is a starting point for future studies. It is intended as a generative tool for structuring design choices, making explicit which aspects of audiovisual complexity are included, controlled, or deliberately omitted, and for motivating incremental changes across studies. Therein, the framework shifts the focus from the scene alone to a more comprehensive overview that also encompasses task and participant characteristics. It could also serve as starting point for future models that aim to predict listening performance from combined auditory, visual, participant-related, and technical information (see [154]).

7.4 Descriptive versus Generative Complexity

The framework of complex acoustic environments by Weisser et al. [484] and the audiovisual complexity framework proposed here differ in their fundamental approaches. The framework by Weisser et al. [484] was designed to *describe* and assess real-life scenes, for example via recordings, with regard to their complexity. In contrast, in this thesis the complexity of experimental scenes was *generated* in order to evaluate established paradigms under increasingly complex conditions. The proposed framework of audiovisual complexity also serves as a guideline for complexity increments. This difference in perspective motivates a differentiation between these approaches. Therefore, this thesis proposes a distinction between *descriptive* and *generative* approaches to complexity.

In descriptive approaches, complex environments (or recordings thereof) are used to observe measures of behaviour, and an effort is made to link findings to aspects of the scene, such as the SNR, stimulus content, or spatial distribution of sources (see, e.g., [400]). By contrast, in generative approaches, complexity is gradually increased starting from simplified laboratory scenarios, as in the present work. This allows for the isolated examination of individual increments and for systematic testing of how specific factors contribute to performance and experience.

Although distinct, both approaches should ideally inform each other in an iterative process: increments in generative approaches should be inspired and constrained by those observed in descriptive studies, while recordings used in descriptive approaches should be selected and analysed with reference to those dimensions of complexity that have proven influential in generative studies.

Because of the different nature of descriptive and generative approaches, the methods employed often differ. Descriptive studies often rely on recordings or

data collection in real-life settings (e.g., ECA, soundwalks). These offer high external validity but limited experimental control. Generative approaches, by contrast, typically aim for higher internal validity, often at the expense of external validity. Simulated scenarios are common in this context, as they allow precise manipulation of individual factors while keeping others constant. As a consequence, there is often a conceptual and qualitative difference between the visual and auditory cues presented in studies following the two approaches.

Through these different methods, descriptive and generative approaches differ with respect to the quality of visual cues and interactivity provided. In recordings visual cues are presented with photorealism, but they do not allow for interactivity with the scene, e.g., through movement, as the position of the participant in the scenario as well as the temporal course of events is predefined. Simulations allow for such interactivity, at a cost of cue fidelity. Theoretically, it is also possible to achieve conditions close to photorealism also in simulations, as demonstrated in contemporary video game development. For example, MetaHuman (Epic Games, Inc., United States of America) advertises "high-fidelity" [120] ECAs that convincingly mimic real-life behaviour in short, pre-rendered videos (e.g., [2, 310]). While this performance may be achievable in offline renderings, real-time applications such as VR HMD presentations are constrained by the computational capabilities of typical computer systems and therefore fall short of these promises. Thus, such a level of detail is often not feasible for research purposes. This limitation was also apparent in the present thesis, where MetaHumans were employed as ECAs (see Chapter 4 [127] Figure 4.7, Chapter 5 [125] Figure 5.1, and Chapter 6 [124] Figure 6.1, as well as the supplementary video in [114]). Although MetaHumans may provide high quality visualisations compared to other available techniques (e.g., [347, 337]), they still did not reach the realism of actual real humans, which was also reflected in participants' critical evaluations of the ECAs in the questionnaires used in Chapter 4 [127]. While it has been shown that virtual speakers can yield audiovisual benefits even if they are evaluated negatively [337], benefit from virtual humans are generally smaller than those provided by video recordings, and it remains unclear how other factors such as motivation, engagement, immersion, or the uncanny valley effect [326] in VR compared to video presentations contribute to a holistic assessment of listening performance and experience. In addition to creation, simulated environments also pose challenges for long-term maintenance, especially given the rapid pace of technological development, as pointed out by Parsons [350].

Furthermore, by employing simulated environments, a risk, but also a strength, of generative approaches is that they may also focus on variations of individual factors that either have limited direct relevance in everyday life or may not occur in reality at all (e.g., audiovisual incongruences as examined in Chapter 4). The

relevance and implications of such variations should be carefully considered — or as Chubala et al. [88] put it humorously: "if the experimenter unexpectedly pokes the subject at a critical time, performance [...] may suffer, but for uninteresting reasons" [88, p. 1]. However, manipulations with clear theoretical motivation can provide valuable information for evaluating theoretical models [88]. In this sense, generative complexity approaches can support model-building processes potentially better than descriptive approaches because they extend beyond typical real-world conditions and allow for testing the boundaries of cognitive mechanisms.

In summary, both descriptive and generative approaches therefore offer advantages and challenges. Which approach is more appropriate ultimately depends on the research question: whether the primary goal is to understand behaviour in a realistic, unconstrained setting, or to isolate and quantify the effects of clearly defined components of complexity. At present, it may seem as if descriptive and generative approaches will not fully converge in terms of the realism and authenticity of the environments they employ. In the future, however, it may be possible that problems that appear difficult or even impossible to solve today, such as achieving convincing photorealism in real-time simulations or limitations in recording and reproduction, will be solved in 10 to 20 years, or even sooner. Then, simulated environments would no longer necessarily be at a disadvantage compared to recorded ones.

At the current point in time, however, studies using simulated (VR) environments almost always entail an additional, often implicit, research question: can the simulation (and/or VR) be used as a valid tool to assess the phenomenon or criterion of interest? If no direct comparison to non-simulated conditions (e.g., videos or real-life recordings) is included, as is also the case in the present thesis, the generalisability of the results is therefore always limited to comparable scenarios. While this is a clear limitation of the current thesis, this generalisability and classification of scenarios opens a new field of research.

7.5 Conclusion and Outlook

This thesis systematically increased the complexity of experimental listening scenarios through audiovisual simulation, with the aim of approximating real-life conditions more closely. Complexity increments were implemented in several domains: by introducing spatial audio reproduction and spatial position changes, by varying background stimuli, by evaluating VR display devices and audiovisual incongruences, by examining room acoustic influences, and by utilising alternative paradigms for assessing memory performance and listening effort.

Most acoustic variations, such as changes in stimulus content and room acoustics, showed clear effects on listening performance, conceptualised in terms of memory performance and listening effort. For example, semantic meaningful background sounds showed a significantly stronger influence on listening performance than non-meaningful speech, and anechoic conditions lead to significantly different performance than conditions with room acoustic simulations. By contrast, the role of visual cues was less clear. Visual information appears to be crucial in some applications, for example when examining speech intelligibility in the presence of background sounds [122], but its contribution is more ambiguous under favourable acoustic conditions. Audiovisual effects also strongly depended on the task and outcome measure (see Chapter 4 [127]) as well as the participant abilities (see Section 7.2), emphasising the importance of considering the participants and their task as a component of audiovisual complexity (see also [484] and Section 7.3).

Importantly, these findings underline that current unimodal complexity considerations [484] are not sufficient any more in interdisciplinary research, since they do not yet incorporate the role of visual information to a sufficient degree. Based on this, and to move towards a more comprehensive model of real-world listening performance, this thesis proposed a novel framework of audiovisual complexity that explicitly considers auditory, visual, and audiovisual characteristics, as well as properties of the participant and the technical implementation (see Section 7.3).

While several aspects of this framework were explicitly examined in this thesis, there are multiple aspects which offer directions for future research which have been addressed in the previous sections. Examples include a more systematic investigation of participant characteristics, such as hearing loss, visual impairment, and language background; the use of more diverse groups of ECAs with respect to age, gender, appearance, and perceived hearing status; and the implementation of more advanced interactivity, in which the behaviour of ECAs dynamically adapts to the listener. Further, future studies could explicitly examine the influence of variations in vocal effort within audiovisual scenes, embed more natural conversational speech material, and decompose embodiment into more detailed visual cues (e.g., gaze, location, mimics). This stresses the importance of interdisciplinary research, especially from the fields of acoustics, computer vision, and psychology. Further studies could also address more senses besides audition and vision, e.g., by providing olfactory and haptic cues.

Another important contribution of the present thesis is the proposed distinction between descriptive and generative complexity approaches (see Section 7.4). In many studies, complexity is addressed in methodologically different ways, e.g., through real-life observations or simulations, potentially limiting comparability.

This highlights the need to connect controlled experimental work with observations from real-life listening situations. The studies presented in this thesis as well as the audiovisual complexity framework proposed here are intended as one step in this direction.

Ultimately, research in complex audiovisual environments should aim to bridge the gap between theoretical model development and practical applicability, ideally helping to address the challenges faced by vulnerable listener groups, such as hearing impaired people, in their everyday life. The findings, the audiovisual complexity framework, and the directions for future research provided in this thesis may contribute to this goal.

Appendices

A

Copyright Clearances

The supplementary material below contains the copyright clearances from the publications incorporated in this thesis.

A.1 Chapter 2

Copyright information for the article: C. A. Ermert, M. Yadav, J. E. Marsh, et al. "Serial recall in spatial acoustic environments: irrelevant sound effect and spatial source alternations". In: *Scientific Reports* 15.1 (2025).

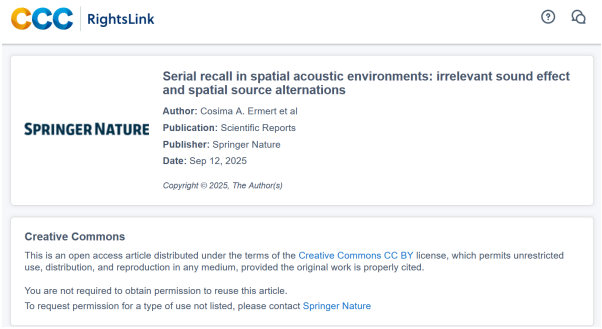


Figure A.1: Copyright of [128]. Accessed 17.02.2026

A.2 Chapter 3

The article C. A. Ermert, L. Vollmer, et al. “Spatial-changing irrelevant speech in visuo-spatial and visuo-verbal serial recall”. In: *tbd* (under Review) is currently under review. Thus, no licence can be obtained.

A.3 Chapter 4

Copyright information for the article: C. A. Ermert, M. Yadav, J. Ehret, et al. “Audiovisual angle and voice incongruence do not affect audiovisual verbal short-term memory in virtual reality”. In: *PLoS One* 20.8 (2025).

Citation: Ermert CA, Yadav M, Ehret J, Mohanathasan C, Bönsch A, Kuhlen TW, et al. (2025) Audiovisual angle and voice incongruence do not affect audiovisual verbal short-term memory in virtual reality. PLoS One 20(8): e0330693. https://doi.org/10.1371/journal.pone.0330693

Editor: Patrick Bruns, University of Hamburg, GERMANY

Received: October 16, 2024; **Accepted:** August 4, 2025; **Published:** August 22, 2025

Copyright: © 2025 Ermert et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Figure A.2: Copyright of [127]. Accessed 17.02.2026

A.4 Chapter 5

Copyright information for the article: C. A. Ermert, S. J. Schlittmeier, et al. “Heard-Text Recall and Listening Effort under Irrelevant Speech and Pseudo-Speech in Virtual Reality”. In: *Acta Acustica* 10 (2026).



This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Figure A.3: Copyright of [125]. Accessed 17.02.2026

A.5 Chapter 6

Copyright information for the article: C. A. Ermert, K. Loh, et al. “Listening Effort In Complex Audiovisual Scenes Under Varying Simulated Room Acoustic Conditions”. In: *Trends in Hearing* 30 (2026). Copyright ©2026 by The Author(s). Reprinted by Permission of Sage Publications.

Rights and permissions

© The Author(s) 2026.



This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage). Request permissions for this article.

Figure A.4: Copyright of [124]. Accessed 26.08.2026

SAGE PUBLICATIONS LICENSE
TERMS AND CONDITIONS

Aug 26, 2026

This Agreement between Cosima Antonia Ermert ("You") and Sage Publications ("Sage Publications") consists of your license details and the terms and conditions provided by Sage Publications and Copyright Clearance Center.

License Number	6335930016394
License date	Aug 25, 2026
Licensed Content Publisher	Sage Publications
Licensed Content Publication	Trends in Hearing
Licensed Content Title	Listening Effort in Complex Audiovisual Scenes Under Varying Simulated Room Acoustic Conditions
Licensed Content Author	Cosima A. Ermert, Karin Loh, Karl Baylan, Konstantin W. Kühlem, et al.
Licensed Content Date	Jan 1, 2026
Licensed Content Volume	30
Licensed Content Pages	19
Type of Use	Dissertation/Thesis
Requestor type	I am the author of the requested article
Format	Print and electronic
Portion	Full article
Will you be translating?	No, only English
Circulation	10000
Title of new work	Simulating complex listening scenarios in audiovisual virtual reality - The impact of audiovisual cues, spatial noise, and room acoustics on listening performance
Institution name	Institute for Hearing Technology and Acoustics, RWTH Aachen University
Expected presentation date	Sep 2026
The Requesting Person / Organization to Appear on the License	Cosima Antonia Ermert
Requestor Location	Cosima A. Ermert Kopernikusstr. 5 Aachen, 52074 Germany
Payment Type	Invoice
Email Address	cosima.ermert@akustik.rwth-aachen.de
Billing Address	Mrs. Cosima Ermert Kopernikusstr. 5 Aachen, Germany 52074
Total	0.00 USD

Terms and Conditions

Sage Terms and Conditions for Permissions Administered Through RightsLink

You, the Requestor, are requesting to use the material specified in the permission request (the "Work"). Your agreement to the terms and conditions herein and completion of a permission request constitutes a Permission Request. You are in no way required to use the Work; however, should you decide not to use the Work after you complete this request, you agree to cancel your order through this website. Under the above conditions, and the following terms and conditions, permission to use Work ("Permission") is granted solely to you:

1. Sage reserves the right to revoke any Permission, at Sage's sole discretion, within two (2) business days of the request.
2. The number of copies ("Copies") for print use is defined as the total number of copies made for distribution or for repurposing, and the number of copies ("Copies") for electronic use is defined as the total number of viewers of the Work, recipients of copies of the Work, and individuals or entities who may have access to the Work. The Copies must not exceed the Copies as stated in the Permission Request.
3. Requests to post a full article on a website, internet, intranet, or any publicly accessible site must be submitted to the Publisher through the Sage Permissions Portal (https://sage-cloud.atlassian.net/servicedesk/customer/portal/10).
4. Permission is granted only for the type of use specified in the Permission Request. If any information pertaining to your Permission Request changes, you must cancel this request and resubmit a new request with the correct and current permission request information. Sage may exercise its right to revoke Permission, if Sage finds, in Sage's sole opinion, that the context in which you have used or repurposed the Work is considered libelous, in violation of any right of privacy, or otherwise unlawful; infringement or in violation of any copyright or other proprietary right of others; or can be construed to possibly cause harm or injury. You agree that use of Work will be professional, in the context of fact-based and generally acceptable professional practices.
5. Permission does not include the use within Custom Publishing Programs, and all use within such programs is explicitly prohibited.
6. Permission does not include use of the material within Massive Open Online Courses (MOOC's). For permission to include material in a MOOC, please contact Sage directly through the Sage Permissions Portal (https://sage-cloud.atlassian.net/servicedesk/customer/portal/10).
7. If your Permission Request includes the right to translate the Work, you agree that the translation of the material shall be made faithfully and accurately, and that abbreviations and/or alterations in the text and/or title of the Work shall be made only with Sage's prior the written consent. Requestor shall not own or acquire any copyright or other proprietary rights in the Work or any other material furnished by Sage, including without limitation translations or transcriptions thereof, all of which rights shall be owned by and/or are hereby assigned to Sage. Requestor shall indemnify Sage against any and all claims, including without limitation attorneys' fees and legal costs, that concern or relate to (a) inaccurate translation or transcription of the Work, (b) infringement claims arising out of the inclusion of material not furnished by Sage or (c) infringement or other claims asserted by anyone retained by Requestor to translate the Work. Requestor agrees that the name of the Author (s), Copyright Holder, and Publisher shall appear in due prominence on the title page of every copy of the translation and in all advertisements for the translation, and that the translation shall include: (1) the Work's original copyright notice and original American title, both in English, and (2) notice in granted translated language in identifying the Sage Publications as the original publisher and stating the translation is published by arrangement with Sage. The rights licensed to Requestor are not transferrable. The translated article reprint must include the following disclaimer in English and in the language of the reprint: "While every effort has been made to ensure that the contents of this publication are factually correct, neither the authors nor the publisher accepts, and they hereby expressly exclude to the fullest extent permissible under applicable law, any and all liability arising from the contents published in this Article, including, without limitation, from any errors, omissions, inaccuracies in original or following translation, or for any consequences arising therefrom. Nothing in this notice shall exclude liability which may not be excluded by law. Approved product information should be reviewed before prescribing any subject medications."
8. Permission is granted for prospective use only, and does not apply to any use that has occurred prior to the date of the Permission Request.
9. Permission does not apply to any material (reprints, illustrations, figures, tables, etc.) not controlled by Sage. Requests for permission to re-use third-party material should be submitted to the original copyright holder, as indicated in the credit line for the materials.
10. Full acknowledgment of your source must appear in every copy of your work as follows:

    Author(s), Book/Journal Title (Journal Volume Number and Issue Number) pp. xx-xx. Copyright © YEAR by (Copyright Holder). Reprinted by Permission of Sage Publications

11. Unless specified in the request or by prior arrangement with Sage, payment is due from you within sixty (60) days after the completion of Permission.
12. It is assumed that the Requester is using the selection in question, and is subject to billing and collections procedures, unless otherwise noted.
13. Permission Requests for reuse of text excerpts shall not exceed 50% of a journal article's content.
14. No more than 20% of any one Sage book or journal issue may be reused at one time.

Other Terms and Conditions:

v4.06

Questions? customercare@copyright.com.

Figure A.5: Licence of [124].

Bibliography

[1] 1DStudio. *Big Office. Asset Package for Unreal Engine*. 2020.

[2] *A Look at MetaHuman in Unreal Engine*. 2025. URL: https://www.youtube.com/watch?v=doax4FU1-t0.

[3] T. Afghah, J. Heeren, L. Hartog, P. Biermann, A. Wulff, A. Warzybok, and K. C. Wagener. "An Open-Access Dataset of Unaided and Hearing Aid-Assisted Auditory Perception Measures in Complex Virtual Acoustic Scenes". In: *International Journal of Audiology* (2025), pp. 1–14. DOI: 10.1080/14992027.2025.2601653.

[4] S. Agrawal, A. Simon, S. Bech, K. Bærentsen, and S. Forchhammer. "Defining Immersion: Literature Review and Implications for Research on Audiovisual Experiences". In: *Journal of the Audio Engineering Society* 68.6 (2020), pp. 404–417. DOI: 10.17743/jaes.2020.0039.

[5] A. Ahrens and K. D. Lund. "Auditory Spatial Analysis in Reverberant Multi-Talker Environments with Congruent and Incongruent Audio-Visual Room Information". In: *The Journal of the Acoustical Society of America* 152.3 (2022), pp. 1586–1594. DOI: 10.1121/10.0013991.

[6] M. Akeroyd, L. Hickson, I. Holube, G. Keidser, G. Naylor, and K. Smeds. "Evaluating the Ecological Validity of Spatial Hearing". In: *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*. Turin, Italy: European Acoustics Association, 2023, pp. 411–414. ISBN: 978-88-88942-67-4. DOI: 10.61782/fa.2023.0457.

[7] D. Alais and D. Burr. "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration". In: *Current Biology* 14.3 (2004), pp. 257–262. DOI: 10.1016/j.cub.2004.01.029.

[8] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown. "A Corpus of Audio-Visual Lombard Speech with Frontal and Profile Views". In: *The Journal of the Acoustical Society of America* 143.6 (2018), EL523–EL529. DOI: 10.1121/1.5042758.

[9] S. Alhanbali, P. Dawes, S. Lloyd, and K. J. Munro. "Self-Reported Listening-Related Effort and Fatigue in Hearing-Impaired Adults". In: *Ear and Hearing* 38.1 (2017), e39. DOI: 10.1097/AUD.0000000000000361.

[10] E. Almadori, S. Mastroberardino, F. Botta, R. Brunetti, J. Lupiáñez, C. Spence, and V. Santangelo. "Crossmodal Semantic Congruence Interacts with Object Contextual Consistency in Complex Visual Scenes to Enhance Short-Term Memory Performance". In: *Brain Sciences* 11.9 (2021), p. 1206. DOI: 10.3390/brainsci11091206.

[11] ANSI. *American National Standard Acoustical Terminology, ANSI S1.1-2013*. New York, NY, USA, 2013.

[12] Arbeitsschutzausschüsse BMAS. *ASR A3.7 Lärm: Technische Regel für Arbeitsstätten*. 2021.

[13] I. Arweiler and J. M. Buchholz. "The Influence of Spectral Characteristics of Early Reflections on Speech Intelligibility". In: *The Journal of the Acoustical Society of America* 130.2 (2011), pp. 996–1005. DOI: 10.1121/1.3609258.

[14] I. Arweiler, J. M. Buchholz, and T. Dau. "The Influence of Masker Type on Early Reflection Processing and Speech Intelligibility (L)". In: *The Journal of the Acoustical Society of America* 133.1 (2013), pp. 13–16. DOI: 10.1121/1.4770249.

[15] L. Aspöck, F. Brinkmann, D. Ackermann, S. Weinzierl, and M. Vorländer. *BRAS - Benchmark for Room Acoustical Simulation*. 2020. DOI: 10.14279/depositonce-6726.3.

[16] L. Aspöck, A. Colsman, M. Kohnen, and M. Vorländer. "Investigating the Immersion of Reproduction Techniques for Room Auralizations". In: *Fortschritte Der Akustik - DAGA 2016*. Berlin, Germany: Deutsche Gesellschaft für Akustik e.V. (DEGA), 2016.

[17] T. O. Atienzar, L. K. Pilgrim, U. N. Sio, and J. E. Marsh. "Replicating and Extending Hemispheric Asymmetries in Auditory Distraction: No Metacognitive Awareness for the Left-Ear Disadvantage for Changing-State Sounds". In: *Journal of Cognitive Psychology* (2024), pp. 1–21. DOI: 10.1080/20445911.2024.2319268.

[18] R. Atkinson and R. Shiffrin. "Human Memory: A Proposed System and Its Control Processes". In: *Psychology of Learning and Motivation*. Vol. 2. Elsevier, 1968, pp. 89–195. ISBN: 978-0-12-543302-0. DOI: 10.1016/S0079-7421(08)60422-3.

[19] E. Au, S. Xiao, C. T. J. Hui, Y. Hioka, H. Masuda, and C. I. Watson. "Speech Intelligibility in Noise with Varying Spatial Acoustics under Ambisonics-based Sound Reproduction System". In: *Applied Acoustics* 174 (2021), p. 107707. DOI: 10.1016/j.apacoust.2020.107707.

[20] J. Badajoz-Davila and J. M. Buchholz. "Effect of Test Realism on Speech-in-noise Outcomes in Bilateral Cochlear Implant Users". In: *Ear and Hearing* 42.6 (2021), p. 1687. DOI: 10.1097/AUD.0000000000001061.

[21] A. D. Baddeley. "Is Working Memory Still Working?" In: *The American Psychologist* 56.11 (2001), pp. 851–864. DOI: 10.1037/0003-066x.56.11.851.

[22] A. Baddeley. "Working Memory and Language: An Overview". In: *Journal of Communication Disorders*. ASHA 2002 36.3 (2003), pp. 189–208. DOI: 10.1016/S0021-9924(03)00019-4.

[23] A. Baddeley. "Working Memory". In: *Current Biology* 20.4 (2010), R136–R140. DOI: 10.1016/j.cub.2009.12.014.

[24] A. Baddeley, G. Hitch, and R. Allen. "A Multicomponent Model of Working Memory". In: *Working Memory: State of the Science*. New York, NY, US: Oxford

University Press, 2021, pp. 10–43. ISBN: 978-0-19-884228-6. DOI: 10.1093/oso/9780198842286.003.0002.

[25] A. D. Baddeley and G. Hitch. "Working Memory". In: *Psychology of Learning and Motivation*. Ed. by G. H. Bower. Vol. 8. Academic Press, 1974, pp. 47–89. DOI: 10.1016/S0079-7421(08)60452-1.

[26] S. P. Banbury, W. J. Macken, S. Tremblay, and D. M. Jones. "Auditory Distraction and Short-Term Memory: Phenomena and Practical Implications". In: *Human Factors* 43.1 (2001), pp. 12–29. DOI: 10.1518/001872001775992462.

[27] L. Barbisan, M. Levorato, and F. Riente. *DynamicSound Simulator for Simulating Moving Sources and Microphone Arrays*. 2026. DOI: 10.48550/arXiv.2601.15433.

[28] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily. "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal". In: *Journal of Memory and Language* 68.3 (2013), pp. 255–278. DOI: 10.1016/j.jml.2012.11.001.

[29] J. Baumeister, S. Y. Ssin, N. A. M. ElSayed, J. Dorrian, D. P. Webb, J. A. Walsh, T. M. Simon, A. Irlitti, R. T. Smith, M. Kohler, and B. H. Thomas. "Cognitive Cost of Using Augmented Reality Displays". In: *IEEE transactions on visualization and computer graphics* 23.11 (2017), pp. 2378–2388. DOI: 10.1109/TVCG.2017.2735098.

[30] C. P. Beaman. "Auditory Distraction from Low-Intensity Noise: A Review of the Consequences for Learning and Workplace Environments". In: *Applied Cognitive Psychology* 19.8 (2005), pp. 1041–1064. DOI: 10.1002/acp.1134.

[31] T. Beechey. "Ecological Validity, External Validity, and Mundane Realism in Hearing Science". In: *Ear & Hearing* 43.5 (2022), pp. 1395–1401. DOI: 10.1097/AUD.0000000000001202.

[32] T. Beechey, J. M. Buchholz, and G. Keidser. "Hearing Aid Amplification Reduces Communication Effort of People With Hearing Impairment and Their Conversation Partners". In: *Journal of Speech, Language, and Hearing Research* 63.4 (2020), pp. 1299–1311. DOI: 10.1044/2020_JSLHR-19-00350.

[33] T. Beechey, J. M. Buchholz, and G. Keidser. "Hearing Impairment Increases Communication Effort During Conversations in Noise". In: *Journal of Speech, Language, and Hearing Research* 63.1 (2020), pp. 305–320. DOI: 10.1044/2019_JSLHR-19-00201.

[34] T. Beechey and G. Naylor. "How Purposeful Adaptive Responses to Adverse Conditions Facilitate Successful Auditory Functioning: A Conceptual Model". In: *Trends in Hearing* 29 (2025), p. 23312165251317010. DOI: 10.1177/23312165251317010.

[35] R. Bell, L. Mieth, J. P. Röer, and A. Buchner. "Auditory Distraction in the Item-color Binding Task: Support for a General Object-based Binding Account of the Changing-state Effect". In: *Auditory Perception & Cognition* 4.3-4 (2021), pp. 165–185. DOI: 10.1080/25742442.2022.2027210.

[36] R. Bell, L. Mieth, J. P. Röer, and A. Buchner. "The Metacognition of Auditory Distraction: Judgments about the Effects of Deviating and Changing Auditory Distractors on Cognitive Performance". In: *Memory & Cognition* 50.1 (2022), pp. 160–173. DOI: 10.3758/s13421-021-01200-2.

[37] R. Bell, J. P. Röer, A.-G. Lang, and A. Buchner. "Distraction by steady-state sounds: Evidence for a graded attentional model of auditory distraction". In: *Journal of Experimental Psychology: Human Perception and Performance* 45.4 (2019), pp. 500–512. DOI: 10.1037/xhp0000623.

[38] A. Berger and M. Kiefer. "Comparison of Different Response Time Outlier Exclusion Methods: A Simulation Study". In: *Frontiers in Psychology* 12 (2021), p. 675558. DOI: 10.3389/fpsyg.2021.675558.

[39] I. Bergström, S. Azevedo, P. Papiotis, N. Saldanha, and M. Slater. "The Plausibility of a String Quartet Performance in Virtual Reality". In: *IEEE Transactions on Visualization and Computer Graphics* 23.4 (2017), pp. 1352–1359. DOI: 10.1109/TVCG.2017.2657138.

[40] C. Bernarding, D. J. Strauss, R. Hannemann, H. Seidler, and F. I. Corona-Strauss. "Objective Assessment of Listening Effort in the Oscillatory EEG: Comparison of Different Hearing Aid Configurations". In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014, pp. 2653–2656. DOI: 10.1109/EMBC.2014.6944168.

[41] M. Berzborn, R. Bomhardt, J. Klein, J.-G. Richter, and M. Vorländer. "The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing". In: *Proceedings of the 43th Annual German Congress on Acoustics, Kiel, Germany*. Vol. 6. 2017.

[42] V. Best, G. Keidser, J. M. Buchholz, and K. Freeston. "An Examination of Speech Reception Thresholds Measured in a Simulated Reverberant Cafeteria Environment". In: *International Journal of Audiology* 54.10 (2015), pp. 682–690. DOI: 10.3109/14992027.2015.1028656.

[43] V. Best, E. J. Ozmeral, N. Kopčo, and B. G. Shinn-Cunningham. "Object Continuity Enhances Selective Auditory Attention". In: *Proceedings of the National Academy of Sciences* 105.35 (2008), pp. 13174–13178. DOI: 10.1073/pnas.0803718105.

[44] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. Revised Edition. Cambridge, Mass. and London: The MIT Press, 1997. ISBN: 978-0-262-02413-6.

[45] A. Bönsch. "Social Wayfinding Strategies to Explore Immersive Virtual Environments". Aachen, Germany: RWTH Aachen University, 2024. DOI: 10.18154/RWTH-2024-07063.

[46] A. Bönsch, S. Radke, H. Overath, L. M. Asche, J. Wendt, T. Vierjahn, U. Habel, and T. W. Kuhlen. "Social VR: How Personal Space Is Affected by Virtual Agents' Emotions". In: *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR 2018)*. Piscataway, NJ: IEEE, 2018, pp. 199–206. ISBN: 978-1-5386-3365-6. DOI: 10.1109/VR.2018.8446480.

[47] A. Bönsch, T. Vierjahn, and T. W. Kuhlen. "Evaluation of Approaching-Strategies of Temporarily Required Virtual Assistants in Immersive Environments". In: *2017 IEEE Symposium on 3D User Interfaces (3DUI 2017)*. Piscataway, NJ: IEEE, 2017, pp. 69–72. ISBN: 978-1-5090-6716-9. DOI: 10.1109/3DUI.2017.7893319.

[48] J. Bortz and N. Döring. *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler*. 4., überarb. Aufl. Springer-Lehrbuch. Heidelberg: Springer-Medizin-Verlag, 2006. ISBN: 978-3-540-33305-0.

[49] F. Botta, V. Santangelo, A. Raffone, D. Sanabria, J. Lupiáñez, and M. O. Belardinelli. "Multisensory Integration Affects Visuo-Spatial Working Memory." In: *Journal of Experimental Psychology: Human Perception and Performance* 37.4 (2011), pp. 1099–1109. DOI: 10.1037/a0023513.

[50] W. P. Boyce, A. Lindsay, A. Zgonnikov, I. Rañó, and K. Wong-Lin. "Optimality and Limitations of Audio-Visual Integration for Cognitive Systems". In: *Frontiers in Robotics and AI* 7 (2020), p. 516605. DOI: 10.3389/frobt.2020.00094.

[51] J. S. Bradley, R. Reich, and S. G. Norcross. "A Just Noticeable Difference in $C50$ for Speech". In: *Applied Acoustics* 58.2 (1999), pp. 99–108. DOI: 10.1016/S0003-682X(98)00075-9.

[52] T. Brand. "Speech Intelligibility". In: *Handbook of Signal Processing in Acoustics*. Ed. by D. Havelock, S. Kuwano, and M. Vorländer. New York, NY: Springer, 2008, pp. 197–204. ISBN: 978-0-387-30441-0. DOI: 10.1007/978-0-387-30441-0_13.

[53] H. S. Braren and J. Fels. *A High-Resolution Head-Related Transfer Function Data Set and 3D-Scan of KEMAR*. 2020. DOI: 10.18154/RWTH-2020-11307.

[54] A. S. Bregman. *Auditory Scene Analysis*. The MIT Press, 1990. ISBN: 0-262-02297-4.

[55] A. S. Bregman. "Auditory Scene Analysis: Hearing in Complex Environments". In: *Thinking in Sound: The Cognitive Psychology of Human Audition*. Ed. by S. McAdams and E. Bigand. Oxford University Press, 1993. ISBN: 978-0-19-852257-7. DOI: 10.1093/acprof:oso/9780198522577.003.0002.

[56] N. E. Breslow and X. Lin. "Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion". In: *Biometrika* 82.1 (1995), pp. 81–91. DOI: 10.1093/biomet/82.1.81.

[57] C. Breuer and J. Fels. "Exploring Auditory Selective Attention Shifts in Virtual Reality: An Approach with Matrix Sentences". In: *The Journal of the Acoustical Society of America* 158.5 (2025), pp. 3805–3813. DOI: 10.1121/10.0039864.

[58] C. Breuer, L. Leist, S. Fremerey, A. Raake, M. Klatte, and J. Fels. *Towards Investigating Listening Comprehension in Virtual Reality*. European Acoustics Association, 2023. DOI: 10.61782/fa.2023.0259.

[59] C. Breuer, L. Leist, S. Fremerey, A. Raake, M. Klatte, and J. Fels. "The influence of reverberation time on auditory selective attention switching in a virtual classroom scene". In: *The Journal of the Acoustical Society of America* 160.1 (2026), pp. 333–344. DOI: 10.1121/10.0044270.

[60] C. Breuer, K. Loh, L. Leist, S. Fremerey, A. Raake, M. Klatte, and J. Fels. "Examining the Auditory Selective Attention Switch in a Child-Suited Virtual Reality Classroom Environment". In: *International Journal of Environmental Research and Public Health* 19.24 (2022), p. 16569. DOI: 10.3390/ijerph192416569.

[61] C. Breuer, R. J. Schmitt, L. Leist, S. Fremerey, A. Raake, M. Klatte, and J. Fels. "The Influence of Complex Classroom Noise on Auditory Selective Attention". In: *Scientific Reports* 15.1 (2025), p. 32926. DOI: 10.1038/s41598-025-18232-2.

[62] C. Breuer, M. Tillmann, and J. Fels. *Complex Noise Scene Generator - CNSG*. Zenodo, 2025. DOI: 10.5281/zenodo.17105834.

[63] C. Breuer, L. J. Vollmer, L. Leist, S. Fremerey, A. Raake, M. Klatte, and J. Fels. "Exploring Cross-Modal Perception in a Virtual Classroom: The Effect of Visual Stimuli on Auditory Selective Attention". In: *Frontiers in Psychology* 16 (2025). DOI: 10.3389/fpsyg.2025.1512851.

[64] F. Brinkmann, A. Lindau, and S. Weinzierl. "On the Authenticity of Individual Dynamic Binaural Synthesis". In: *The Journal of the Acoustical Society of America* 142.4 (2017), pp. 1784–1795. DOI: 10.1121/1.5005606.

[65] D. E. Broadbent. *Perception and Communication*. Perception and Communication. Elmsford, NY, US: Pergamon Press, 1958, pp. v, 340. DOI: 10.1037/10037-000.

[66] A. W. Bronkhorst. "The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions". In: *Acta Acustica united with Acustica* 86.1 (2000), pp. 117–128.

[67] A. W. Bronkhorst. "The Cocktail-Party Problem Revisited: Early Processing and Selection of Multi-Talker Speech". In: *Attention, Perception, & Psychophysics* 77.5 (2015), pp. 1465–1487. DOI: 10.3758/s13414-015-0882-9.

[68] A. W. Bronkhorst and R. Plomp. "A Clinical Test for the Assessment of Binaural Speech Perception in Noise". In: *Audiology* 29.5 (1990), pp. 275–285. DOI: 10.3109/00206099009072858.

[69] M. E. Brooks, K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. Skaug J., M. Mächler, and B. Bolker M. "glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling". In: *The R Journal* 9.2 (2017), p. 378. DOI: 10.32614/RJ-2017-066.

[70] V. A. Brown. "An Introduction to Linear Mixed-Effects Modeling in R". In: *Advances in Methods and Practices in Psychological Science* 4.1 (2021). DOI: 10.1177/2515245920960351.

[71] V. A. Brown and J. F. Strand. ""Paying" Attention to Audiovisual Speech: Do Incongruent Stimuli Incur Greater Costs?" In: *Attention, Perception, & Psychophysics* 81.6 (2019), pp. 1743–1756. DOI: 10.3758/s13414-019-01772-x.

[72] V. A. Brown, K. J. Van Engen, and J. E. Peelle. "Face Mask Type Affects Audiovisual Speech Intelligibility and Subjective Listening Effort in Young and

Older Adults". In: *Cognitive Research: Principles and Implications* 6.1 (2021), p. 49. DOI: 10.1186/s41235-021-00314-0.

[73] D. Brungart, N. Iyer, E. R. Thompson, B. D. Simpson, S. Gordon-Salant, J. Schurman, C. Vogel, and K. Grant. "Interactions between Listening Effort and Masker Type on the Energetic and Informational Masking of Speech Stimuli". In: *Proceedings of Meetings on Acoustics* 19.1 (2013), p. 060146. DOI: 10.1121/1.4800033.

[74] D. S. Brungart. "Informational and Energetic Masking Effects in the Perception of Two Simultaneous Talkers". In: *The Journal of the Acoustical Society of America* 109.3 (2001), pp. 1101–1109. DOI: 10.1121/1.1345696.

[75] E. Brunswik. "Organismic Achievement and Environmental Probability". In: *Psychological Review* 50.3 (1943), pp. 255–272. DOI: 10.1037/h0060889.

[76] E. Brunswik. "Representative Design and Probabilistic Theory in a Functional Psychology". In: *Psychological Review* 62.3 (1955), pp. 193–217. DOI: 10.1037/h0047470.

[77] J. M. Buchholz, C. Davis, J. Beadle, and J. Kim. "Developing a Real-Time Test to Investigate Conversational Speech Understanding". In: *Journal of Speech, Language, and Hearing Research* 65.12 (2022), pp. 4520–4538. DOI: 10.1044/2022_JSLHR-22-00218.

[78] J. Buchholz, P. Derleth, K. Miles, and R. Ibrahim. "Visual Speech Benefit Provided by Realistic Sentences in Noise". In: 1. AUDICTIVE Conference. RWTH-2023-08867. RWTH Aachen University, Institute for Hearing Technology and Acoustics, 2023. DOI: 10.18154/RWTH-2023-08867.

[79] P.-C. Bürkner. "Brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80 (2017), pp. 1–28. DOI: 10.18637/jss.v080.i01.

[80] R. Campbell, J. Garwood, and S. Rosen. "Adding Sound to Lipread Lists: The Effects on Serial Recall of Adding an Auditory Pulse Train and a Pure Tone to Silently Lipread Lists". In: *Memory & Cognition* 16.3 (1988), pp. 210–219. DOI: 10.3758/BF03197754.

[81] T. Campbell, C. P. Beaman, and D. C. Berry. "Auditory Memory and the Irrelevant Sound Effect: Further Evidence for Changing-State Disruption". In: *Memory* 10.3 (2002), pp. 199–214. DOI: 10.1080/09658210143000335.

[82] J. L. Campos, D. McCumber, B. Chapnik, G. Singh, S.-T. Lau, K. Z. H. Li, V. Nieborowska, and K. Pichora-Fuller. "Perspectives On How Acoustical, Non-Acoustical, And User Characteristics Should Be Considered In Multimodal Virtual Reality Research And Application". In: *Canadian Acoustics* 46.3 (2018), pp. 29–42.

[83] J. Cassell. "Embodied Conversational Interface Agents". In: *Commun. ACM* 43.4 (2000), pp. 70–78. DOI: 10.1145/332051.332075.

[84] L. Chen and J. Vroomen. "Intersensory Binding across Space and Time: A Tutorial Review". In: *Attention, Perception, & Psychophysics* 75.5 (2013), pp. 790–811. DOI: 10.3758/s13414-013-0475-4.

[85] E. C. Cherry. "Some Experiments on the Recognition of Speech, with One and with Two Ears". In: *Journal of the Acoustical Society of America* 25 (1953), pp. 975–979. DOI: 10.1121/1.1907229.

[86] W. Choi, L. Li, S. Satoh, and K. Hachimura. "Multisensory Integration in the Virtual Hand Illusion with Active Movement". In: *BioMed Research International* 2016 (2016), e8163098. DOI: 10.1155/2016/8163098.

[87] G. Chollet, A. Esposito, A. Gentes, P. Horain, W. Karam, Z. Li, C. Pelachaud, P. Perrot, D. Petrovska-Delacrétaz, D. Zhou, and L. Zouari. "Multimodal Human Machine Interactions in Virtual and Augmented Reality". In: *Multimodal Signals: Cognitive and Algorithmic Issues*. Vol. 5398. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–23. ISBN: 978-3-642-00524-4. DOI: 10.1007/978-3-642-00525-1_1.

[88] C. Chubala, A. M. Surprenant, I. Neath, and P. T. Quinlan. "Does Dynamic Visual Noise Eliminate the Concreteness Effect in Working Memory?" In: *Journal of Memory and Language* 102 (2018), pp. 97–114. DOI: 10.1016/j.jml.2018.05.009.

[89] H. A. Colle and A. Welsh. "Acoustic Masking in Primary Memory". In: *Journal of Verbal Learning and Verbal Behavior* 15.1 (1976), pp. 17–31. DOI: 10.1016/S0022-5371(76)90003-7.

[90] B. Collin and M. Lavandier. "Binaural Speech Intelligibility in Rooms with Variations in Spatial Location of Sources and Modulation Depth of Noise Interferers". In: *The Journal of the Acoustical Society of America* 134.2 (2013), pp. 1146–1159. DOI: 10.1121/1.4812248.

[91] A. Colsman, L. Aspöck, M. Kohnen, and M. Vorländer. "Development of a Questionnaire to Investigate Immersion of Virtual Acoustic Environments". In: *Fortschritte Der Akustik - DAGA 2016*. Berlin, Germany: Deutsche Gesellschaft für Akustik e.V. (DEGA), 2016.

[92] A. R. A. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle. "Working Memory Span Tasks: A Methodological Review and User's Guide". In: *Psychonomic Bulletin & Review* 12.5 (2005), pp. 769–786. DOI: 10.3758/BF03196772.

[93] M. Cooke, M. L. Garcia Lecumberri, and J. Barker. "The Foreign Language Cocktail Party Problem: Energetic and Informational Masking Effects in Non-Native Speech Perception". In: *The Journal of the Acoustical Society of America* 123.1 (2008), pp. 414–427. DOI: 10.1121/1.2804952.

[94] M. Cooke, C. Mayo, and J. Villegas. "The Contribution of Durational and Spectral Changes to the Lombard Speech Intelligibility Benefit". In: *The Journal of the Acoustical Society of America* 135.2 (2014), pp. 874–883. DOI: 10.1121/1.4861342.

[95] J. K. Cooper, J. Vanthornhout, A. van Wieringen, and T. Francart. "Objectively Measuring Audiovisual Effects in Noise Using Virtual Human Speakers". In: *Trends in Hearing* 29 (2025). DOI: 10.1177/23312165251333528.

[96] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti. "Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE". In: *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*. 1993, pp. 135–142.

[97] J. F. Culling and M. A. Stone. "Energetic Masking and Masking Release". In: *The Auditory System at the Cocktail Party*. Ed. by J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay. Cham: Springer International Publishing, 2017, pp. 41–73. ISBN: 978-3-319-51662-2. DOI: 10.1007/978-3-319-51662-2_3.

[98] J. J. Cummings and J. N. Bailenson. "How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence". In: *Media Psychology* 19.2 (2016), pp. 272–309. DOI: 10.1080/15213269.2015.1015740.

[99] M. Daeglau, J. Otten, G. Grimm, B. Mirkovic, V. Hohmann, and S. Debener. "Neural Speech Tracking in a Virtual Acoustic Environment: Audio-Visual Benefit for Unscripted Continuous Speech". In: *Frontiers in Human Neuroscience* 19 (2025). DOI: 10.3389/fnhum.2025.1560558.

[100] F. Delogu and P. McMurray. "Where Did That Noise Come from? Memory for Sound Locations Is Exceedingly Eccentric Both in Front and in Rear Space". In: *Cognitive Processing* 20.4 (2019), pp. 479–494. DOI: 10.1007/s10339-019-00922-1.

[101] F. Delogu, T. C. Nijboer, and A. Postma. "Encoding Location and Serial Order in Auditory Working Memory: Evidence for Separable Processes". In: *Cognitive Processing* 13.3 (2012), pp. 267–276. DOI: 10.1007/s10339-012-0442-3.

[102] F. Delogu, A. Raffone, and M. O. Belardinelli. "Semantic Encoding in Working Memory: Is There a (Multi)Modality Effect?" In: *Memory* 17.6 (2009), pp. 655–663. DOI: 10.1080/09658210902998054.

[103] J. L. Desjardins and K. A. Doherty. "Age-Related Changes in Listening Effort for Various Types of Masker Noises". In: *Ear and Hearing* 34.3 (2013), p. 261. DOI: 10.1097/AUD.0b013e31826d0ba4.

[104] A. Devesse, A. Dudek, A. van Wieringen, and J. Wouters. "Speech Intelligibility of Virtual Humans". In: *International Journal of Audiology* 57.12 (2018), pp. 914–922. DOI: 10.1080/14992027.2018.1511922.

[105] A. Devesse, A. van Wieringen, and J. Wouters. "AVATAR Assesses Speech Understanding and Multitask Costs in Ecologically Relevant Listening Situations". In: *Ear and Hearing* 41.3 (2020), p. 521. DOI: 10.1097/AUD.0000000000000778.

[106] A. Dimitrijevic, M. L. Smith, D. S. Kadis, and D. R. Moore. "Neural Indices of Listening Effort in Noisy Environments". In: *Scientific Reports* 9.1 (2019), p. 11278. DOI: 10.1038/s41598-019-47643-1.

[107] DIN 18041. *Acoustic Quality in Rooms - Specifications and Instructions for the Room Acoustic Design*. DIN Media GmbH, 2016. DOI: 10.31030/2395845.

[108] W. Divin, K. Coyle, and D. T. T. James. "The Effects of Irrelevant Speech and Articulatory Suppression on the Serial Recall of Silently Presented Lipread Digits". In: *British Journal of Psychology* 92.4 (2001), pp. 593–616. DOI: 10.1348/000712601162365.

[109] R. Doggett, E. J. Sander, J. Birt, M. Ottley, and O. Baumann. "Using Virtual Reality to Evaluate the Impact of Room Acoustics on Cognitive Performance and Well-Being". In: *Frontiers in Virtual Reality* 2 (2021), p. 620503. DOI: 10.3389/frvir.2021.620503.

[110] N. I. Durlach, C. R. Mason, B. G. Shinn-Cunningham, T. L. Arbogast, H. S. Colburn, and G. Kidd Jr. "Informational Masking: Counteracting the Effects of Stimulus Uncertainty by Decreasing Target-Masker Similarity". In: *The Journal of the Acoustical Society of America* 114.1 (2003), pp. 368–379. DOI: 10.1121/1.1577562.

[111] M. Eckert, T. Robotham, E. A. P. Habets, and O. S. Rummukainen. "Pupillary Light Reflex Correction for Robust Pupillometry in Virtual Reality". In: *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5.2 (2022), pp. 1–16. DOI: 10.1145/3530798.

[112] 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: *ACM Transactions on Applied Perception* 18.4 (2021), 21:1–21:15. DOI: 10.1145/3486580.

[113] J. Ehret, A. Bönsch, J. Fels, S. J. Schlittmeier, and T. W. Kuhlen. "StudyFramework: Comfortably Setting up and Conducting Factorial-Design Studies Using the Unreal Engine". In: *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2024, pp. 442–449. DOI: 10.1109/VRW62533.2024.00087.

[114] 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: *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. IVA '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1–8. ISBN: 978-1-4503-9994-4. DOI: 10.1145/3570945.3607312.

[115] 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: *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2024, pp. 61–67. DOI: 10.1109/VRW62533.2024.00017.

[116] 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: *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2025, pp. 748–752. DOI: 10.1109/VRW66409.2025.00151.

[117] W. Ellermeier and K. Zimmer. "The Psychoacoustics of the Irrelevant Sound Effect". In: *Acoustical Science and Technology* 35.1 (2014), pp. 10–16. DOI: 10.1250/ast.35.10.

[118] K. Enge, M. Frank, and R. Holdrich. "Listening Experiment on the Plausibility of Acoustic Modeling in Virtual Reality". In: *Fortschritte Der Akustik - DAGA 2020*. Berlin, Germany: Deutsche Gesellschaft für Akustik e.V. (DEGA), 2020, pp. 13–16.

[119] G. Enzner, Chr. Antweiler, and S. Spors. "Trends in Acquisition of Individual Head-Related Transfer Functions". In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Berlin, Heidelberg: Springer, 2013, pp. 57–92. ISBN: 978-3-642-37762-4. DOI: 10.1007/978-3-642-37762-4_3.

[120] Epic Games. *MetaHuman | High-Fidelity Digital Humans Made Easy*. MetaHuman. URL: https://www.metahuman.com/en-US.

[121] N. P. Erber. "Auditory, Visual, and Auditory-Visual Recognition of Consonants by Children with Normal and Impaired Hearing". In: *Journal of Speech and Hearing Research* 15.2 (1972), pp. 413–422. DOI: 10.1044/jshr.1502.413.

[122] C. A. Ermert, A. Bönsch, T. W. Kuhlen, and J. Fels. "The impact of coverbal visual cues on speech intelligibility and cognitive load in virtual reality environments". In: *The Journal of the Acoustical Society of America*. Vol. 157. 2025, A159. DOI: 10.1121/10.0037748.

[123] C. A. Ermert, J. Ehret, C. Mohanathasan, A. Bönsch, T. W. Kuhlen, S. J. Schlittmeier, and J. Fels. "Influence of (non-) intelligible background speech on memory and listening effort in conversational situations". In: *Proceedings of DAS/DAGA 2025*. Ed. by T. Dau and B. Epp. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2025, pp. 159–160. ISBN: 978-3-939296-23-2. DOI: 10.71568/dasdaga2025.145.

[124] C. A. Ermert, K. Loh, K. Baylan, K. W. Kühlem, A. Bönsch, T. W. Kuhlen, and J. Fels. "Listening Effort In Complex Audiovisual Scenes Under Varying Simulated Room Acoustic Conditions". In: *Trends in Hearing* 30 (2026). DOI: 10.1177/23312165261468224.

[125] C. A. Ermert, S. J. Schlittmeier, A. Bönsch, T. W. Kuhlen, and J. Fels. "Heard-Text Recall and Listening Effort under Irrelevant Speech and Pseudo-Speech in Virtual Reality". In: *Acta Acustica* 10 (2026), p. 7. ISSN: 2681-4617. DOI: 10.1051/aacus/2026003.

[126] C. A. Ermert, L. Vollmer, H. Dewi, S. J. Schlittmeier, and J. Fels. "Spatial-changing irrelevant speech in visuo-spatial and visuo-verbal serial recall". In: *tbd* (under Review).

[127] C. A. Ermert, M. Yadav, J. Ehret, C. Mohanathasan, A. Bönsch, T. W. Kuhlen, S. J. Schlittmeier, and J. Fels. "Audiovisual angle and voice incongruence do not affect audiovisual verbal short-term memory in virtual reality". In: *PLoS One* 20.8 (2025), e0330693. DOI: 10.1371/journal.pone.0330693.

[128] C. A. Ermert, M. Yadav, J. E. Marsh, S. J. Schlittmeier, T. W. Kuhlen, and J. Fels. "Serial recall in spatial acoustic environments: irrelevant sound effect and

spatial source alternations". In: *Scientific Reports* 15.1 (2025). Publisher: Nature Publishing Group, p. 32473. ISSN: 2045-2322. DOI: 10.1038/s41598-025-18592-9.

[129] C. A. Ermert, C. Mohanathasan, J. Ehret, S. J. Schlittmeier, T. Kuhlen, and J. Fels. *AuViST - An Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm*. RWTH-2023-05543. RWTH Aachen University, 2023. DOI: 10.18154/RWTH-2023-05543.

[130] C. A. Ermert, C. Mohanathasan, S. J. Schlittmeier, and J. Fels. "Examining the Influence of Varying Audio Reproduction Methods on Verbal Short-Term Memory". In: *Fortschritte der Akustik - DAGA 2022*. Ed. by P. Leistner. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2022. ISBN: 978-3-939296-20-1. DOI: 10.18154/RWTH-2022-04013.

[131] H. Ershaid, M. Lizarazu, D. McLaughlin, M. Cooke, O. Simantiraki, M. Koutsogiannaki, and M. Lallier. "Contributions of Listening Effort and Intelligibility to Cortical Tracking of Speech in Adverse Listening Conditions". In: *Cortex* 172 (2024), pp. 54–71. DOI: 10.1016/j.cortex.2023.11.018.

[132] European Broadcasting Union. *Loudness Normalisation and Permitted Maximum Level of Audio Signals*. 2011.

[133] S. Fichna, T. Biberger, B. U. Seeber, and S. D. Ewert. "Effect of Acoustic Scene Complexity and Visual Scene Representation on Auditory Perception in Virtual Audio-Visual Environments". In: *Immersive 3D Audio: From Archit. Automot., I3DA*. Institute of Electrical and Electronics Engineers Inc., 2021. ISBN: 978-1-6654-0998-8. DOI: 10.1109/I3DA48870.2021.9610916.

[134] S. Fichna, S. van de Par, and S. D. Ewert. "Evaluation of Virtual Acoustic Environments with Different Acoustic Level of Detail". In: *2023 Immersive and 3D Audio: From Architecture to Automotive (I3DA)*. 2023, pp. 1–6. DOI: 10.1109/I3DA57090.2023.10289496.

[135] E. Fintor, L. Aspöck, J. Fels, and S. 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". In: *International Journal of Audiology* 61.5 (2022), pp. 371–379. DOI: 10.1080/14992027.2021.1922765.

[136] J. T. Fleming, R. K. Maddox, and B. G. Shinn-Cunningham. "Spatial Alignment between Faces and Voices Improves Selective Attention to Audio-Visual Speech". In: *The Journal of the Acoustical Society of America* 150.4 (2021), pp. 3085–3100. DOI: 10.1121/10.0006415.

[137] J. T. Fleming, A. L. Noyce, and B. G. Shinn-Cunningham. "Audio-Visual Spatial Alignment Improves Integration in the Presence of a Competing Audio-Visual Stimulus". In: *Neuropsychologia* 146 (2020), p. 107530. DOI: 10.1016/j.neuropsychologia.2020.107530.

[138] D. Fogerty, A. Alghamdi, and W.-Y. Chan. "The Effect of Simulated Room Acoustic Parameters on the Intelligibility and Perceived Reverberation of Mono-

syllabic Words and Sentences". In: *The Journal of the Acoustical Society of America* 147.5 (2020), EL396–EL402. DOI: 10.1121/10.0001217.

[139] T. Francart, A. van Wieringen, and J. Wouters. "Comparison of Fluctuating Maskers for Speech Recognition Tests". In: *International Journal of Audiology* 50.1 (2011), pp. 2–13. DOI: 10.3109/14992027.2010.505582.

[140] A. L. Francis and J. Love. "Listening Effort: Are We Measuring Cognition or Affect, or Both?" In: *WIREs Cognitive Science* 11.1 (2020), e1514. DOI: 10.1002/wcs.1514.

[141] A. L. Francis, M. K. MacPherson, B. Chandrasekaran, and A. M. Alvar. "Autonomic Nervous System Responses During Perception of Masked Speech May Reflect Constructs Other than Subjective Listening Effort". In: *Frontiers in Psychology* 7 (2016), p. 170291. DOI: 10.3389/fpsyg.2016.00263.

[142] N. Frangenberg, K. Waldow, and A. Fuhrmann. "Investigating the Influence of Sound Source Visualization on the Ventriloquism Effect in an Auralized Virtual Reality Environment". In: *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). 2021, pp. 609–610. DOI: 10.1109/VRW52623.2021.00187.

[143] M. D. Franzen and K. L. Wilhelm. "Conceptual Foundations of Ecological Validity in Neuropsychological Assessment". In: *Ecological Validity of Neuropsychological Testing*. Ed. by Sbordone, R. J. and Long, C. J. Gr Press/St Lucie Press, Inc., 1996, pp. 91–112.

[144] S. Fraser, é. J.-P. Gagn, M. Alepins, and P. Dubois. "Evaluating the Effort Expended to Understand Speech in Noise Using a Dual-Task Paradigm: The Effects of Providing Visual Speech Cues". In: *Journal of Speech, Language, and Hearing Research* 53.1 (2010), pp. 18–33. DOI: 10.1044/1092-4388(2009/08-0140).

[145] S. Freitag, B. Weyers, and T. W. Kuhlen. "Interactive Exploration Assistance for Immersive Virtual Environments Based on Object Visibility and Viewpoint Quality". In: *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2018, pp. 355–362. DOI: 10.1109/VR.2018.8447553.

[146] S. Fremerey, C. Breuer, L. Leist, M. Klatte, J. Fels, and A. Raake. "Speaker-Story Mapping as a Method to Evaluate Audiovisual Scene Analysis in a Virtual Classroom Scenario". In: *Frontiers in Psychology* 16 (2025). DOI: 10.3389/fpsyg.2025.1520630.

[147] S. Fremerey, R. F. U. Zaman, T. Ashraf, R. R. R. Rao, S. Göring, and A. Raake. "Towards Evaluation of Immersion, Visual Comfort and Exploration Behaviour for Non-Stereoscopic and Stereoscopic 360° Videos". In: *2023 IEEE International Symposium on Multimedia (ISM)*. 2023, pp. 131–138. DOI: 10.1109/ISM59092.2023.00025.

[148] J. B. Frtusova, A. H. Winneke, and N. A. Phillips. "ERP Evidence That Auditory–Visual Speech Facilitates Working Memory in Younger and Older Adults." In: *Psychology and Aging* 28.2 (2013), pp. 481–494. DOI: 10.1037/a0031243.

[149] "Virtual Reality". In: *Encyclopedia of Multimedia*. Ed. by B. Furht. Boston, MA: Springer US, 2008, pp. 968–968. ISBN: 978-0-387-74724-8. DOI: 10.1007/978-0-387-78414-4_255.

[150] J.-P. Gagné, J. Besser, and U. Lemke. "Behavioral Assessment of Listening Effort Using a Dual-Task Paradigm: A Review". In: *Trends in Hearing* 21 (2017), p. 2331216516687287. DOI: 10.1177/2331216516687287.

[151] M. Garnier, L. Ménard, and B. Alexandre. "Hyper-Articulation in Lombard Speech: An Active Communicative Strategy to Enhance Visible Speech Cues?" In: *The Journal of the Acoustical Society of America* 144.2 (2018), pp. 1059–1074. DOI: 10.1121/1.5051321.

[152] S. Gatehouse and W. Noble. "The Speech, Spatial and Qualities of Hearing Scale (SSQ)". In: *International Journal of Audiology* 43.2 (2004), pp. 85–99. DOI: 10.1080/14992020400050014.

[153] M. Georgi, L. Leist, M. Klatte, and S. J. Schlittmeier. "Investigating the Disturbance Impact of Background Speech on Verbal and Visual-Spatial Short-Term Memory: On the Differential Contributions of Changing-State and Phonology to the Irrelevant Sound Effect". In: *Auditory Perception & Cognition* 6.1-2 (2023), pp. 52–71. DOI: 10.1080/25742442.2022.2127988.

[154] M. Gerken, J. Schütze, C. Kirsch, S. D. Ewert, B. U. Seeber, J. Heeren, T. Afghah, K. C. Wagener, B. Kollmeier, and A. Warzybok. "Perceptual Measures of Normal-Hearing and Hearing-Impaired Listeners across Defined Virtual Acoustic Scenes". In: *International Journal of Audiology* 0.0 (2026), pp. 1–33. DOI: 10.1080/14992027.2026.2615093.

[155] J. Gervain. "Development of Speech Perception". In: *Speech Perception*. Ed. by L. L. Holt, J. E. Peelle, A. B. Coffin, A. N. Popper, and R. R. Fay. Cham: Springer International Publishing, 2022, pp. 201–226. ISBN: 978-3-030-81542-4. DOI: 10.1007/978-3-030-81542-4_8.

[156] J. C. Gil-Carvajal, J. Cubick, S. Santurette, and T. Dau. "Spatial Hearing with Incongruent Visual or Auditory Room Cues". In: *Scientific Reports* 6.1 (2016), p. 37342. DOI: 10.1038/srep37342.

[157] D. Gilbert, D. Rupp, M. Krüger, J. Ehret, K. Helwig, T. Römer, V. Wolf, V. Reality, and I. V. Group. *RWTH VR Group Unreal Engine Toolkit*. Version UE5.3-2023.1-rc2. Zenodo, 2024. DOI: 10.5281/zenodo.10817754.

[158] S. D. Goldinger, D. B. Pisoni, and J. S. Logan. "On the Nature of Talker Variability Effects on Recall of Spoken Word Lists". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17.1 (1991), pp. 152–162. DOI: 10.1037/0278-7393.17.1.152.

[159] E. J. Golob and J. R. Mock. "Dynamics of Auditory Spatial Attention Gradients". In: *Cognition* 194 (2020), p. 104058. DOI: 10.1016/j.cognition.2019.104058.

[160] M. Gonzalez-Franco, A. Maselli, D. Florencio, N. Smolyanskiy, and Z. Zhang. "Concurrent Talking in Immersive Virtual Reality: On the Dominance of Visual Speech Cues". In: *Scientific Reports* 7.3817 (2017). DOI: 10.1038/s41598-017-04201-x.

[161] A. Gorini, C. S. Capideville, G. De Leo, F. Mantovani, and G. Riva. "The Role of Immersion and Narrative in Mediated Presence: The Virtual Hospital Experience". In: *Cyberpsychology, Behavior, and Social Networking* 14.3 (2011), pp. 99–105. DOI: 10.1089/cyber.2010.0100.

[162] P. A. Gosselin and J.-P. Gagné. "Older Adults Expend More Listening Effort than Young Adults Recognizing Audiovisual Speech in Noise". In: *International Journal of Audiology* 50.11 (2011), pp. 786–792. DOI: 10.3109/14992027.2011.599870.

[163] K. W. Grant, L. H. Ardell, P. K. Kuhl, and D. W. Sparks. "The Contribution of Fundamental Frequency, Amplitude Envelope, and Voicing Duration Cues to Speechreading in Normal-Hearing Subjects". In: *The Journal of the Acoustical Society of America* 77.2 (1985), pp. 671–677. DOI: 10.1121/1.392335.

[164] K. W. Grant. "The Effect of Speechreading on Masked Detection Thresholds for Filtered Speech". In: *The Journal of the Acoustical Society of America* 109.5 (2001), pp. 2272–2275. DOI: 10.1121/1.1362687.

[165] K. W. Grant and J. G. W. Bernstein. "Toward a Model of Auditory-Visual Speech Intelligibility". In: *Multisensory Processes*. Ed. by A. K. C. Lee, M. T. Wallace, A. B. Coffin, A. N. Popper, and R. R. Fay. Vol. 68. Cham: Springer International Publishing, 2019, pp. 33–57. ISBN: 978-3-030-10461-0. DOI: 10.1007/978-3-030-10461-0_3.

[166] P. Green and C. J. MacLeod. "Simr: An R Package for Power Analysis of Generalised Linear Mixed Models by Simulation". In: *Methods in Ecology and Evolution* 7.4 (2016), pp. 493–498. DOI: 10.1111/2041-210X.12504.

[167] D. Guitard and J. Saint-Aubin. "A Replication of "Functional Equivalence of Verbal and Spatial Information in Serial Short-Term Memory (1995; Experiments 2 and 3)"." In: *The Quantitative Methods for Psychology* 11.2 (2015), r4–r7. DOI: 10.20982/tqmp.11.2.r004.

[168] S. J. Gustafson, C. Ortiz, and L. Nelson. "Test–Retest Reliability and the Role of Task Instructions When Measuring Listening Effort Using a Verbal Response Time Paradigm". In: *Seminars in Hearing* 44.02 (2023), pp. 140–154. DOI: 10.1055/s-0043-1767667.

[169] M. A. A. Gutiérrez, F. Vexo, and D. Thalmann. *Stepping into Virtual Reality*. SpringerLink: Springer e-Books. London: Springer-Verlag London Limited, 2008. ISBN: 978-1-84800-117-6.

[170] A. Haapakangas, E. Kankkunen, V. Hongisto, P. Virjonen, D. Oliva, and E. Keskinen. "Effects of Five Speech Masking Sounds on Performance and Acoustic Satisfaction. Implications for Open-Plan Offices". In: *Acta Acustica united with Acustica* 97.4 (2011), pp. 641–655. DOI: 10.3813/AAA.918444.

[171] A. Haapakangas, V. Hongisto, and A. Liebl. "The Relation between the Intelligibility of Irrelevant Speech and Cognitive Performance—A Revised Model Based on Laboratory Studies". In: *Indoor Air* 30.6 (2020), pp. 1130–1146. DOI: 10.1111/ina.12726.

[172] L. V. Hadley, W. O. Brimijoin, and W. M. Whitmer. "Speech, Movement, and Gaze Behaviours during Dyadic Conversation in Noise". In: *Scientific Reports* 9.1 (2019), p. 10451. DOI: 10.1038/s41598-019-46416-0.

[173] E. T. Hall. *The Hidden Dimension: Man's Use of Space in Public and Private*. Bodley Head, 1969. 201 pp. ISBN: 978-0-370-01308-4.

[174] J. R. Hanley and C. Broadbent. "The Effect of Unattended Speech on Serial Recall Following Auditory Presentation". In: *British Journal of Psychology* 78.3 (1987), pp. 287–297. DOI: 10.1111/j.2044-8295.1987.tb02247.x.

[175] J. R. Hanley and A. Hayes. "The irrelevant sound effect under articulatory suppression: Is it a suffix effect?" In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38.2 (2012), pp. 482–487. DOI: 10.1037/a0025600.

[176] C. Harms and F. Biocca. "Internal Consistency and Reliability of the Networked Minds Measure of Social Presence". In: *Seventh Annual International Workshop: Presence 2004.* (2004).

[177] S. G. Hart and L. E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Advances in Psychology*. Vol. 52. Elsevier, 1988, pp. 139–183. ISBN: 978-0-444-70388-0. DOI: 10.1016/S0166-4115(08)62386-9.

[178] F. Hartig. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. Manual. 2022.

[179] A. J. Harvey and C. P. Beaman. "Input and Output Modality Effects in Immediate Serial Recall". In: *Memory* 15.7 (2007), pp. 693–700. DOI: 10.1080/09658210701644677.

[180] R. P. Hawkins. "Is Social Validity What We Are Interested in? Argument for a Functional Approach". In: *Journal of Applied Behavior Analysis* 24.2 (1991), pp. 205–213. DOI: 10.1901/jaba.1991.24-205.

[181] A. F. Healy. "Short-Term Memory for Order Information". In: *Psychology of Learning and Motivation*. Ed. by G. H. Bower. Vol. 16. Academic Press, 1982, pp. 191–238. DOI: 10.1016/S0079-7421(08)60550-2.

[182] E. Hedlund, L. Muller, and W. Kropp. "Influence of Speech Exposure on Working Memory: Exploring Reverberation Time in Open Plan Office Environments". In: Fortschritte Der Akustik - DAGA 2024. Berlin, Germany: Deutsche Gesellschaft für Akustik e.V. (DEGA), 2024.

[183] T. Heittola, A. Mesaros, and T. Virtanen. *TAU Urban Acoustic Scenes 2019, Development Dataset*. Zenodo, 2019. DOI: 10.5281/zenodo.2589280.

[184] K. S. Helfer and R. L. Freyman. "The Role of Visual Speech Cues in Reducing Energetic and Informational Masking". In: *The Journal of the Acoustical Society of America* 117.2 (2005), pp. 842–849. DOI: 10.1121/1.1836832.

[185] M. M. E. Hendrikse, G. Llorach, G. Grimm, and V. Hohmann. "Influence of Visual Cues on Head and Eye Movements during Listening Tasks in Multi-Talker Audiovisual Environments with Animated Characters". In: *Speech Communication* 101 (2018), pp. 70–84. DOI: 10.1016/j.specom.2018.05.008.

[186] M. M. E. Hendrikse, G. Llorach, V. Hohmann, and G. Grimm. "Movement and Gaze Behavior in Virtual Audiovisual Listening Environments Resembling Everyday Life". In: *Trends in Hearing* 23 (2019). DOI: 10.1177/2331216519872362.

[187] J. Henrich, S. J. Heine, and A. Norenzayan. "Most People Are Not WEIRD". In: *Nature* 466.7302 (2010), pp. 29–29. DOI: 10.1038/466029a.

[188] B. Herrmann and I. S. Johnsrude. "A Model of Listening Engagement (MoLE)". In: *Hearing Research* 397 (2020), p. 108016. DOI: 10.1016/j.heares.2020.108016.

[189] H. Himmelein, M. von Berg, C. Pörschmann, and J. Steffens. "Influence of Mental Effort on Sound Evaluations in Virtual and Real Experimental Environments". In: *Frontiers in Virtual Reality* 6 (2025). DOI: 10.3389/frvir.2025.1672595.

[190] Ľ. Hládek, S. D. Ewert, and B. U. Seeber. "Communication Conditions in Virtual Acoustic Scenes in an Underground Station". In: *2021 Immersive and 3D Audio: From Architecture to Automotive (I3DA)*. 2021, pp. 1–8. DOI: 10.1109/I3DA48870.2021.9610843.

[191] G. A. Holleman, I. T. C. Hooge, C. Kemner, and R. S. Hessels. "The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity". In: *Frontiers in Psychology* 11 (2020), p. 721. DOI: 10.3389/fpsyg.2020.00721.

[192] I. Holube, S. Taesler, S. Ibelings, M. Hansen, and J. Ooster. "Automated Measurement of Speech Recognition, Reaction Time, and Speech Rate and Their Relation to Self-Reported Listening Effort for Normal-Hearing and Hearing-Impaired Listeners Using Various Maskers". In: *Trends in Hearing* 28 (2024), p. 23312165241276435. DOI: 10.1177/23312165241276435.

[193] M. Hoppe, J. Karolus, F. Dietz, P. W. Woźniak, A. Schmidt, and T.-K. Machulla. "VRsneaky: Increasing Presence in VR Through Gait-Aware Auditory Feedback". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–9. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300776.

[194] B. W. Y. Hornsby. "The Effects of Hearing Aid Use on Listening Effort and Mental Fatigue Associated With Sustained Speech Processing Demands". In: *Ear and Hearing* 34.5 (2013), p. 523. DOI: 10.1097/AUD.0b013e31828003d8.

[195] I. P. Howard and W. B. Templeton. *Human Spatial Orientation*. Human Spatial Orientation. Oxford, England: John Wiley & Sons, 1966, p. 533. 533 pp.

[196] H. Huang, I. Oosthuizen, and E. M. Picou. "Dual-Task Paradigm Measures of Listening Effort: To Include or Not to Include Secondary Task Responses with Incorrect Primary Task Responses". In: *Seminars in Hearing* 44.2 (2023), pp. 155–165. DOI: 10.1055/s-0043-1766140.

[197] H. Huang, T. A. Ricketts, B. W. Y. Hornsby, and E. M. Picou. "Effects of Critical Distance and Reverberation on Listening Effort in Adults". In: *Journal of Speech, Language, and Hearing Research* 65.12 (2022), pp. 4837–4851. DOI: 10.1044/2022_JSLHR-22-00109.

[198] W. Huang, R. D. Roscoe, M. C. Johnson-Glenberg, and S. D. Craig. "Motivation, Engagement, and Performance across Multiple Virtual Reality Sessions and Levels of Immersion". In: *Journal of Computer Assisted Learning* 37.3 (2021), pp. 745–758. DOI: 10.1111/jcal.12520.

[199] R. W. Hughes. "Auditory Distraction: A Duplex-Mechanism Account". In: *PsyCh Journal* 3.1 (2014), pp. 30–41. DOI: 10.1002/pchj.44.

[200] R. W. Hughes, C. Chamberland, S. Tremblay, and D. M. Jones. "Perceptual-Motor Determinants of Auditory-Verbal Serial Short-Term Memory". In: *Journal of Memory and Language* 90 (2016), pp. 126–146. DOI: 10.1016/j.jml.2016.04.006.

[201] R. W. Hughes and J. E. Marsh. "The Functional Determinants of Short-Term Memory: Evidence from Perceptual-Motor Interference in Verbal Serial Recall". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43.4 (2017), pp. 537–551. DOI: 10.1037/xlm0000325.

[202] R. W. Hughes, J. E. Marsh, and D. M. Jones. "Perceptual–Gestural (Mis)Mapping in Serial Short-Term Memory: The Impact of Talker Variability". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.6 (2009), pp. 1411–1425. DOI: 10.1037/a0017008.

[203] R. W. Hughes, F. Vachon, and D. M. Jones. "Disruption of Short-Term Memory by Changing and Deviant Sounds: Support for a Duplex-Mechanism Account of Auditory Distraction". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33.6 (2007), pp. 1050–1061. DOI: 10.1037/0278-7393.33.6.1050.

[204] T. Huisman, A. Ahrens, and E. MacDonald. "Ambisonics Sound Source Localization With Varying Amount of Visual Information in Virtual Reality". In: *Frontiers in Virtual Reality* 2 (2021). DOI: 10.3389/frvir.2021.722321.

[205] T. Huisman, T. Dau, T. Piechowiak, E. MacDonald, T. Dau, T. Piechowiak, and E. MacDonald. "The Ventriloquist Effect Is Not Consistently Affected by Stimulus Realism". In: *Journal of Perceptual Imaging* 5 (2022), pp. 1–10. DOI: 10.2352/J.Percept.Imaging.2022.5.000404.

[206] M. J. Hurlstone. "Serial Recall". In: *The Oxford Handbook of Human Memory* (2022). DOI: 10.31219/osf.io/4w8fu.

[207] M. J. Hurlstone, G. J. Hitch, and A. D. Baddeley. "Memory for Serial Order across Domains: An Overview of the Literature and Directions for Future Research". In: *Psychological Bulletin* 140 (2014), pp. 339–373. DOI: 10.1037/a0034221.

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

[209] Institute for Hearing Technology and Acoustics, RWTH Aachen University, P. Schäfer, J. Stienen, L. Aspöck, and M. Vorländer. *Virtual Acoustics - A Real-*

Time Auralization Framework for Scientific Research. Zenodo, 2021. DOI: 10.5281/zenodo.13744493.

[210] International Organization for Standardization. *BS ISO 22955: Acoustics. Acoustic quality of open office spaces*. 2021.

[211] S. Irshad, A. E. Lai, G. H. Alene, A. Moraru, A. Krupicka, G. Kapellmann-Zafra, A. Perkis, O. Bruland, and C. A. Klöckner. "Beyond Presence: Understanding the Role of Identity and Emotions in Virtual Reality for Natural Hazards Preparedness". In: *Virtual Reality* 30.1 (2025), p. 28. DOI: 10.1007/s10055-025-01288-w.

[212] C. E. Jack and W. R. Thurlow. "Effects of Degree of Visual Association and Angle of Displacement on the "Ventriloquism" Effect". In: *Perceptual and Motor Skills* 37.3 (1973), pp. 967–979. DOI: 10.2466/pms.1973.37.3.967.

[213] A. B. Jagadeesh and A. K. Uppunda. "Speech-on-Speech Masking: Effect of Maskers with Different Degrees of Linguistic Information." In: *Canadian Journal of Speech-Language Pathology & Audiology* 45.2 (2021).

[214] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton. "Measuring and Defining the Experience of Immersion in Games". In: *International Journal of Human-Computer Studies* 66.9 (2008), pp. 641–661. DOI: 10.1016/j.ijhcs.2008.04.004.

[215] P. Jianxin and J. Peng. "The Effects of the Noise and Reverberation on the Working Memory Span of Children". In: *Archives of Acoustics* 43.1 (2018), pp. 123–128. DOI: 10.24425/118087.

[216] C. Jicol, C. Clarke, E. Tor, R. M. Dakin, T. C. Lancaster, S. T. Chang, K. Petrini, E. O'Neill, M. J. Proulx, and C. Lutteroth. "Realism and Field of View Affect Presence in VR but Not the Way You Think". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, 2023, pp. 1–17. ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3581448.

[217] S. Johnston. "Virtual Reality as New Media". In: *Encyclopedia of Computer Graphics and Games*. Ed. by N. Lee. Cham: Springer International Publishing, 2024, pp. 2013–2015. ISBN: 978-3-031-23159-9. DOI: 10.1007/978-3-031-23161-2_87.

[218] D. M. Jones, C. Madden, and C. Miles. "Privileged Access by Irrelevant Speech to Short-term Memory: The Role of Changing State". In: *The Quarterly Journal of Experimental Psychology Section A* 44.4 (1992), pp. 645–669. DOI: 10.1080/14640749208401304.

[219] D. Jones. "Objects, Streams, and Threads of Auditory Attention". In: *Attention: Selection, Awareness, and Control: A Tribute to Donald Broadbent*. New York, NY, US: Clarendon Press/Oxford University Press, 1993, pp. 87–104. ISBN: 978-0-19-852259-1.

[220] D. Jones, P. Farrand, G. Stuart, and N. Morris. "Functional Equivalence of Verbal and Spatial Information in Serial Short-Term Memory". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21 (1995), pp. 1008–1018. DOI: 10.1037/0278-7393.21.4.1008.

[221] D. M. Jones. "Disruption of Memory for Lip-Read Lists by Irrelevant Speech: Further Support for the Changing State Hypothesis". In: *The Quarterly Journal of Experimental Psychology Section A* 47.1 (1994), pp. 143–160. DOI: 10.1080/14640749408401147.

[222] D. M. Jones, P. Bcaman, and W. J. Macken. "The Object-Oriented Episodic Record Model". In: *Models Of Short-Term Memory*. Psychology Press, 1996.

[223] D. M. Jones, R. W. Hughes, and W. J. Macken. "Auditory Distraction and Serial Memory: The Avoidable and the Ineluctable". In: *Noise and Health* 12.49 (2010), p. 201. DOI: 10.4103/1463-1741.70497.

[224] D. M. Jones and W. J. Macken. "Irrelevant Tones Produce an Irrelevant Speech Effect: Implications for Phonological Coding in Working Memory". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19.2 (1993), pp. 369–381. DOI: 10.1037/0278-7393.19.2.369.

[225] D. M. Jones and W. J. Macken. "Auditory Babble and Cognitive Efficiency: Role of Number of Voices and Their Location". In: *Journal of Experimental Psychology: Applied* 1 (1995), pp. 216–226. DOI: 10.1037/1076-898X.1.3.216.

[226] D. M. Jones and W. J. Macken. "Organizational factors in the effect of irrelevant speech: The role of spatial location and timing". In: *Memory & Cognition* 23.2 (1995), pp. 192–200. DOI: 10.3758/BF03197221.

[227] D. M. Jones, W. J. Macken, and A. P. Nicholls. "The Phonological Store of Working Memory: Is It Phonological and Is It a Store?" In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30 (2004), pp. 656–674. DOI: 10.1037/0278-7393.30.3.656.

[228] D. M. Jones and S. Tremblay. "Interference in Memory by Process or Content? A Reply to Neath (2000)". In: *Psychonomic Bulletin & Review* 7.3 (2000), pp. 550–558. DOI: 10.3758/BF03214370.

[229] G. L. Jones and R. Y. Litovsky. "A Cocktail Party Model of Spatial Release from Masking by Both Noise and Speech Interferers". In: *The Journal of the Acoustical Society of America* 130.3 (2011), pp. 1463–1474. DOI: 10.1121/1.3613928.

[230] D. Kahneman. *Attention and Effort*. Prentice-Hall Series in Experimental Psychology. Englewood Cliffs and Hemel Hempstead: Prentice-Hall, 1973. ISBN: 0-13-050518-8.

[231] A. Kassambara. *Rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. Manual. 2023.

[232] G. Keidser, G. Naylor, D. S. Brungart, A. Caduff, J. Campos, S. Carlile, M. G. Carpenter, G. Grimm, V. Hohmann, I. Holube, S. Launer, T. Lunner, R. Mehra, F. Rapport, M. Slaney, and K. Smeds. "The Quest for Ecological Validity in Hearing Science: What It Is, Why It Matters, and How to Advance It". In: *Ear and Hearing* 41 (2020), 5S. DOI: 10.1097/AUD.0000000000000944.

[233] G. Kidd and H. S. Colburn. "Informational Masking in Speech Recognition". In: *The Auditory System at the Cocktail Party*. Ed. by J. C. Middlebrooks, J. Z.

Simon, A. N. Popper, and R. R. Fay. Cham: Springer International Publishing, 2017, pp. 75–109. ISBN: 978-3-319-51662-2. DOI: 10.1007/978-3-319-51662-2_4.

[234] G. Kidd Jr., C. R. Mason, J. Swaminathan, E. Roverud, K. K. Clayton, and V. Best. "Determining the Energetic and Informational Components of Speech-on-Speech Masking". In: *The Journal of the Acoustical Society of America* 140.1 (2016), pp. 132–144. DOI: 10.1121/1.4954748.

[235] H. Kim and I.-K. Lee. "Studying the Effects of Congruence of Auditory and Visual Stimuli on Virtual Reality Experiences". In: *IEEE Transactions on Visualization and Computer Graphics* 28.5 (2022), pp. 2080–2090. DOI: 10.1109/TVCG.2022.3150514.

[236] M. Klatte, J. Hellbrück, J. Seidel, and P. Leistner. "Effects of Classroom Acoustics on Performance and Well-Being in Elementary School Children: A Field Study". In: *Environment and Behavior* 42.5 (2010), pp. 659–692. DOI: 10.1177/0013916509336813.

[237] M. Klatte, T. Lachmann, and M. Meis. "Effects of Noise and Reverberation on Speech Perception and Listening Comprehension of Children and Adults in a Classroom-like Setting". In: *Noise and Health* 12.49 (2010), p. 270. DOI: 10.4103/1463-1741.70506.

[238] M. Klatte, N. Lee, and J. Hellbrück. "Effects of Irrelevant Speech and Articulatory Suppression on Serial Recall of Heard and Read Materials". In: *Psychologische Beiträge* 44.2 (2002), p. 166.

[239] T. Koelewijn, H. de Kluiver, B. G. Shinn-Cunningham, A. A. Zekveld, and S. E. Kramer. "The Pupil Response Reveals Increased Listening Effort When It Is Difficult to Focus Attention". In: *Hearing Research* 323 (2015), pp. 81–90. DOI: 10.1016/j.heares.2015.02.004.

[240] T. Koelewijn, A. A. Zekveld, J. M. Festen, and S. E. Kramer. "Pupil Dilation Uncovers Extra Listening Effort in the Presence of a Single-Talker Masker". In: *Ear and Hearing* 33.2 (2012), p. 291. DOI: 10.1097/AUD.0b013e3182310019.

[241] T. Koelewijn, A. A. Zekveld, J. M. Festen, and S. E. Kramer. "The Influence of Informational Masking on Speech Perception and Pupil Response in Adults with Hearing Impairment". In: *The Journal of the Acoustical Society of America* 135.3 (2014), pp. 1596–1606. DOI: 10.1121/1.4863198.

[242] M. Kohnen, B. S. Masiero, and M. Vorländer. "Combining Reproduction Techniques for Room Auralizations". RWTH Aachen University, 2024. DOI: 10.18154/RWTH-2025-01606.

[243] L. Kroczek, S. Roßkopf, F. Stärz, M. Blau, S. V. de Par, and A. Mühlberger. "The Influence of Affective Voice on Sound Distance Perception". In: *OSF* (2023). DOI: 10.31234/osf.io/qdw9p.

[244] M. Krueger, M. Schulte, T. Brand, and I. Holube. "Development of an Adaptive Scaling Method for Subjective Listening Effort". In: *The Journal of the Acoustical Society of America* 141.6 (2017), pp. 4680–4693. DOI: 10.1121/1.4986938.

[245] M. Krueger, M. Schulte, M. A. Zokoll, K. C. Wagener, M. Meis, T. Brand, and I. Holube. "Relation Between Listening Effort and Speech Intelligibility in Noise". In: *American Journal of Audiology* 26 (2017), pp. 378–392. DOI: 10.1044/2017_AJA-16-0136.

[246] J. K. Kruschke and T. M. Liddell. "The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective". In: *Psychonomic Bulletin & Review* 25.1 (2018), pp. 178–206. DOI: 10.3758/s13423-016-1221-4.

[247] S. E. Kuchinsky, J. B. Ahlstrom, K. I. Vaden JR., S. L. Cute, L. E. Humes, J. R. Dubno, and M. A. Eckert. "Pupil Size Varies with Word Listening and Response Selection Difficulty in Older Adults with Hearing Loss". In: *Psychophysiology* 50.1 (2013), pp. 23–34. DOI: 10.1111/j.1469-8986.2012.01477.x.

[248] S. E. Kuchinsky, F. J. Gallun, and A. K. C. Lee. "Note on the Dual-Task Paradigm and Its Use to Measure Listening Effort". In: *Trends in Hearing* 28 (2024), p. 23312165241292215. DOI: 10.1177/23312165241292215.

[249] T. W. Kuhlen and B. Hentschel. "Quo Vadis CAVE: Does Immersive Visualization Still Matter?" In: *IEEE Computer Graphics and Applications* 34.5 (2014), pp. 14–21. DOI: 10.1109/MCG.2014.97.

[250] C. Kuhn-Rahloff. *Realitätstreue, Natürlichkeit, Plausibilität: Perzeptive Beurteilungen in der Elektroakustik*. Berlin, Heidelberg: Springer, 2012. ISBN: 978-3-642-22071-5. DOI: 10.1007/978-3-642-22072-2.

[251] H. Kuttruff. *Room Acoustics*. Sixth edition. Boca Raton London New York: CRC Press, Taylor & Francis Group, 2017. 1 p. ISBN: 978-1-4822-6043-4.

[252] A. Kuusinen, K. Kondraciuk, and T. Lokki. "Effects of Masker Type and Reverberation on Speech-in-Noise Recognition Thresholds and Listening Effort as Indexed by Pupil Dilation Responses". In: *Journal of the Audio Engineering Society* 10650 (2023).

[253] T. Kvetnaya. "Registered Replication Report: Testing Disruptive Effects of Irrelevant Speech on Visual-Spatial Working Memory". In: *Journal of European Psychology Students* 9.1 (2018). DOI: 10.5334/jeps.450.

[254] J.-S. Kyong, C. Kwak, W. Han, M.-W. Suh, and J. Kim. "Effect of Speech Degradation and Listening Effort in Reverberating and Noisy Environments Given N400 Responses". In: *Journal of Audiology & Otology* 24.3 (2020), pp. 119–126. DOI: 10.7874/jao.2019.00514.

[255] E. B. Lange. "Disruption of Attention by Irrelevant Stimuli in Serial Recall". In: *Journal of Memory and Language* 53.4 (2005), pp. 513–531. DOI: 10.1016/j.jml.2005.07.002.

[256] P. Larm and V. Hongisto. "Experimental Comparison between Speech Transmission Index, Rapid Speech Transmission Index, and Speech Intelligibility Index". In: *The Journal of the Acoustical Society of America* 119.2 (2006), pp. 1106–1117. DOI: 10.1121/1.2146112.

[257] B. Larsby, M. Hällgren, and B. Lyxell. "The Interference of Different Background Noises on Speech Processing in Elderly Hearing Impaired Subjects". In: *International Journal of Audiology* 47 (2008), S83–S90. DOI: 10.1080/14992020802301159.

[258] D. C. LeCompte. "Irrelevant speech, serial rehearsal, and temporal distinctiveness: A new approach to the irrelevant speech effect". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22.5 (1996), pp. 1154–1165. DOI: 10.1037/0278-7393.22.5.1154.

[259] A. K. C. Lee, R. K. Maddox, and J. K. Bizley. "An Object-Based Interpretation of Audiovisual Processing". In: *Multisensory Processes*. Ed. by A. K. C. Lee, M. T. Wallace, A. B. Coffin, A. N. Popper, and R. R. Fay. Vol. 68. Cham: Springer International Publishing, 2019, pp. 59–83. ISBN: 978-3-030-10459-7. DOI: 10.1007/978-3-030-10461-0_4.

[260] A. K. C. Lee, M. T. Wallace, A. B. Coffin, A. N. Popper, R. R. Fay, and A. N. Popper, eds. *Multisensory Processes: The Auditory Perspective*. Corrected publication. Springer Handbook of Auditory Research Volume 68. Cham, Switzerland: Springer, 2020. 1 p. ISBN: 978-3-030-10461-0.

[261] G. Lehnert and H. D. Zimmer. "Auditory and Visual Spatial Working Memory". In: *Memory & Cognition* 34.5 (2006), pp. 1080–1090. DOI: 10.3758/BF03193254.

[262] G. Lehnert and H. D. Zimmer. "Modality and Domain Specific Components in Auditory and Visual Working Memory Tasks". In: *Cognitive Processing* 9.1 (2008), pp. 53–61. DOI: 10.1007/s10339-007-0187-6.

[263] D. Leiner. *SoSci Survey*. 2019. URL: https://www.soscisurvey.de/.

[264] L. Leist, C. Breuer, M. Yadav, S. Fremerey, J. Fels, A. Raake, T. Lachmann, S. J. Schlittmeier, and M. Klatte. "Differential Effects of Task-Irrelevant Monaural and Binaural Classroom Scenarios on Children's and Adults' Speech Perception, Listening Comprehension, and Visual–Verbal Short-Term Memory". In: *International Journal of Environmental Research and Public Health* 19.23 (2022), p. 15998. DOI: 10.3390/ijerph192315998.

[265] L. Leist, T. Lachmann, S. J. Schlittmeier, M. Georgi, and M. Klatte. "Irrelevant Speech Impairs Serial Recall of Verbal but Not Spatial Items in Children and Adults". In: *Memory & Cognition* (2022). DOI: 10.3758/s13421-022-01359-2.

[266] U. Lemke and J. Besser. "Cognitive Load and Listening Effort: Concepts and Age-Related Considerations". In: *Ear and Hearing* 37 (2016), 77S. DOI: 10.1097/AUD.0000000000000304.

[267] R. V. Lenth, B. Banfai, B. Bolker, P. Buerkner, I. Giné-Vázquez, M. Herve, M. Jung, J. Love, F. Miguez, J. Piaskowski, H. Riebl, and H. Singmann. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. 2025. DOI: 10.32614/CRAN.package.emmeans.

[268] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff. "A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory". In: *Presence: Teleoperators and Virtual Environments* 10.3 (2001), pp. 282–297. DOI: 10.1162/105474601300343612.

[269] S. C. Levinson and F. Torreira. "Timing in Turn-Taking and Its Implications for Processing Models of Language". In: *Frontiers in Psychology* 6 (2015), p. 731. DOI: 10.3389/fpsyg.2015.00731.

[270] G. Li, J. A. Anguera, S. V. Javed, M. A. Khan, G. Wang, and A. Gazzaley. "Enhanced Attention Using Head-Mounted Virtual Reality". In: *Journal of Cognitive Neuroscience* 32.8 (2020), pp. 1438–1454. DOI: 10.1162/jocn_a_01560.

[271] A. Liebl, J. Haller, B. Jödicke, H. Baumgartner, S. Schlittmeier, and J. Hellbrück. "Combined Effects of Acoustic and Visual Distraction on Cognitive Performance and Well-Being". In: *Applied Ergonomics*. Special Section on Product Comfort 43.2 (2012), pp. 424–434. DOI: 10.1016/j.apergo.2011.06.017.

[272] S. Y. Lim, B. Jun, H. J. Lim, J. H. Ahn, Y. W. Yi, and H. Kim. "Display Resolution and Human Factors for Presence and Motion Sickness in HMD Experiences". In: *SID Symposium Digest of Technical Papers* 50.1 (2019), pp. 1142–1145. DOI: 10.1002/sdtp.13131.

[273] X. Lin and N. E. Breslow. "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion". In: *Journal of the American Statistical Association* 91.435 (1996), pp. 1007–1016. DOI: 10.1080/01621459.1996.10476971.

[274] A. Lindau and S. Weinzierl. "Assessing the Plausibility of Virtual Acoustic Environments". In: *Acta Acustica united with Acustica* 98.5 (2012), pp. 804–810. DOI: 10.3813/AAA.918562.

[275] R. Y. Litovsky, M. J. Goupell, R. R. Fay, and A. N. Popper, eds. *Binaural Hearing*. Vol. 73. Springer Handbook of Auditory Research. Cham: Springer International Publishing, 2021. ISBN: 978-3-030-57099-6. DOI: 10.1007/978-3-030-57100-9.

[276] R. Ljung and A. Kjellberg. "Long Reverberation Time Decreases Recall of Spoken Information". In: *Building Acoustics* 16.4 (2009), pp. 301–311. DOI: 10.1260/135101009790291273.

[277] R. H. Logie, G. M. Zucco, and A. D. Baddeley. "Interference with Visual Short-Term Memory". In: *Acta Psychologica* 75.1 (1990), pp. 55–74. DOI: 10.1016/0001-6918(90)90066-O.

[278] K. Loh, E. Fintor, S. Nolden, and J. Fels. "Children's Intentional Switching of Auditory Selective Attention in Spatial and Noisy Acoustic Environments in Comparison to Adults". In: *Developmental Psychology* 58.1 (2022), pp. 69–82. DOI: 10.1037/dev0001239.

[279] D. Lüdecke, D. Makowski, M. S. Ben-Shachar, I. Patil, P. Waggoner, B. M. Wiernik, and R. Thériault. *performance: Assessment of Regression Models Performance*. 2019. DOI: 10.32614/CRAN.package.performance.

[280] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, and A. Spriet. "Multicenter Evaluation of Signal Enhancement Algorithms for Hearing Aids". In: *The Journal of the Acoustical Society of America* 127.3 (2010), pp. 1491–1505. DOI: 10.1121/1.3299168.

[281] W. J. Macken, F. G. Phelps, and D. M. Jones. "What Causes Auditory Distraction?" In: *Psychonomic Bulletin & Review* 16.1 (2009), pp. 139–144. DOI: 10.3758/PBR.16.1.139.

[282] C. L. Mackersie and H. Cones. "Subjective and Psychophysiological Indexes of Listening Effort in a Competing-Talker Task". In: *Journal of the American Academy of Audiology* 22.2 (2011), pp. 113–122. DOI: 10.3766/jaaa.22.2.6.

[283] A. Macleod and Q. Summerfield. "Quantifying the Contribution of Vision to Speech Perception in Noise". In: *British Journal of Audiology* 21.2 (1987), pp. 131–141. DOI: 10.3109/03005368709077786.

[284] M. C. Macpherson, K. Ma. Miles, A. Weisser, J. Buchholz, S. Carlile, and M. J. Richardson. "Exploring the Impact of Background Noise on Fractal Dynamics and Complexity Matching during Social Interaction". In: *OSF* (2025). DOI: 10.31234/osf.io/cwgpx_v1. Pre-published.

[285] R. K. Maddox and B. G. Shinn-Cunningham. "Influence of Task-Relevant and Task-Irrelevant Feature Continuity on Selective Auditory Attention". In: *Journal of the Association for Research in Otolaryngology* 13.1 (2012), pp. 119–129. DOI: 10.1007/s10162-011-0299-7.

[286] E. Magosso, F. De Crescenzio, G. Ricci, S. Piastra, and M. Ursino. "EEG Alpha Power Is Modulated by Attentional Changes during Cognitive Tasks and Virtual Reality Immersion". In: *Changes* 33 (2019), p. 34.

[287] S. Mahalakshmi, K. Valarmathi, L. S. Karthick, K. Balasubramanian, B. Mohan, and P. A. Mathina. "Respond User Mindset for Enhanced VR Experiences: Integrating GSR Sensors with Virtual Reality Technology". In: *2025 3rd International Conference on Advancement in Computation; Computer Technologies (InCACCT)*. Gharuan, India: IEEE, 2025, pp. 856–862. ISBN: 979-8-3315-4389-1. DOI: 10.1109/InCACCT65424.2025.11011445.

[288] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke. "bayestestR: Describing Effects and Their Uncertainty, Existence and Significance within the Bayesian Framework." In: *Journal of Open Source Software* 4.40 (2019), p. 1541. DOI: 10.21105/joss.01541.

[289] E. Malbos, R. M. Rapee, and M. Kavakli. "Behavioral Presence Test in Threatening Virtual Environments". In: *Presence: Teleoperators and Virtual Environments* 21.3 (2012), pp. 268–280. DOI: 10.1162/PRES_a_00112.

[290] S. Malpica, A. Serrano, M. Allue, M. G. Bedia, and B. Masia. "Crossmodal Perception in Virtual Reality". In: *Multimedia Tools and Applications* 79.5 (2020), pp. 3311–3331. DOI: 10.1007/s11042-019-7331-z.

[291] S. Marcos-Pablos, E. González-Pablos, C. Martín-Lorenzo, L. A. Flores, J. Gómez-García-Bermejo, and E. Zalama. "Virtual Avatar for Emotion Recognition in Patients with Schizophrenia: A Pilot Study". In: *Frontiers in Human Neuroscience* 10 (2016). DOI: 10.3389/fnhum.2016.00421.

[292] F. Marmel and D. Lelic. "Influence of Hearing Aid Processing on Acoustic Features Associated with Emotional Speech: Acoustic Analyses and Perception by

Listeners with Normal Hearing". In: *JASA Express Letters* 5.9 (2025), p. 095203. DOI: 10.1121/10.0039220.

[293] J. E. Marsh, R. W. Hughes, and D. M. Jones. "Interference by Process, Not Content, Determines Semantic Auditory Distraction". In: *Cognition* 110.1 (2009), pp. 23–38. DOI: 10.1016/j.cognition.2008.08.003.

[294] J. E. Marsh, R. Ljung, H. Jahncke, D. MacCutcheon, F. Pausch, L. J. Ball, and F. Vachon. "Why Are Background Telephone Conversations Distracting?" In: *Journal of Experimental Psychology. Applied* 24.2 (2018), pp. 222–235. DOI: 10.1037/xap0000170.

[295] J. E. Marsh, M. Hurlstone, A. Marois, L. Ball, S. Moore, F. Vachon, S. Schlittmeier, J. Röer, A. Buchner, F. Aust, and R. Bell. "Changing-State Irrelevant Speech Disrupts Visual-Verbal but Not Visual-Spatial Serial Recall". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* (2023). DOI: 10.1037/xlm0001360.

[296] C. S. Martin, J. W. Mullennix, D. B. Pisoni, and W. V. Summers. "Effects of Talker Variability on Recall of Spoken Word Lists". In: *Journal of Experimental Psychology. Learning, Memory, and Cognition* 15.4 (1989), pp. 676–684. DOI: 10.1037//0278-7393.15.4.676.

[297] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano. "Multimodality in VR: A Survey". In: *ACM Computing Surveys* 54 (2022), 216:1–216:36. DOI: 10.1145/3508361.

[298] B. Masiero and J. Fels. "Perceptually Robust Headphone Equalization for Binaural Reproduction". In: Audio Engineering Society, 2011.

[299] D. W. Massaro, M. M. Cohen, and P. M. Smeele. "Perception of Asynchronous and Conflicting Visual and Auditory Speech". In: *The Journal of the Acoustical Society of America* 100.3 (1996), pp. 1777–1786. DOI: 10.1121/1.417342.

[300] H. Matuschek, R. Kliegl, S. Vasishth, H. Baayen, and D. Bates. "Balancing Type I Error and Power in Linear Mixed Models". In: *Journal of Memory and Language* 94 (2017), pp. 305–315. DOI: 10.1016/j.jml.2017.01.001.

[301] T. May, S. Van De Par, and A. Kohlrausch. "Binaural Localization and Detection of Speakers in Complex Acoustic Scenes". In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 397–425. ISBN: 978-3-642-37761-7. DOI: 10.1007/978-3-642-37762-4_15.

[302] J. McConnell and J. G. Quinn. "Complexity Factors in Visuo-spatial Working Memory". In: *Memory* 12.3 (2004), pp. 338–350. DOI: 10.1080/09658210344000035.

[303] R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay. "Listening Effort and Fatigue: What Exactly Are We Measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'White Paper'". In: *International Journal of Audiology* 53.7 (2014), pp. 433–440. DOI: 10.3109/14992027.2014.890296.

[304] M. McGillycuddy, D. I. Warton, G. Popovic, and B. M. Bolker. "Parsimoniously Fitting Large Multivariate Random Effects in glmmTMB". In: *Journal of Statistical Software* 112.1 (2025). DOI: 10.18637/jss.v112.i01.

[305] M. McGrath and Q. Summerfield. "Intermodal Timing Relations and Audio-visual Speech Recognition by Normal-hearing Adults". In: *The Journal of the Acoustical Society of America* 77.2 (1985), pp. 678–685. DOI: 10.1121/1.392336.

[306] H. Mcgurk and J. Macdonald. "Hearing Lips and Seeing Voices". In: *Nature* 264.5588 (1976), pp. 746–748. DOI: 10.1038/264746a0.

[307] G. McKeown, C. Spencer, A. Patterson, T. Creaney, and D. Dupré. "Comparing Virtual Reality with Computer Monitors as Rating Environments for Affective Dimensions in Social Interactions". In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2017, pp. 464–469. DOI: 10.1109/ACII.2017.8273640.

[308] D. McNeish. "On Using Bayesian Methods to Address Small Sample Problems". In: *Structural Equation Modeling: A Multidisciplinary Journal* 23.5 (2016), pp. 750–773. DOI: 10.1080/10705511.2016.1186549.

[309] H. Meister, S. Rählmann, U. Lemke, and J. Besser. "Verbal Response Times as a Potential Indicator of Cognitive Load During Conventional Speech Audiometry With Matrix Sentences". In: *Trends in Hearing* 22 (2018), p. 2331216518793255. DOI: 10.1177/2331216518793255.

[310] *MetaHuman Creator: High-Fidelity Digital Humans Made Easy | Unreal Engine*. 2021. URL: https://www.youtube.com/watch?v=S3F1vZYpH8c.

[311] N. Meyer-Kahlen, S. J. Schlecht, S. Amengual Garí, and T. Lokki. "Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility, and Authenticity". In: *Journal of the Audio Engineering Society* 72.11 (2024), pp. 797–812. DOI: 10.17743/jaes.2022.0178.

[312] J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay, eds. *The Auditory System at the Cocktail Party*. Vol. 60. Springer Handbook of Auditory Research. Cham: Springer International Publishing, 2017. ISBN: 978-3-319-51660-8. DOI: 10.1007/978-3-319-51662-2.

[313] K. Miles, T. Beechey, V. Best, and J. Buchholz. "Measuring Speech Intelligibility and Hearing-Aid Benefit Using Everyday Conversational Sentences in Real-World Environments". In: *Frontiers in Neuroscience* 16 (2022), p. 789565. DOI: 10.3389/fnins.2022.789565.

[314] K. Miles, V. Best, and J. M. Buchholz. "Feasibility of an Adaptive Version of the Everyday Conversational Sentences in Noise Test". In: *Journal of Speech, Language, and Hearing Research* 67.2 (2024), pp. 680–687. DOI: 10.1044/2023_JSLHR-23-00507.

[315] K. Miles, A. Weisser, R. W. Kallen, M. Varlet, M. J. Richardson, and J. M. Buchholz. "Behavioral Dynamics of Conversation, (Mis)Communication and Coordination in Noisy Environments". In: *Scientific Reports* 13.1 (2023), p. 20271. DOI: 10.1038/s41598-023-47396-y.

[316] K. M. Miles, G. Keidser, K. Freeston, T. Beechey, V. Best, and J. M. Buchholz. "Development of the Everyday Conversational Sentences in Noise Test". In: *The Journal of the Acoustical Society of America* 147.3 (2020), pp. 1562–1576. DOI: 10.1121/10.0000780.

[317] A. W. Mills. "On the Minimum Audible Angle". In: *The Journal of the Acoustical Society of America* 30.4 (1958), pp. 237–246. DOI: 10.1121/1.1909553.

[318] C. Mogensen, L. M. Miller, and S. Roodenrys. "Not so Fast! Talker Variability in Serial Recall at Standard Presentation Rates". In: *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale* 69.1 (2015), pp. 39–53. DOI: 10.1037/cep0000044.

[319] C. Mohanathasan, J. Ehret, C. A. Ermert, J. Fels, T. Kuhlen, and S. J. Schlittmeier. "Towards More Realistic Listening Research in Virtual Environments: The Effect of Spatial Separation of Two Talkers in Conversations on Memory and Listening Effort". In: *Fortschritte der Akustik - DAGA 2023*. Ed. by O. von Estorff and S. Lippert. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2023. ISBN: 978-3-939296-21-8. DOI: 10.18154/RWTH-2023-05116.

[320] C. Mohanathasan, C. A. Ermert, J. Fels, T. W. Kuhlen, and S. J. Schlittmeier. "Exploring short-term memory and listening effort in two-talker conversations: The influence of soft and moderate background noise". In: *PLOS ONE* 20.2 (2025), e0318821. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0318821.

[321] C. Mohanathasan, J. Fels, and S. J. Schlittmeier. "Listening to Two-Talker Conversations in Quiet Settings: The Role of Listeners' Cognitive Processing Capabilities for Memory and Listening Effort". In: *Scientific Reports* 14.1 (2024), p. 22764. DOI: 10.1038/s41598-024-74085-1.

[322] C. Mohanathasan, P. B. Koleva, J. Ehret, A. Bönsch, J. Fels, T. W. Kuhlen, and S. J. Schlittmeier. "Beyond Words: The Impact of Static and Animated Faces as Visual Cues on Memory Performance and Listening Effort during Two-Talker Conversations". In: *Acta Psychologica* 263 (2026), p. 106295. DOI: 10.1016/j.actpsy.2026.106295.

[323] C. Mohanathasan, S. J. Schlittmeier, and I. S. Schiller. *Difficulty Indices for 34 Texts in the Heard Text Recall (HTR) Paradigm: Measuring Text Comprehension and Memory*. RWTH-2025-05207. Lehr- und Forschungsgebiet Arbeits- und Ingenieurpsychologie, 2025. DOI: 10.18154/RWTH-2025-05207.

[324] A. H. Moore, T. Green, M. Brookes, and P. A. Naylor. "Measuring Audio-Visual Speech Intelligibility under Dynamic Listening Conditions Using Virtual Reality". In: *Proceedings of the AES International Conference*. AES 2022 International Audio for Virtual and Augmented Reality Conference. Vol. 2022-A. Audio Engineering Society, 2022.

[325] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. 6th ed. Leiden, The Netherlands: Brill, 2013.

[326] M. Mori, K. F. MacDorman, and N. Kageki. "The Uncanny Valley [From the Field]". In: *IEEE Robotics & Automation Magazine* 19.2 (2012), pp. 98–100. DOI: 10.1109/MRA.2012.2192811.

[327] A. Moulin, J. Vergne, S. Gallego, and C. Micheyl. "A New Speech, Spatial, and Qualities of Hearing Scale Short-Form: Factor, Cluster, and Comparative Analyses". In: *Ear and Hearing* 40.4 (2019), p. 938. DOI: 10.1097/AUD.0000000000000675.

[328] I. Muhammad, M. Vorländer, and S. J. Schlittmeier. "Audio-Video Virtual Reality Environments in Building Acoustics: An Exemplary Study Reproducing Performance Results and Subjective Ratings of a Laboratory Listening Experiment". In: *The Journal of the Acoustical Society of America* 146.3 (2019), EL310–EL316. DOI: 10.1121/1.5126598.

[329] "Mundane Realism". In: R. Baumeister and K. Vohs. *Encyclopedia of Social Psychology*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2007. ISBN: 978-1-4129-1670-7. DOI: 10.4135/9781412956253.n357.

[330] J. H. Murray. *Hamlet on the Holodeck, Updated Edition: The Future of Narrative in Cyberspace*. MIT Press, 2017. 435 pp. ISBN: 978-0-262-53348-5.

[331] J. Mütterlein and T. Hess. "Immersion, Presence, Interactivity: Towards a Joint Understanding of Factors Influencing Virtual Reality Acceptance and Use". In: *AMCIS 2017 Proceedings* (2017).

[332] S. Nakagawa and H. Schielzeth. "A general and simple method for obtaining R^2 from generalized linear mixed-effects models". In: *Methods in Ecology and Evolution* 4.2 (2013), pp. 133–142. DOI: 10.1111/j.2041-210x.2012.00261.x.

[333] D. Nardo, V. Santangelo, and E. Macaluso. "Spatial Orienting in Complex Audiovisual Environments". In: *Human Brain Mapping* 35.4 (2014), pp. 1597–1614. DOI: 10.1002/hbm.22276.

[334] S. Naumann, M. L. Byrne, A. De La Fuente, A. Harrewijn, T. Nugiel, M. Rosen, N. Van Atteveldt, and P. J. Matusz. "Assessing the Degree of Ecological Validity of Your Study: Introducing the Multidimensional Assessment of Research in Context (MARC) Tool". In: *Mind, Brain, and Education* 16.3 (2022), pp. 228–238. DOI: 10.1111/mbe.12318.

[335] I. Neath. "Modeling the Effects of Irrelevant Speech on Memory". In: *Psychonomic Bulletin & Review* 7.3 (2000), pp. 403–423. DOI: 10.3758/bf03214356.

[336] E. H. N. Ng, M. Rudner, T. Lunner, M. S. Pedersen, and J. Rönnberg. "Effects of Noise and Working Memory Capacity on Memory Processing of Speech for Hearing-Aid Users". In: *International Journal of Audiology* 52.7 (2013), pp. 433–441. DOI: 10.3109/14992027.2013.776181.

[337] J. Nirme, B. Sahlén, V. Lyberg Åhlander, J. Brännström, and M. Haake. "Audio-Visual Speech Comprehension in Noise with Real and Virtual Speakers". In: *Speech Communication* 116 (2020), pp. 44–55. DOI: 10.1016/j.specom.2019.11.005.

[338] L. C. Nygaard, M. S. Sommers, and D. B. Pisoni. "Effects of Stimulus Variability on Perception and Representation of Spoken Words in Memory". In: *Perception & Psychophysics* 57.7 (1995), pp. 989–1001. DOI: 10.3758/bf03205458.

[339] J. Oberem and J. Fels. *Speech Material for a Paradigm on the Intentional Switching of Auditory Selective Attention*. RWTH-2020-02105. Lehr- und Forschungsgebiet für Medizinische Akustik, RWTH Aachen Publications, 2020. DOI: 10.18154/RWTH-2020-02105.

[340] J. Oberem, B. Masiero, and J. Fels. "Experiments on Authenticity and Plausibility of Binaural Reproduction via Headphones Employing Different Recording Methods". In: *Applied Acoustics* 114 (2016), pp. 71–78. DOI: 10.1016/j.apacoust.2016.07.009.

[341] J. Oberem, J.-G. Richter, D. Setzer, J. Seibold, I. Koch, and J. Fels. "Experiments on Localization Accuracy with Non-Individual and Individual HRTFs Comparing Static and Dynamic Reproduction Methods". In: *bioRxiv* (2020), p. 2020.03.31.011650. DOI: 10.1101/2020.03.31.011650.

[342] J. Oberem, J. Seibold, I. Koch, and J. Fels. "Intentional Switching in Auditory Selective Attention: Exploring Attention Shifts with Different Reverberation Times". In: *Hearing Research* 359 (2018), pp. 32–39. DOI: 10.1016/j.heares.2017.12.013.

[343] D. Oberfeld, N. Kalmbach, C. Von Castell, S. Boustila, L. Aspöck, and D. Bechmann. "Reverberation Makes Interior Spaces Appear Larger". In: *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*. European Acoustics Association, 2023, pp. 5003–5007.

[344] B. Ohlenforst, D. Wendt, S. E. Kramer, G. Naylor, A. A. Zekveld, and T. Lunner. "Impact of SNR, Masker Type and Noise Reduction Processing on Sentence Recognition Performance and Listening Effort as Indicated by the Pupil Dilation Response". In: *Hearing Research* 365 (2018), pp. 90–99. DOI: 10.1016/j.heares.2018.05.003.

[345] B. Ohlenforst, A. A. Zekveld, E. P. Jansma, Y. Wang, G. Naylor, A. Lorens, T. Lunner, and S. E. Kramer. "Effects of Hearing Impairment and Hearing Aid Amplification on Listening Effort: A Systematic Review". In: *Ear and Hearing* 38.3 (2017), p. 267. DOI: 10.1097/AUD.0000000000000396.

[346] B. Ohlenforst, A. A. Zekveld, T. Lunner, D. Wendt, G. Naylor, Y. Wang, N. J. Versfeld, and S. E. Kramer. "Impact of Stimulus-Related Factors and Hearing Impairment on Listening Effort as Indicated by Pupil Dilation". In: *Hearing Research* 351 (2017), pp. 68–79. DOI: 10.1016/j.heares.2017.05.012.

[347] J. Otten, G. Llorach Tó, G. Grimm, and V. Hohmann. "Evaluation of Lip Synchronization Algorithms via SRT Measurements in Virtual Environments". In: *Fortschritte Der Akustik - DAGA 2025*. DAS|DAGA 2025. Berlin, Germany: Deutsche Gesellschaft für Akustik e.V. (DEGA), 2025, pp. 144–147. DOI: 10.71568/dasdaga2025.461.

[348] K. K. Padayachee, S. P. Levitt, and K. J. Nixon. "The Effects of Encountered-Type Haptics on Immersion, Presence, User Experience, and Task Performance in Virtual Control Panel Environments". In: *IEEE Access* 13 (2025), pp. 160288–160312. DOI: 10.1109/ACCESS.2025.3609542.

[349] F. B. R. Parmentier and D. M. Jones. “Functional Characteristics of Auditory Temporal—Spatial Short-Term Memory: Evidence from Serial Order Errors”. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26.1 (2000), pp. 222–238. DOI: 10.1037/0278-7393.26.1.222.

[350] T. D. Parsons. “Virtual Reality for Enhanced Ecological Validity and Experimental Control in the Clinical, Affective and Social Neurosciences”. In: *Frontiers in Human Neuroscience* 9 (2015). DOI: 10.3389/fnhum.2015.00660.

[351] A. Pastusiak, Ł. Błasiński, and J. Kociński. “Listening Effort in Reverberant Rooms: A Comparative Study of Subjective Perception and Objective Acoustic Metrics”. In: *Archives of Acoustics* 50.3 (2025), pp. 321–329. DOI: 10.24425/aoa.2025.154824.

[352] E. Patrick, D. Cosgrove, A. Slavkovic, J. A. Rode, T. Verratti, and G. Chiselko. “Using a Large Projection Screen as an Alternative to Head-Mounted Displays for Virtual Environments”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’00. New York, NY, USA: Association for Computing Machinery, 2000, pp. 478–485. ISBN: 978-1-58113-216-8. DOI: 10.1145/332040.332479.

[353] F. Pausch. *Documentation of the experimental environments and hardware used in the dissertation "Spatial audio reproduction for hearing aid research: System design, evaluation and application"*. RWTH-2022-01536. Lehrstuhl für Hörtechnik und Akustik, RWTH Aachen Publications, 2022. DOI: 10.18154/RWTH-2022-01536.

[354] F. Pausch, G. Behler, and J. Fels. “SCaLAr – A Surrounding Spherical Cap Loudspeaker Array for Flexible Generation and Evaluation of Virtual Acoustic Environments”. In: *Acta Acustica* 4 (2020), p. 19. DOI: 10.1051/aacus/2020014.

[355] E. Payne, K. Battel, N. Baker, and N. Ward. “Considerations When Trying to Quantify Immersion with Secondary Tasks”. In: *Frontiers in Virtual Reality* 5 (2024), p. 1212216. DOI: 10.3389/frvir.2024.1212216.

[356] J. E. Peelle. “Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior”. In: *Ear and Hearing* 39.2 (2018), p. 204. DOI: 10.1097/AUD.0000000000000494.

[357] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv. “PsychoPy2: Experiments in Behavior Made Easy”. In: *Behavior Research Methods* 51.1 (2019), pp. 195–203. DOI: 10.3758/s13428-018-01193-y.

[358] Z. E. Peng and L. M. Wang. “Listening Effort by Native and Nonnative Listeners Due to Noise, Reverberation, and Talker Foreign Accent During English Speech Perception”. In: *Journal of Speech, Language, and Hearing Research* 62.4 (2019), pp. 1068–1081. DOI: 10.1044/2018_JSLHR-H-17-0423.

[359] G. Perfect, K. Smeds, A. Heinrich, and G. H. Saunders. “Responses to Listening Difficulty in In-Person Conversation: A Survey-Based Synthesis of Communication Strategies Used by People with Hearing Loss, Communication Partners, and

Hearing-Care Professionals". In: *International Journal of Audiology* 0.0 (2025), pp. 1–13. DOI: 10.1080/14992027.2025.2558689.

[360] E. B. Petersen. "Hearing-Loss Related Variations in Turn-Taking Time Affect How Conversations Are Perceived". In: *PLOS ONE* 20.6 (2025), e0325244. DOI: 10.1371/journal.pone.0325244.

[361] M. K. Pichora-Fuller, S. E. Kramer, M. A. Eckert, B. Edwards, B. W. Y. Hornsby, L. E. Humes, U. Lemke, T. Lunner, M. Matthen, C. L. Mackersie, G. Naylor, N. A. Phillips, M. Richter, M. Rudner, M. S. Sommers, K. L. Tremblay, and A. Wingfield. "Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL)". In: *Ear and Hearing* 37 Suppl 1.1 (2016), 5S–27S. DOI: 10.1097/AUD.0000000000000312.

[362] E. M. Picou, B. Bean, S. C. Marcrum, T. A. Ricketts, and B. W. Y. Hornsby. "Moderate Reverberation Does Not Increase Subjective Fatigue, Subjective Listening Effort, or Behavioral Listening Effort in School-Aged Children". In: *Frontiers in Psychology* 10 (2019). DOI: 10.3389/fpsyg.2019.01749.

[363] E. M. Picou, J. Gordon, and T. A. Ricketts. "The Effects of Noise and Reverberation on Listening Effort in Adults With Normal Hearing". In: *Ear and Hearing* 37.1 (2016), p. 1. DOI: 10.1097/AUD.0000000000000222.

[364] E. M. Picou, T. A. Ricketts, and B. W. Y. Hornsby. "Visual Cues and Listening Effort: Individual Variability". In: *Journal of Speech, Language, and Hearing Research* 54.5 (2011), pp. 1416–1430. DOI: 10.1044/1092-4388(2011/10-0154).

[365] E. M. Picou, T. A. Ricketts, and B. W. Y. Hornsby. "How Hearing Aids, Background Noise, and Visual Cues Influence Objective Listening Effort". In: *Ear and Hearing* 34.5 (2013), e52. DOI: 10.1097/AUD.0b013e31827f0431.

[366] E. M. Picou, G. Singh, H. Goy, F. Russo, L. Hickson, A. J. Oxenham, G. H. Buono, T. A. Ricketts, and S. Launer. "Hearing, Emotion, Amplification, Research, and Training Workshop: Current Understanding of Hearing Loss and Emotion Perception and Priorities for Future Research". In: *Trends in Hearing* 22 (2018), p. 2331216518803215. DOI: 10.1177/2331216518803215.

[367] L. Pieri, G. Tosi, and D. Romano. "Virtual Reality Technology in Neuropsychological Testing: A Systematic Review". In: *Journal of Neuropsychology* 17.2 (2023), pp. 382–399. DOI: 10.1111/jnp.12304.

[368] D. Pietschmann. "Presence, Immersion". In: *Handbuch Online-Kommunikation*. Springer VS, Wiesbaden, 2025, pp. 1–19. ISBN: 978-3-658-18017-1. DOI: 10.1007/978-3-658-18017-1_29-1.

[369] Poeschl Sandra and Doering Nicola. "Measuring Co-Presence and Social Presence in Virtual Environments - Psychometric Construction of a German Scale for a Fear of Public Speaking Scenario". In: *Studies in Health Technology and Informatics*. IOS Press, 2015. DOI: 10.3233/978-1-61499-595-1-58.

[370] C. Popp and D. T. Murphy. "Speech Intelligibility versus Congruency: User Preferences of the Acoustics of Virtual Reality Game Spaces". In: *Virtual Worlds* 3.1 (2024), pp. 40–61. DOI: 10.3390/virtualworlds3010003.

[371] B. N. J. Postma and B. F. G. Katz. "The Influence of Visual Distance on the Room-Acoustic Experience of Auralizations". In: *The Journal of the Acoustical Society of America* 142.5 (2017), pp. 3035–3046. DOI: 10.1121/1.5009554.

[372] N. Prodi and C. Visentin. "A Slight Increase in Reverberation Time in the Classroom Affects Performance and Behavioral Listening Effort". In: *Ear and Hearing* 43.2 (2022), p. 460. DOI: 10.1097/AUD.0000000000001110.

[373] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, 2023.

[374] R Core Team. *R: A Language and Environment for Statistical Computing*. manual. Vienna, Austria: R Foundation for Statistical Computing, 2024.

[375] J. Radun, H. Maula, I.-K. Tervahartiala, V. Rajala, S. Schlittmeier, and V. Hongisto. "The Effects of Irrelevant Speech on Physiological Stress, Cognitive Performance, and Subjective Experience – Focus on Heart Rate Variability". In: *International Journal of Psychophysiology* 200 (2024), p. 112352. DOI: 10.1016/j.ijpsycho.2024.112352.

[376] N. Ram, M. Brinberg, A. L. Pincus, and D. E. Conroy. "The Questionable Ecological Validity of Ecological Momentary Assessment: Considerations for Design and Analysis". In: *Research in Human Development* 14.3 (2017), pp. 253–270. DOI: 10.1080/15427609.2017.1340052.

[377] J. Reader and M. Savin-Baden. "Ethical Conundrums and Virtual Humans". In: *Postdigital Science and Education* 2.2 (2020), pp. 289–301. DOI: 10.1007/s42438-019-00095-2.

[378] F. Rebelo, P. Noriega, E. Duarte, and M. Soares. "Using Virtual Reality to Assess User Experience". In: *Human Factors* 54.6 (2012), pp. 964–982. DOI: 10.1177/0018720812465006.

[379] E. Redlinger, B. Glas, and Y. Rong. "Enhanced Cognitive Training Using Virtual Reality: Examining a Memory Task Modified for Use in Virtual Environments". In: *2021 5th International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. AIVR 2021. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–8. ISBN: 978-1-4503-8414-8. DOI: 10.1145/3480433.3480435.

[380] R. Rehman, C. Dreier, J. Heck, J. Llorca-Bofí, and M. Vorländer. "Comparison of Virtual Reality and Web-Based Listening Experiments on the Perception in Complex Auralized Environments". In: *The Journal of the Acoustical Society of America* 157.3 (2025), pp. 2001–2017. DOI: 10.1121/10.0036147.

[381] J. Rennies, V. Best, E. Roverud, and G. Kidd. "Energetic and Informational Components of Speech-on-Speech Masking in Binaural Speech Intelligibility and Perceived Listening Effort". In: *Trends in Hearing* 23 (2019), p. 2331216519854597. DOI: 10.1177/2331216519854597.

[382] J. Rennies, H. Schepker, I. Holube, and B. Kollmeier. "Listening Effort and Speech Intelligibility in Listening Situations Affected by Noise and Reverberation". In: *The Journal of the Acoustical Society of America* 136.5 (2014), pp. 2642–2653. DOI: 10.1121/1.4897398.

[383] T. Renz, P. Leistner, and A. Liebl. "The Effect of Spatial Separation of Sound Masking and Distracting Speech Sounds on Working Memory Performance and Annoyance". In: *Acta Acustica united with Acustica* 104.4 (2018), pp. 611–622. DOI: 10.3813/AAA.919201.

[384] J.-G. Richter, G. Behler, and J. Fels. "Evaluation of a Fast HRTF Measurement System". In: *140th Audio Engineering Society Convention*. Audio Engineering Society, 2016. URL: http://www.aes.org/e-lib/browse.cfm?elib=18197.

[385] S. G. Roberts, F. Torreira, and S. C. Levinson. "The Effects of Processing and Sequence Organization on the Timing of Turn Taking: A Corpus Study". In: *Frontiers in Psychology* 6 (2015), p. 136319. DOI: 10.3389/fpsyg.2015.00509.

[386] T. Robotham, A. Singla, O. S. Rummukainen, A. Raake, and E. A. P. Habets. "Audiovisual Database with 360° Video and Higher-Order Ambisonics Audio for Perception, Cognition, Behavior, and QoE Evaluation Research". In: *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*. 2022, pp. 1–6. DOI: 10.1109/QoMEX55416.2022.9900893.

[387] B. E. Roe and D. R. Just. "Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments, and Field Data". In: *American Journal of Agricultural Economics* 91.5 (2009), pp. 1266–1271.

[388] J. P. Röer, R. Bell, U. Körner, and A. Buchner. "A semantic mismatch effect on serial recall: Evidence for interlexical processing of irrelevant speech". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45.3 (2019), pp. 515–525. DOI: 10.1037/xlm0000596.

[389] J. Rönnberg. "Cognition in the Hearing Impaired and Deaf as a Bridge between Signal and Dialogue: A Framework and a Model". In: *International Journal of Audiology* 42 (2003), pp. 68–76. DOI: 10.3109/14992020309074626.

[390] J. Rönnberg, E. Holmer, and M. Rudner. "Cognitive Hearing Science and Ease of Language Understanding". In: *International Journal of Audiology* 58.5 (2019), pp. 247–261. DOI: 10.1080/14992027.2018.1551631.

[391] J. Rönnberg, T. Lunner, A. Zekveld, P. Sörqvist, H. Danielsson, B. Lyxell, O. Dahlström, C. Signoret, S. Stenfelt, M. K. Pichora-Fuller, and M. Rudner. "The Ease of Language Understanding (ELU) Model: Theoretical, Empirical, and Clinical Advances". In: *Frontiers in Systems Neuroscience* 7 (2013), p. 31. DOI: 10.3389/fnsys.2013.00031.

[392] J. Rönnberg, M. Rudner, C. Foo, and T. Lunner. "Cognition Counts: A Working Memory System for Ease of Language Understanding (ELU)". In: *International Journal of Audiology* 47 (2008), S99–S105. DOI: 10.1080/14992020802301167.

[393] S. Rosen, P. Souza, C. Ekelund, and A. A. Majeed. "Listening to Speech in a Background of Other Talkers: Effects of Talker Number and Noise Vocoding". In: *The Journal of the Acoustical Society of America* 133.4 (2013), pp. 2431–2443. DOI: 10.1121/1.4794379.

[394] S. Roßkopf, A. Mühlberger, F. Stärz, S. Van De Par, M. Blau, and L. O. Kroczek. "Impact of Visual Virtual Scene and Localization Task on Auditory Distance

Perception in Virtual Reality". In: *IEEE Transactions on Visualization and Computer Graphics* 31.5 (2025), pp. 2464–2474. DOI: 10.1109/TVCG.2025.3549855.

[395] W. C. Sabine. *Collected Papers on Acoustics*. Courier Dover Publications, 2015. ISBN: 978-0-486-78627-8.

[396] P. Salamé and A. Baddeley. "Disruption of Short-Term Memory by Unattended Speech: Implications for the Structure of Working Memory". In: *Journal of Verbal Learning and Verbal Behavior* 21.2 (1982), pp. 150–164. DOI: 10.1016/S0022-5371(82)90521-7.

[397] M. Salanger, D. Lewis, T. Vallier, T. McDermott, and A. Dergan. "Applying Virtual Reality to Audiovisual Speech Perception Tasks in Children". In: *American Journal of Audiology* 29.2 (2020), pp. 244–258. DOI: 10.1044/2020_AJA-19-00004.

[398] H. Sato, M. Morimoto, H. Sato, and M. Wada. "Relationship between Listening Difficulty and Acoustical Objective Measures in Reverberant Sound Fields". In: *Journal of the Acoustical Society of America* 123.4 (2008), pp. 2087–2093. DOI: 10.1121/1.2885750.

[399] I. S. Schiller, C. Breuer, L. Aspöck, J. Ehret, A. Bönsch, T. W. Kuhlen, J. Fels, and S. J. Schlittmeier. "A Lecturer's Voice Quality and Its Effect on Memory, Listening Effort, and Perception in a VR Environment". In: *Scientific Reports* 14.1 (2024), p. 12407. DOI: 10.1038/s41598-024-63097-6.

[400] N. Schinkel-Bielefeld, L. Burke, I. Holube, M. Iankilevitch, L. M. Jenstad, D. Lelic, G. Naylor, G. Singh, K. Smeds, P. von Gablenz, F. Wolters, and Y.-H. Wu. "Implementing Ecological Momentary Assessment in Audiological Research: Opportunities and Challenges". In: *American Journal of Audiology* 33.3 (2024), pp. 648–673. DOI: 10.1044/2024_AJA-23-00249.

[401] S. J. Schlittmeier, J. Hellbrück, and M. Klatte. "Can the Irrelevant Speech Effect Turn into a Stimulus Suffix Effect?" In: *Quarterly Journal of Experimental Psychology* 61.5 (2008), pp. 665–673. DOI: 10.1080/17470210701774168.

[402] S. J. Schlittmeier, N. Weisz, and O. Bertrand. "What characterizes changing-state speech in affecting short-term memory? An EEG study on the irrelevant sound effect". In: *Psychophysiology* 48.12 (2011), pp. 1669–1680. DOI: 10.1111/j.1469-8986.2011.01263.x.

[403] S. J. Schlittmeier and E. Fintor. "Remembering Landmarks in a Virtual Maze: Does the Disturbance Impact of Background Speech Depend on the Spatial Information Inherent in the Speech Signal?" In: *Proceedings of the 23rd International Congress on Acoustics*. Aachen, Germany, 2019.

[404] S. J. Schlittmeier, J. Hellbrück, and M. Klatte. "Does Irrelevant Music Cause an Irrelevant Sound Effect for Auditory Items?" In: *European Journal of Cognitive Psychology* 20.2 (2008), pp. 252–271. DOI: 10.1080/09541440701427838.

[405] S. J. Schlittmeier, J. Hellbrück, R. Thaden, and M. Vorländer. "The Impact of Background Speech Varying in Intelligibility: Effects on Cognitive Performance

and Perceived Disturbance". In: *Ergonomics* 51.5 (2008), pp. 719–736. DOI: 10.1080/00140130701745925.

[406] S. J. Schlittmeier, I. S. Schiller, C. Mohanathasan, and A. Liebl. *Measuring Text Comprehension and Memory: A Comprehensive Database for Heard Text Recall (HTR) and Read Text Recall (RTR) Paradigms, with Optional Note-Taking and Graphical Displays*. RWTH-2023-05285. RWTH Publications, 2023. DOI: 10.18154/RWTH-2023-05285.

[407] S. J. Schlittmeier, T. Weißgerber, S. Kerber, H. Fastl, and J. Hellbrück. "Algorithmic Modeling of the Irrelevant Sound Effect (ISE) by the Hearing Sensation Fluctuation Strength". In: *Attention, Perception, & Psychophysics* 74.1 (2012), pp. 194–203. DOI: 10.3758/s13414-011-0230-7.

[408] A. Schmitz. "Ein Neues Digitales Kunstkopfmeßsystem". In: *Acustica: International Journal on Acoustics* 81.4 (1995), pp. 416–420.

[409] M. Schoeffler, J. L. Gernert, M. Neumayer, S. Westphal, and J. Herre. "On the Validity of Virtual Reality-Based Auditory Experiments: A Case Study about Ratings of the Overall Listening Experience". In: *Virtual Reality* 19.3 (2015), pp. 181–200. DOI: 10.1007/s10055-015-0270-8.

[410] D. Schröder. *Physically Based Real-Time Auralization of Interactive Virtual Environments*. Vol. 11. Aachener Beiträge Zur Technischen Akustik. Aachen: Hochschulbibliothek der Rheinisch-Westfälischen Technischen Hochschule Aachen, 2011. ISBN: 978-3-8325-3031-0.

[411] B. Schulte-Fortkamp, A. Fiebig, J. A. Sisneros, A. N. Popper, and R. R. Fay, eds. *Soundscapes: Humans and Their Acoustic Environment*. Vol. 76. Springer Handbook of Auditory Research. Cham: Springer International Publishing, 2023. ISBN: 978-3-031-22778-3. DOI: 10.1007/978-3-031-22779-0.

[412] M. Schutte, S. D. Ewert, and L. Wiegrebe. "The Percept of Reverberation Is Not Affected by Visual Room Impression in Virtual Environments". In: *The Journal of the Acoustical Society of America* 145.3 (2019), EL229–EL235. DOI: 10.1121/1.5093642.

[413] J. Schütze, S. D. Ewert, C. Kirsch, and B. Kollmeier. "Unaided and Aided Speech Intelligibility in a Real and Virtual Acoustic Environment". In: *Trends in Hearing* 29 (2025), p. 23312165251389112. DOI: 10.1177/23312165251389112.

[414] J. Schütze, C. Kirsch, B. Kollmeier, and S. D. Ewert. "Comparison of Speech Intelligibility in a Real and Virtual Living Room Using Loudspeaker and Headphone Presentations". In: *Acta Acustica* 9 (2025), p. 6. DOI: 10.1051/aacus/2024068.

[415] J. Seitz, K. Loh, and J. Fels. "Listening Effort in Children and Adults in Classroom Noise". In: *Scintific Reports* 14.1 (2024), p. 25200. DOI: 10.1038/s41598-024-76932-7.

[416] H. Y. Seol, S. Kang, J. Lim, S. H. Hong, and I. J. Moon. "Feasibility of Virtual Reality Audiological Testing: Prospective Study". In: *JMIR Serious Games* 9.3 (2021), e26976. DOI: 10.2196/26976.

[417] W. Setti, I. A.-M. Engel, L. F. Cuturi, M. Gori, and L. Picinali. "The Audio-Corsi: An Acoustic Virtual Reality-Based Technological Solution for Evaluating Audio-Spatial Memory Abilities". In: *Journal on Multimodal User Interfaces* 16.2 (2022), pp. 207–218. DOI: 10.1007/s12193-021-00383-x.

[418] K. Sewell, V. Brown, G. Farwell, M. Rogers, X. Zhang, and J. Strand. "The Effects of Temporal Cues, Point-Light Displays, and Faces on Speech Identification and Listening Effort". In: *PLOS One* 18 (2023), e0290826. DOI: 10.1371/journal.pone.0290826.

[419] S. A. Shamma, M. Elhilali, and C. Micheyl. "Temporal Coherence and Attention in Auditory Scene Analysis". In: *Trends in Neurosciences* 34.3 (2011), pp. 114–123. DOI: 10.1016/j.tins.2010.11.002.

[420] K. Shavit-Cohen and E. Zion Golumbic. "The Dynamics of Attention Shifts Among Concurrent Speech in a Naturalistic Multi-speaker Virtual Environment". In: *Frontiers in Human Neuroscience* 13 (2019). DOI: 10.3389/fnhum.2019.00386.

[421] S. W. Sheffield, H. J. Wheeler, D. S. Brungart, and J. G. W. Bernstein. "The Effect of Sound Localization on Auditory-Only and Audiovisual Speech Recognition in a Simulated Multitalker Environment". In: *Trends in Hearing* 27 (2023), p. 23312165231186040. DOI: 10.1177/23312165231186040.

[422] B. Shinn-Cunningham, V. Best, and A. K. C. Lee. "Auditory Object Formation and Selection". In: *The Auditory System at the Cocktail Party*. Ed. by J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay. Cham: Springer International Publishing, 2017, pp. 7–40. ISBN: 978-3-319-51662-2. DOI: 10.1007/978-3-319-51662-2_2.

[423] B. G. Shinn-Cunningham. "Object-Based Auditory and Visual Attention". In: *Trends in Cognitive Sciences* 12.5 (2008), pp. 182–186. DOI: 10.1016/j.tics.2008.02.003.

[424] D. E. Shub, S. P. Carr, Y. Kong, and H. S. Colburn. "Discrimination and Identification of Azimuth Using Spectral Shape". In: *The Journal of the Acoustical Society of America* 124.5 (2008), pp. 3132–3141. DOI: 10.1121/1.2981634.

[425] A. Siddig, P. W. Sun, M. Parker, and A. Hines. "Perception Deception: Audio-Visual Mismatch in Virtual Reality Using The McGurk Effect". In: *AICS*. Vol. 2019. 2019, pp. 176–187.

[426] D. J. Simons. "Attentional Capture and Inattentional Blindness". In: *Trends in Cognitive Sciences* 4.4 (2000), pp. 147–155. DOI: 10.1016/S1364-6613(00)01455-8.

[427] A. Singla, T. Robotham, A. Bhattacharya, W. Menz, E. A. P. Habets, and A. Raake. "Saliency of Omnidirectional Videos with Different Audio Presentations: Analyses and Dataset". In: *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023, pp. 264–269. DOI: 10.1109/QoMEX58391.2023.10178588.

[428] H. Singmann, B. Bolker, J. Westfall, F. Aust, and M. S. Ben-Shachar. *Afex: Analysis of Factorial Experiments*. manual. 2025.

[429] T. Sivula, M. Magnusson, A. A. Matamoros, and A. Vehtari. "Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison". In: *Bayesian Anal.* (2025). DOI: 10.1214/25-BA1569.

[430] A. Skulmowski, S. Nebel, M. Remmele, and G. D. Rey. "Is a Preference for Realism Really Naive After All? A Cognitive Model of Learning with Realistic Visualizations". In: *Educational Psychology Review* 34.2 (2022), pp. 649–675. DOI: 10.1007/s10648-021-09638-1.

[431] M. Slater. "Place Illusion and Plausibility Can Lead to Realistic Behaviour in Immersive Virtual Environments". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1535 (2009), pp. 3549–3557. DOI: 10.1098/rstb.2009.0138.

[432] M. Slater. "Immersion and the Illusion of Presence in Virtual Reality". In: *British Journal of Psychology* 109.3 (2018), pp. 431–433. DOI: 10.1111/bjop.12305.

[433] M. Slater and S. Wilbur. "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments". In: *Presence: Teleoperators and Virtual Environments* 6.6 (1997), pp. 603–616. DOI: 10.1162/pres.1997.6.6.603.

[434] D. A. Slutsky and G. H. Recanzone. "Temporal and Spatial Dependency of the Ventriloquism Effect". In: *Neuroreport* 12.1 (2001), pp. 7–10. DOI: 10.1097/00001756-200101220-00009.

[435] H. Snellen. *Probebuchstaben Zur Bestimmung Der Sehschärfe*. H. Peters, 1873.

[436] C. Spence. "Audiovisual Multisensory Integration". In: *Acoustical Science and Technology* 28.2 (2007), pp. 61–70. DOI: 10.1250/ast.28.61.

[437] C. Spence. "Just How Important Is Spatial Coincidence to Multisensory Integration? Evaluating the Spatial Rule". In: *Annals of the New York Academy of Sciences* 1296 (2013), pp. 31–49. DOI: 10.1111/nyas.12121.

[438] B. Stahl and G. Marentakis. "Does Serial Memory of Locations Benefit from Spatially Congruent Audiovisual Stimuli? Investigating the Effect of Adding Spatial Sound to Visuospatial Sequences". In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ICMI '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 326–330. ISBN: 978-1-4503-5543-8. DOI: 10.1145/3136755.3136773.

[439] B. E. Stein and M. A. Meredith. *The Merging of the Senses*. The Merging of the Senses. Cambridge, MA, US: The MIT Press, 1993. xv, 211. ISBN: 978-0-262-19331-3.

[440] V. Stenbäck, E. Marsja, M. Hällgren, B. Lyxell, and B. Larsby. "Informational Masking and Listening Effort in Speech Recognition in Noise: The Role of Working Memory Capacity and Inhibitory Control in Older Adults With and Without Hearing Impairment". In: *Journal of Speech, Language, and Hearing Research* 65.11 (2022), pp. 4417–4428. DOI: 10.1044/2022_JSLHR-21-00674.

[441] K. N. Stevens. "Toward a Model for Speech Recognition". In: *The Journal of the Acoustical Society of America* 32.1 (1960), pp. 47–55. DOI: 10.1121/1.1907874.

[442] B. Stodt, D. Neudek, S. Getzmann, E. Wascher, and R. Martin. "Comparing Auditory Distance Perception in Real and Virtual Environments and the Role of the Loudness Cue: A Study Based on Event-Related Potentials". In: *Hearing Research* 444 (2024), p. 108968. DOI: 10.1016/j.heares.2024.108968.

[443] D. J. Strauss, F. I. Corona-Strauss, C. Trenado, C. Bernarding, W. Reith, M. Latzel, and M. Froehlich. "Electrophysiological Correlates of Listening Effort: Neurodynamical Modeling and Measurement". In: *Cognitive Neurodynamics* 4.2 (2010), pp. 119–131. DOI: 10.1007/s11571-010-9111-3.

[444] J. R. Stroop. "Studies of Interference in Serial Verbal Reactions". In: *Journal of Experimental Psychology* 18.6 (1935), pp. 643–662. DOI: 10.1037/h0054651.

[445] W. H. Sumby and I. Pollack. "Visual Contribution to Speech Intelligibility in Noise". In: *The Journal of the Acoustical Society of America* 26.2 (1954), pp. 212–215. DOI: 10.1121/1.1907309.

[446] Q. Summerfield. "Lipreading and Audio-Visual Speech Perception". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 335.1273 (1992), pp. 71–78. DOI: 10.1098/rstb.1992.0009.

[447] A. M. Surprenant, I. Neath, and D. C. LeCompte. "Irrelevant Speech, Phonological Similarity, and Presentation Modality". In: *Memory (Hove, England)* 7.4 (1999), pp. 405–420. DOI: 10.1080/741944920.

[448] J. L. Szalma and P. A. Hancock. "Noise Effects on Human Performance: A Meta-Analytic Synthesis". In: *Psychological Bulletin* 137.4 (2011), pp. 682–707. DOI: 10.1037/a0023987.

[449] K. A. Szerszen. "The Audio/Visual Mismatch and the Uncanny Valley: An Investigation Using a Mismatch in the Human Realism of Facial and Vocal Aspects of Stimuli". Thesis. Indianapolis, USA: Indiana University-Purdue University Indianapolis (IUPUI), 2011. DOI: 10.7912/c2/915.

[450] H. Tadayyoni, M. S. Ramirez Campos, A. J. U. Quevedo, and B. A. Murphy. "Biomarkers of Immersion in Virtual Reality Based on Features Extracted from the EEG Signals: A Machine Learning Approach". In: *Brain Sciences* 14.5 (2024), p. 470. DOI: 10.3390/brainsci14050470.

[451] S. Taylor and G. Asmundson. "Internal and External Validity in Clinical Research". In: *Handbook of Research Methods in Abnormal and Clinical Psychology*. SAGE Publications Inc, 2008, pp. 23–34.

[452] R. Thézé, M. A. Gadiri, L. Albert, A. Provost, A.-L. Giraud, and P. Mégevand. "Animated Virtual Characters to Explore Audio-Visual Speech in Controlled and Naturalistic Environments". In: *Scientific Reports* 10.1 (2020). DOI: 10.1038/s41598-020-72375-y.

[453] J. Thiemann, N. Ito, and E. Vincent. "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A Database of Multichannel Environmental Noise Recordings". In: *Proceedings of Meetings on Acoustics* 19.1 (2013), p. 035081. DOI: 10.1121/1.4799597.

[454] W. R. Thurlow and C. E. Jack. "Certain Determinants of the "Ventriloquism Effect"". In: *Perceptual and Motor Skills* 36.3 (1973), pp. 1171–1184. DOI: 10.2466/pms.1973.36.3c.1171.

[455] T. Q. Tran, T. Langlotz, and H. Regenbrecht. "A Survey On Measuring Presence in Mixed Reality". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, 2024, pp. 1–38. ISBN: 979-8-4007-0330-0. DOI: 10.1145/3613904.3642383.

[456] A. M. Treisman. "Shifting Attention between the Ears". In: *Quarterly Journal of Experimental Psychology* 23.2 (1971), pp. 157–167. DOI: 10.1080/14640747108400236.

[457] S. Tremblay, W. J. MacKen, and D. M. Jones. "The Impact of Broadband Noise on Serial Memory: Changes in Band-Pass Frequency Increase Disruption". In: *Memory* 9.4–6 (2001), pp. 323–331. DOI: 10.1080/09658210143000010.

[458] J. Tromp, D. Peeters, A. S. Meyer, and P. Hagoort. "The Combined Use of Virtual Reality and EEG to Study Language Processing in Naturalistic Environments". In: *Behavior Research Methods* 50.2 (2017), pp. 862–869. DOI: 10.3758/s13428-017-0911-9.

[459] O. Tuomainen, S. Rosen, L. Taschenberger, and V. Hazan. "The Effects of Informational and Energetic/Modulation Masking on the Efficiency and Ease of Speech Communication across the Lifespan". In: *Speech Communication* 162 (2024), p. 103101. DOI: 10.1016/j.specom.2024.103101.

[460] K. Ueda, Y. Nakajima, F. Kattner, and W. Ellermeier. "Irrelevant Speech Effects with Locally Time-Reversed Speech: Native vs Non-Native Language". In: *The Journal of the Acoustical Society of America* 145.6 (2019), pp. 3686–3694. DOI: 10.1121/1.5112774.

[461] S. Ullas, M. Bonte, E. Formisano, and J. Vroomen. "Adaptive Plasticity in Perceiving Speech Sounds". In: *Speech Perception*. Ed. by L. L. Holt, J. E. Peelle, A. B. Coffin, A. N. Popper, and R. R. Fay. Cham: Springer International Publishing, 2022, pp. 173–199. ISBN: 978-3-030-81542-4. DOI: 10.1007/978-3-030-81542-4_7.

[462] M. Usoh, E. Catena, S. Arman, and M. Slater. "Using Presence Questionnaires in Reality". In: *Presence: Teleoperators and Virtual Environments* 9.5 (2000), pp. 497–503. DOI: 10.1162/105474600566989.

[463] F. Vachon, K. Labonté, and J. E. Marsh. "Attentional Capture by Deviant Sounds: A Noncontingent Form of Auditory Distraction?" In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43.4 (2017), pp. 622–634. DOI: 10.1037/xlm0000330.

[464] J. T. Valderrama, J. Mejia, A. Wong, N. C. Herbert, and B. Edwards. "Reduced Listening Effort with Adaptive Binaural Beamforming in Realistic Noisy Environments". In: *Scientific Reports* 15.1 (2025), p. 17998. DOI: 10.1038/s41598-025-95045-3.

[465] S. van de Par, S. D. Ewert, L. Hladek, C. Kirsch, J. Schütze, J. Llorca-Bofí, G. Grimm, M. M. E. Hendrikse, B. Kollmeier, and B. U. Seeber. "Auditory-

Visual Scenes for Hearing Research". In: *Acta Acustica* 6 (2022), p. 55. DOI: 10.1051/aacus/2022032.

[466] S. C. Van Hedger and I. S. Johnsrude. "Speech Perception Under Adverse Listening Conditions". In: *Speech Perception*. Ed. by L. L. Holt, J. E. Peelle, A. B. Coffin, A. N. Popper, and R. R. Fay. Cham: Springer International Publishing, 2022, pp. 141–171. ISBN: 978-3-030-81542-4. DOI: 10.1007/978-3-030-81542-4_6.

[467] M. Van Zyl, M. Wright, and T. Fujioka. "Room Acoustics Affects Immersion and Musical Behavior in Virtual Renderings of Real-World Spaces". In: *Proc. - IEEE Conf. Virtual Real. 3D User Interfaces Abstr. Workshops, VRW*. Institute of Electrical and Electronics Engineers Inc., 2025, pp. 632–635. DOI: 10.1109/VRW66409.2025.00131.

[468] A. Vatakis and C. Spence. "Crossmodal Binding: Evaluating the "Unity Assumption" Using Audiovisual Speech Stimuli". In: *Perception & Psychophysics* 69.5 (2007), pp. 744–756. DOI: 10.3758/bf03193776.

[469] S. Villard, T. Perrachione, S.-J. Lim, A. Alam, and G. Kidd. "Listening Effort Elicited by Energetic versus Informational Masking". In: *Proceedings of Meetings on Acoustics* 45.1 (2022), p. 050002. DOI: 10.1121/2.0001546.

[470] S. Villard, T. K. Perrachione, S.-J. Lim, A. Alam, and G. Kidd Jr. "Energetic and Informational Masking Place Dissociable Demands on Listening Effort: Evidence from Simultaneous Electroencephalography and Pupillometry". In: *The Journal of the Acoustical Society of America* 154.2 (2023), pp. 1152–1167. DOI: 10.1121/10.0020539.

[471] Virtual Reality & Immersive Visualization Group and Institute for Hearing Technology and Acoustics, RWTH Aachen University. *Virtual Acoustics Plugin*. 2023. URL: https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/unreal-va-plugin.

[472] Virtual Reality & Immersive Visualization Group, RWTH Aachen University. *Character Plugin*. 2022. URL: https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/character-plugin.

[473] R. Viveros Muñoz, L. Aspöck, and J. Fels. "Spatial Release From Masking Under Different Reverberant Conditions in Young and Elderly Subjects: Effect of Moving or Stationary Maskers at Circular and Radial Conditions". In: *Journal of Speech, Language, and Hearing Research* 62.9 (2019), pp. 3582–3595. DOI: 10.1044/2019_JSLHR-H-19-0092.

[474] M. Vorländer. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. 1st ed. Berlin: Springer, 2008. ISBN: 978-3-540-48829-3.

[475] K. W. Kühlem, J. Ehret, T. W. Kuhlen, and A. Bönsch. "A Latency-Optimized LLM-based Multimodal Dialogue System for Embodied Conversational Agents in VR". In: *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*. IVA '25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 1–3. ISBN: 979-8-4007-1508-2. DOI: 10.1145/3717511.3749287.

[476] K. Wagener, T. Brand, and B. Kollmeier. "Entwicklung Und Evaluation Eines Satztests Für Die Deutsche Sprache II: Optimierung des Oldenburger Satztests". In: *Zeitschrift für Audiologie* 38.2 (1999), pp. 44–56.

[477] K. Wagener, T. Brand, and B. Kollmeier. "Entwicklung Und Evaluation Eines Satztests Für Die Deutsche Sprache III: Evaluation des Oldenburger Satztests". In: *Zeitschrift für Audiologie* 38.3 (1999), pp. 86–95.

[478] K. Wagener, V. Kuehnel, and B. Kollmeier. "Entwicklung Und Evaluation Eines Satztests Für Die Deutsche Sprache I: Design Des Oldenburger Satztests". In: *Zeitschrift für Audiologie* 38.1 (1999), pp. 4–15.

[479] K. C. Wagener and T. Brand. "Sentence Intelligibility in Noise for Listeners with Normal Hearing and Hearing Impairment: Influence of Measurement Procedure and Masking Parameters". In: *International Journal of Audiology* 44.3 (2005), pp. 144–156. DOI: 10.1080/14992020500057517.

[480] K. C. Wagener, S. Hochmuth, M. Ahrlich, M. A. Zokoll, and B. Kollmeier. "Der Weibliche Oldenburger Satztest". In: *17. Jahrestagung der deutschen Gesellschaft für Audiologie*. Oldenburg, 2014, 4 pages.

[481] A. S. Walker-Andrews, L. E. Bahrick, S. S. Raglioni, and I. Diaz. "Infants' Bimodal Perception of Gender". In: *Ecological Psychology* 3.2 (1991), pp. 55–75. DOI: 10.1207/s15326969eco0302_1.

[482] S. Weech, S. Kenny, and M. Barnett-Cowan. "Presence and Cybersickness in Virtual Reality Are Negatively Related: A Review". In: *Frontiers in Psychology* 10 (2019). DOI: 10.3389/fpsyg.2019.00158.

[483] A. Weisser and J. M. Buchholz. "Conversational Speech Levels and Signal-to-Noise Ratios in Realistic Acoustic Conditions". In: *The Journal of the Acoustical Society of America* 145.1 (2019), pp. 349–360. DOI: 10.1121/1.5087567.

[484] A. Weisser, J. M. Buchholz, and G. Keidser. "Complex Acoustic Environments: Review, Framework, and Subjective Model". In: *Trends in Hearing* 23 (2019), p. 2331216519881346. DOI: 10.1177/2331216519881346.

[485] A. Weisser, J. M. Buchholz, C. Oreinos, J. Badajoz-Davila, J. Galloway, T. Beechey, and G. Keidser. "The Ambisonic Recordings of Typical Environments (ARTE) Database". In: *Acta Acustica united with Acustica* 105.4 (2019), pp. 695–713. DOI: 10.3813/AAA.919349.

[486] D. Wendt, R. K. Hietkamp, and T. Lunner. "Impact of Noise and Noise Reduction on Processing Effort: A Pupillometry Study". In: *Ear and Hearing* 38.6 (2017), p. 690. DOI: 10.1097/AUD.0000000000000454.

[487] D. Wendt, T. Koelewijn, P. Książek, S. E. Kramer, and T. Lunner. "Toward a More Comprehensive Understanding of the Impact of Masker Type and Signal-to-Noise Ratio on the Pupillary Response While Performing a Speech-in-Noise Test". In: *Hearing Research* 369 (2018), pp. 67–78. DOI: 10.1016/j.heares.2018.05.006.

[488] N. Wenk, J. Penalver-Andres, K. A. Buetler, T. Nef, R. M. Müri, and L. Marchal-Crespo. "Effect of Immersive Visualization Technologies on Cognitive Load, Moti-

vation, Usability, and Embodiment". In: *Virtual Reality* 27.1 (2023), pp. 307–331. DOI: 10.1007/s10055-021-00565-8.

[489] R. Whelan. "Effective Analysis of Reaction Time Data". In: *The Psychological Record* 58.3 (2008), pp. 475–482. DOI: 10.1007/BF03395630.

[490] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. "Welcome to the Tidyverse". In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: 10.21105/joss.01686.

[491] E. Wiedenmann, M. Daeglau, J. Otten, B. Mirkovic, G. Grimm, V. Hohmann, and S. Debener. "The Influence of Likeability Ratings of Audio-Visual Stimuli on Cortical Speech Tracking with Mobile EEG in Virtual Environments". In: *Fortschritte Der Akustik - DAGA 2023*. 2023.

[492] Wiederhold Brenda K, Davis Renee, and Wiederhold Mark D. "The Effects of Immersiveness on Physiology". In: *Studies in Health Technology and Informatics*. IOS Press, 1998. DOI: 10.3233/978-1-60750-902-8-52.

[493] M. Wilkinson, S. Brantley, and J. Feng. "A Mini Review of Presence and Immersion in Virtual Reality". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 65.1 (2021), pp. 1099–1103. DOI: 10.1177/1071181321651148.

[494] M. B. Winn, J. R. Edwards, and R. Y. Litovsky. "The Impact of Auditory Spectral Resolution on Listening Effort Revealed by Pupil Dilation". In: *Ear and Hearing* 36.4 (2015), e153. DOI: 10.1097/AUD.0000000000000145.

[495] B. G. Witmer and M. J. Singer. "Measuring Presence in Virtual Environments: A Presence Questionnaire". In: *Presence: Teleoperators and Virtual Environments* 7.3 (1998), pp. 225–240. DOI: 10.1162/105474698565686.

[496] I. Wohlgenannt, A. Simons, and S. Stieglitz. "Virtual Reality". In: *Business & Information Systems Engineering* 62.5 (2020), pp. 455–461. DOI: 10.1007/s12599-020-00658-9.

[497] F. Wolters, K. Smeds, E. Schmidt, E. K. Christensen, and C. Norup. "Common Sound Scenarios: A Context-Driven Categorization of Everyday Sound Environments for Application in Hearing-Device Research". In: *Journal of the American Academy of Audiology* 27.7 (2016), pp. 527–540. DOI: 10.3766/jaaa.15105.

[498] World Health Organization. *Report of the Informal Working Group on Prevention of Deafness and Hearing Impairment Programme Planning, Geneva, 18-21 June 1991*. PDH/91.1. 1991.

[499] Y.-H. Wu, N. Aksan, M. Rizzo, E. Stangl, X. Zhang, and R. Bentler. "Measuring Listening Effort: Driving Simulator Versus Simple Dual-Task Paradigm". In: *Ear and Hearing* 35.6 (2014), pp. 623–632. DOI: 10.1097/AUD.0000000000000079.

[500] M. Yadav, M. Georgi, L. Leist, M. Klatte, S. J. Schlittmeier, and J. Fels. "Cognitive Performance in Open-Plan Office Acoustic Simulations: Effects of Room

Acoustics and Semantics but Not Spatial Separation of Sound Sources". In: *Applied Acoustics* 211 (2023), p. 109559. DOI: 10.1016/j.apacoust.2023.109559.

[501] M. Yadav, J. Kim, D. Cabrera, and R. de Dear. "Auditory Distraction in Open-Plan Office Environments: The Effect of Multi-Talker Acoustics". In: *Applied Acoustics* 126 (2017), pp. 68–80. DOI: 10.1016/j.apacoust.2017.05.011.

[502] M. Zaglauer, H. Drotleff, and A. Liebl. "Background Babble in Open-Plan Offices: A Natural Masker of Disruptive Speech?" In: *Applied Acoustics* 118 (2017), pp. 1–7. DOI: 10.1016/j.apacoust.2016.11.004.

[503] A. A. Zekveld and S. E. Kramer. "Cognitive Processing Load across a Wide Range of Listening Conditions: Insights from Pupillometry". In: *Psychophysiology* 51.3 (2014), pp. 277–284. DOI: 10.1111/psyp.12151.

[504] S. Zojaji, A. Steed, and C. Peters. "Impact of Immersiveness on Persuasiveness, Politeness, and Social Adherence in Human-Agent Interactions within Small Groups". In: *ICAT-EGVE 2023 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. Ed. by J.-M. Normand, M. Sugimoto, and V. Sundstedt. The Eurographics Association, 2023. ISBN: 978-3-03868-218-9. DOI: 10.2312/egve.20231315.

[505] S. A. Zollinger and H. Brumm. "The Lombard Effect". In: *Current Biology* 21.16 (2011), R614–R615. DOI: 10.1016/j.cub.2011.06.003.

[506] F. Zotter and M. Frank. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Vol. 19. Springer Topics in Signal Processing. Cham: Springer International Publishing, 2019. ISBN: 978-3-030-17206-0. DOI: 10.1007/978-3-030-17207-7.

Curriculum Vitae

Personal Data

Cosima Antonia Ermert

September 21, 1997 born in Aachen, Germany

Education

July 10, 2026	Doctoral Examination for Dr.-Ing.
03/2021 – 07/2026	Doctoral Candidate RWTH Aachen University, Aachen, Germany
04/2019 – 02/2021	M. Sc. Electrical Engineering RWTH Aachen University, Aachen, Germany
10/2015 – 09/2018	B. Sc. Electrical Engineering RWTH Aachen University, Aachen, Germany
09/2007 – 06/2015	General Higher Education Entrance Qualification (Abitur) Heilig-Geist Gymnasium, Würselen, Germany

Work Experience

03/2021 – 06/2026	Research assistant Institute for Hearing Technology and Acoustics RWTH Aachen University, Aachen, Germany
10/2020 – 02/2021	Research Assistant (Internship) Oticon Centre for Applied Audiology Research Demant A/S, Smørum, Denmark
10/2016 – 02/2017	Student research assistant Clinic for Phoniatrics and Pediatric Audiology University Hospital Aachen University, Aachen, Germany

Aachen, Germany, September 3, 2026

Publications

Journal articles

[1] C. A. Ermert, A. Bönsch, T. W. Kuhlen, and J. Fels. "The impact of visual cues on speech intelligibility and listening effort in virtual reality". In: *tbd* (under Review).

[2] C. A. Ermert, L. Vollmer, H. Dewi, S. J. Schlittmeier, and J. Fels. "Spatial-changing irrelevant speech in visuo-spatial and visuo-verbal serial recall". In: *tbd* (under Review).

[3] C. A. Ermert, K. Loh, K. Baylan, K. W. Kühlem, A. Bönsch, T. W. Kuhlen, and J. Fels. "Listening Effort In Complex Audiovisual Scenes Under Varying Simulated Room Acoustic Conditions". In: *Trends in Hearing* 30 (2026). DOI: 10.1177/23312165261468224.

[4] C. A. Ermert, S. J. Schlittmeier, A. Bönsch, T. W. Kuhlen, and J. Fels. "Heard-Text Recall and Listening Effort under Irrelevant Speech and Pseudo-Speech in Virtual Reality". In: *Acta Acustica* 10 (2026), p. 7. ISSN: 2681-4617. DOI: 10.1051/aacus/2026003.

[5] J. Ehret, J. Schüppen, C. Mohanathasan, C. A. Ermert, J. Fels, S. J. Schlittmeier, T. W. Kuhlen, and A. Bönsch. "Objectifying Social Presence: Evaluating Multimodal Degraders in ECAs Using the Heard Text Recall Paradigm". In: *IEEE Transactions on Visualization and Computer Graphics* (2025), pp. 1–15. DOI: 10.1109/TVCG.2025.3636079.

[6] C. A. Ermert, M. Yadav, J. Ehret, C. Mohanathasan, A. Bönsch, T. W. Kuhlen, S. J. Schlittmeier, and J. Fels. "Audiovisual angle and voice incongruence do not affect audiovisual verbal short-term memory in virtual reality". In: *PLoS One* 20.8 (2025), e0330693. DOI: 10.1371/journal.pone.0330693.

[7] C. A. Ermert, M. Yadav, J. E. Marsh, S. J. Schlittmeier, T. W. Kuhlen, and J. Fels. "Serial recall in spatial acoustic environments: irrelevant sound effect and spatial source alternations". In: *Scientific Reports* 15.1 (2025). Publisher: Nature Publishing Group, p. 32473. ISSN: 2045-2322. DOI: 10.1038/s41598-025-18592-9.

[8] C. Mohanathasan, C. A. Ermert, J. Fels, T. W. Kuhlen, and S. J. Schlittmeier. "Exploring short-term memory and listening effort in two-talker conversations: The influence of soft and moderate background noise". In: *PLOS ONE* 20.2 (2025), e0318821. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0318821.

[9] S. Doma, C. A. Ermert, and J. Fels. “A magnitude-based parametric model predicting the audibility of HRTF variation”. In: *Journal of the Audio Engineering Society* 71.4 (2023), pp. 155–172. DOI: 10.17743/jaes.2022.0080.

Conference papers

[1] C. A. Ermert and J. Fels. “Listening Effort in Virtual Reality: The Role of Background Speech, Room Acoustics, and Talker Visualization”. In: *Fortschritte der Akustik - DAGA 2026*. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2026. DOI: 10.71568/daga2026.128.

[2] C. A. Ermert, J. Ehret, C. Mohanathasan, A. Bönsch, T. W. Kuhlen, S. J. Schlittmeier, and J. Fels. “Influence of (non-) intelligible background speech on memory and listening effort in conversational situations”. In: *Proceedings of DAS/DAGA 2025*. Ed. by T. Dau and B. Epp. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2025, pp. 159–160. ISBN: 978-3-939296-23-2. DOI: 10.71568/dasdaga2025.145.

[3] S. Doma, C. A. Ermert, and J. Fels. “On the monaural contribution to HRTF discrimination”. In: *Fortschritte der Akustik - DAGA 2023*. Ed. by O. von Estorff and S. Lippert. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2023. ISBN: 978-3-939296-21-8. URL: https://publications.rwth-aachen.de/record/956681.

[4] 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: *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. IVA ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1–8. ISBN: 978-1-4503-9994-4. DOI: 10.1145/3570945.3607312.

[5] J. Ehret, S. J. Schlittmeier, C. Mohanathasan, C. A. Ermert, J. Fels, and T. Kuhlen. “Towards Plausible Cognitive Research in Virtual Environments: The Effect of Audiovisual Cues on Short-Term Memory in Two-Talker Conversations”. In: *Proceedings of the 1st AUDICTIVE Conference*. Ed. by J. Balint and J. Fels. 2023. DOI: 10.18154/RWTH-2023-08409.

[6] C. A. Ermert, J. Ehret, T. Kuhlen, C. Mohanathasan, S. J. Schlittmeier, and J. Fels. “Audio-Visual Content Mismatches in the Serial Recall Paradigm”. In: *Fortschritte der Akustik - DAGA 2023*. Ed. by O. von Estorff and S. Lippert. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2023. ISBN: 978-3-939296-21-8. URL: https://publications.rwth-aachen.de/record/956451.

[7] C. Mohanathasan, J. Ehret, C. A. Ermert, J. Fels, T. Kuhlen, and S. J. Schlittmeier. “Towards More Realistic Listening Research in Virtual Environments: The Effect of Spatial Separation of Two Talkers in Conversations on Memory and Listening Effort”. In: *Fortschritte der Akustik - DAGA 2023*. Ed. by O. von Estorff and S. Lippert. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2023. ISBN: 978-3-939296-21-8. DOI: 10.18154/RWTH-2023-05116.

[8] C. A. Ermert, J. Ehret, T. Kuhlen, C. Mohanathasan, S. J. Schlittmeier, and J. Fels. “Spatial audio-visual congruency effects in virtual reality environments”.

In: *International Conference on Acoustics (ICA)*. Gyeongju (South Korea), 2022. URL: https://publications.rwth-aachen.de/record/888679.

[9] C. A. Ermert, C. Mohanathasan, S. J. Schlittmeier, and J. Fels. "Examining the Influence of Varying Audio Reproduction Methods on Verbal Short-Term Memory". In: *Fortschritte der Akustik - DAGA 2022*. Ed. by P. Leistner. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2022. ISBN: 978-3-939296-20-1. DOI: 10.18154/RWTH-2022-04013.

[10] L. J. Vollmer, J. Burger, C. A. Ermert, and J. Fels. "Stimulus Onset Asynchronies and Audiovisual Serial Recall Performance". In: *Fortschritte der Akustik - DAGA 2022*. Ed. by P. Leistner. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2022. ISBN: 978-3-939296-20-1. URL: https://publications.rwth-aachen.de/record/845415.

[11] C. A. Ermert, J. Fels, and S. Doma. "Examining the influence of controlled HRTF-variation on various distance metrics". In: *Fortschritte der Akustik - DAGA 2021*. Ed. by H. W. und Peter Balazs. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2021. ISBN: 978-3-939296-18-8. URL: https://publications.rwth-aachen.de/record/829378.

[12] J. Fels, C. A. Ermert, J. Ehret, C. Mohanathasan, A. Bönsch, T. Kuhlen, and S. J. Schlittmeier. "Listening to, and remembering conversations between two talkers: Cognitive research using embodied conversational agents in audiovisual virtual environments". In: *Fortschritte der Akustik - DAGA 2021*. Ed. by H. W. und Peter Balazs. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2021. ISBN: 978-3-939296-18-8. URL: https://publications.rwth-aachen.de/record/829385.

Oral presentations and Posters

[1] C. A. Ermert and J. Fels. "Simulating Real-World Aspects of Communication Situations". In: *53rd Erlangen Colloquium for audiological research and development*. Erlangen, Germany, 2026. URL: https://publications.rwth-aachen.de/record/1028834.

[2] C. A. Ermert, A. Bönsch, T. W. Kuhlen, and J. Fels. "The impact of coverbal visual cues on speech intelligibility and cognitive load in virtual reality environments". In: *The Journal of the Acoustical Society of America*. Vol. 157. 2025, A159. DOI: 10.1121/10.0037748.

[3] C. A. Ermert, K. Baylan, K. Loh, T. Kuhlen, A. Bönsch, K. W. Kühlem, and J. Fels. "Listening Effort In Populated Audiovisual Scenes Under Plausible Room Acoustic Conditions". In: *10th International Symposium on Auditory and Audiological Research - ISAAR 2025*. 2025. URL: https://publications.rwth-aachen.de/record/1023109.

[4] C. A. Ermert and J. Fels. "Bringing real-life into the lab: Investigating speech in virtual auditory environments". In: *Workshop in Communication in Hearing Science*. Lynge, Denmark, 2025. URL: https://publications.rwth-aachen.de/record/1028838.

[5] C. Mohanathasan, J. Ehret, A. Bönsch, C. A. Ermert, J. Fels, T. Kuhlen, P. B. Koleva, and S. J. Schlittmeier. "The Role of Talker Visibility on Memory and Listening Effort during Conversations". In: *Fortschritte der Akustik - DAGA 2024*. Ed. by J. Peissig and R. Rolfes. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2024. ISBN: 978-3-939296-22-5. DOI: 10.18154/RWTH-2024-03432.

[6] C. A. Ermert. "Short-Term Memory Performance under Gender and Spatial Audio-Visual Mismatch". In: *27th International Student Conference on Electrical Engineering - POSTER 2023*. 2023.

[7] C. Mohanathasan, J. Ehret, C. A. Ermert, J. Fels, T. Kuhlen, and S. J. Schlittmeier. "Memory and Listening Effort in Conversations: The Role of Spatial Cues and Cognitive Functions". In: *65. Tagung experimentell arbeitender Psycholog:innen - TeaP 2023*. 2023. DOI: 10.18154/RWTH-2023-05305.

[8] C. Mohanathasan, S. J. Schlittmeier, J. Fels, J. Ehret, C. A. Ermert, and T. Kuhlen. "Memory and Listening Effort in Two-Talker Conversations: Does Face Visibility Help Us Remember?" In: *Conference Name: 23. Conference of the European Society for Cognitive Psychology - ESCoP 2023*. 2023. DOI: 10.18154/RWTH-2023-08097.

[9] C. Mohanathasan, C. A. Ermert, J. Ehret, J. Fels, and S. J. Kuhlen Torsten andSchlittmeier. "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 (ESCOP)*. 2022. DOI: 10.18154/RWTH-2022-08380.

[10] C. A. Ermert, L. Xia, B. M. K. Loong, J. Fels, and S. Santurette. "Bringing ecological validity to the technical evaluation of hearing aids". In: *8th International Symposium on Auditory and Audiological Research - ISAAR 2021*. 2021.

Data publications

[1] C. A. Ermert, H. K. Dewi, K. Smeds, and J. Fels. *SEVQuest - A subjective questionnaire on ecological validity for participant studies*. Zenodo, 2026. DOI: 10.5281/zenodo.21062865.

[2] C. A. Ermert, C. Mohanathasan, J. Ehret, S. J. Schlittmeier, T. Kuhlen, and J. Fels. *AuViST - An Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm*. RWTH-2023-05543. RWTH Aachen University, 2023. DOI: 10.18154/RWTH-2023-05543.

Acknowledgements

Receiving my PhD marks the end of a long and eventful journey. Although the PhD itself may seem like the final goal, the journey toward it has been just as important to me - both personally and professionally. Along the way, many people accompanied me, supported me, and encouraged me. I would like to take this opportunity to thank them.

First and foremost, I would like to thank my supervisor, Janina Fels. A little more than five years ago, she saw potential in me and gave me the opportunity to conduct my research at IHTA. Throughout these years, she always supported my decisions and had confidence in my work. Thank you very much for your guidance and trust. I would also like to thank my co-examiner, Stephan Getzmann, for accompanying me during the examination process and for his valuable feedback.

My sincere thanks go to my fellow travellers at IHTA. Whenever I had a question, I could always count on receiving advice, encouragement, or simply an open ear. I would like to thank Karin, Ellen, Kim, and Emilie for helping me with all my organisational questions, and Rolf, Richard, Norbert, Marc, and Thomas for their support with technical matters. I am furthermore very grateful to my wonderful scientific colleagues; especially Julia, Caro, Karin, Jamilla, Lukas V., Shaimaa, Pascal, Anne, Simon, Philipp, Marco, Hark, Manuj, Michael, and Lukas A. Thank you for the many discussions, your support, and the time we shared. Further thanks go to Marco, Pascal, and Anne for joining me in my Delicious-Acoustic shenanigans. I would also like to thank my students, especially Hera and Karl. You were a tremendous help, and I truly enjoyed working with you. All of you made my time at IHTA very special and enjoyable.

I would also like to thank my project partners in Audictive: Jonathan, Andi, Thorsten, Chinthusa, Isabel, and Sabine. I learned a lot from you, and I am proud of what we achieved together. Further thanks go to my co-authors and to the people who advised and mentored me during my PhD, namely Axel, Karolina, and Veronika for their valuable input and support.

Last but not least, I would like to thank my friends and family, who were always there for me. Special thanks go to Leonie and Stefan, Anna, Nina and Julian, and Gregori and Wendy. Thank you for your participation, friendship, and encouragement.

I would like to express my deepest gratitude to my parents. Thank you for always believing in me and for always being there for me. Without you, I would not have become the person I am today.

And Marius - thank you for everything. I know that you will always stand behind me, and this means the world to me. I could not have completed this journey without you.

Dissertations at the Institute of Technical Acoustics and the Institute for Hearing Technology and Acoustics since 1998

1. Fedtke, Thomas
   Breitbandige Methoden zur Bestimmung der komplexen Übertragungsfunktion von Mikrophonen
   ISBN: 3-89701-244-8

2. Müller, Swen
   Digitale Signalverarbeitung für Lautsprecher
   `https://publications.rwth-aachen.de/record/56081`

3. Meier, Andreas
   Die Bedeutung des Verlustfaktors bei der Bestimmung der Schalldämmung im Prüfstand
   `https://publications.rwth-aachen.de/record/62169`

4. Mauer, Günter
   Messungen und Modellbildung zur Untersuchung der amplitude modulation following response
   `https://publications.rwth-aachen.de/record/56646`

5. Kob, Malte
   Physical modeling of the singing voice
   `https://publications.rwth-aachen.de/record/59935`

6. Bartsch, Guido
   Effiziente Methoden für die niederfrequente Schallfeldsimulation
   ISBN: 978-3-936328-07-3

7. Abou-Elleal, Esam Ragab Esmail
Raumakustik - Interaktion visueller und auditiver Wahrnehmungen
https://publications.rwth-aachen.de/record/58908

8. Stephenson, Uwe Martin
Beugungssimulation ohne Rechenzeitexplosion: Die Methode der quantisierten Pyramidenstrahlen: ein neues Berechnungsverfahren für Raumakustik und Lärmimmissionsprognose
https://publications.rwth-aachen.de/record/62024

9. Klemenz, Martin
Die Geräuschqualität bei der Anfahrt elektrischer Schienenfahrzeuge
https://publications.rwth-aachen.de/record/59680

10. Thaden, Rainer
Auralisation in building acoustics
https://publications.rwth-aachen.de/record/59629

11. Makarski, Michael
Tools for the Professional Development of Horn Loudspeakers
https://publications.rwth-aachen.de/record/61507

12. Diniz Guimarães, João Henrique
Modelling the dynamic interactions of rolling bearings
https://publications.rwth-aachen.de/record/50357

13. Lentz, Tobias
Binaural technology for virtual reality
https://publications.rwth-aachen.de/record/51299

14. Fels, Janina
From children to adults: How binaural cues and ear canal impedances grow
https://publications.rwth-aachen.de/record/51298

15. Kling, Christoph
Investigations into damping in building acoustics by use of downscaled models
https://publications.rwth-aachen.de/record/94549

16. Franck, Andreas
Finite-Elemente-Methoden, Lösungsalgorithmen und Werkzeuge für die akustische Simulationstechnik
https://publications.rwth-aachen.de/record/51586

17. Fingerhuth, Sebastián
Tonalness and consonance of technical sounds
https://publications.rwth-aachen.de/record/51771

18. Schröder, Dirk
Physically Based Real-Time Auralization of Interactive Virtual Environments
https://publications.rwth-aachen.de/record/50580

19. Aretz, Marc
Combined Wave And Ray Based Room Acoustic Simulations Of Small Rooms
https://publications.rwth-aachen.de/record/197461

20. Sanches Masiero, Bruno
Individualized binaural technology: measurement, equalization and perceptual evaluation
https://publications.rwth-aachen.de/record/197486

21. Shabalina, Elena D.
The Propagation of Low Frequency Sound through an Audience
https://publications.rwth-aachen.de/record/229210

22. Lievens, Matthias
Structure-borne sound sources in buildings
https://publications.rwth-aachen.de/record/229320

23. Feistel, Stefan
Modeling the Radiation of Modern Sound Reinforcement Systems in High Resolution
https://publications.rwth-aachen.de/record/444852

24. Scharrer, Roman
Acoustic field analysis in small microphone arrays
https://publications.rwth-aachen.de/record/229274

25. Dietrich, Pascal
Uncertainties in Acoustical Transfer Functions: Modeling, Measurement and Derivation of Parameters for Airborne and Structure-borne Sound
https://publications.rwth-aachen.de/record/229384

26. Wang, Xun
Model based signal enhancement for impulse response measurement
https://publications.rwth-aachen.de/record/229544

27. Vitale, Renzo
Perceptual aspects of sound scattering in concert halls
https://publications.rwth-aachen.de/record/479369

28. Wefers, Frank
Partitioned convolution algorithms for real-time auralization
https://doi.org/10.18154/RWTH-2015-02185

29. Pollow, Martin
Directivity patterns for room acoustical measurements and simulations
https://publications.rwth-aachen.de/record/483805

30. Müller-Trapet, Markus
Measurement of surface reflection properties : concepts and uncertainties
https://doi.org/10.18154/RWTH-2015-05681

31. Guski, Martin
Influences of external error sources on measurements of room acoustic parameters
https://publications.rwth-aachen.de/record/564354

32. Nau, Clemens
Beamforming in modalen Schallfeldern von Fahrzeuginnenräumen
https://doi.org/10.18154/RWTH-2016-11927

33. Philippen, Bernd
Transfer path analysis based on in-situ measurements for automotive applications
ISBN: 978-3-8325-4435-5

34. Mohamady, Samira
Uncertainties of transfer path analysis and sound design for electrical drives
ISBN: 978-3-8325-4431-7

35. Bomhardt, Ramona
Anthropometric individualization of head-related transfer functions analysis and modeling
ISBN: 978-3-8325-4543-7

36. Meng, Fanyu
Modeling of moving sound sources based on array measurements
https://doi.org/10.18154/RWTH-2018-227109

37. Richter, Jan-Gerrit
Fast measurement of individual head-related transfer functions
https://doi.org/10.18154/RWTH-2019-04006

38. Viveros Munoz, Rhoddy Angel
Speech perception in complex acoustic environments: evaluating moving maskers using virtual acoustics
https://doi.org/10.18154/RWTH-2019-07497

39. Brezas, Spyros
Investigation on the dissemination of unit watt in airborne sound and applications
https://doi.org/10.18154/RWTH-2019-10636

40. Oberem, Josefa
Examining auditory selective attention: From dichotic towards realistic environments
https://doi.org/10.18154/RWTH-2020-04889

41. Klein, Johannes Christian
Directional room impulse response measurement
https://doi.org/10.18154/RWTH-2020-07829

42. Aspöck, Lukas
Validation of room acoustic simulation models
https://doi.org/10.18154/RWTH-2020-12146

43. Engel, Margret Sibylle
Determination of the soundscape cost index through soundscape data collection and the use of psychoacoustic metrics
https://doi.org/10.18154/RWTH-2022-01435

44. Pausch, Florian
Spatial audio reproduction for hearing aid research: System design, evaluation and application
https://doi.org/10.18154/RWTH-2022-03745

45. Witew, Ingo B.
Measurements in room acoustics: Uncertainties and influence of the measurement position
https://doi.org/10.18154/RWTH-2022-08992

46. Imran, Muhammad
Virtual building acoustics: auralization with contextual and interactive features
https://doi.org/10.18154/RWTH-2022-08992

47. Stienen, Jonas
Real-time auralisation of outdoor sound propagation
https://doi.org/10.18154/RWTH-2023-04171

48. Knappe, Fabian
Measurements in room acoustics: Uncertainties and influence of the measurement position
https://doi.org/10.18154/RWTH-2023-03062

49. Gomes Lobato, Thiago Henrique
Improving sound source localization and characterization with machine learning techniques
https://doi.org/10.18154/RWTH-2024-04316

50. Abi Raad, Elie
Real-time monitoring of ultrasonic metal welding: using horn and anvil vibrations and sound measurements
https://doi.org/10.18154/RWTH-2025-01322

51. Schäfer, Philipp
Real-time auralization of aircraft noise considering curved sound propagation
https://doi.org/10.18154/RWTH-2025-00309

52. Kohnen, Michael
Combining reproduction techniques for room auralizations
https://doi.org/10.18154/RWTH-2025-01606

53. Loh, Karin
Noise exposure in pre- and primary schools: exploring noise effects on young children
https://doi.org/10.18154/RWTH-2025-05615

54. Berzborn, Marco
Measurement and quantification of directional sound field decay
https://doi.org/10.18154/RWTH-2025-06560

55. Dreier, Christian Georg
Virtual Acoustics: Synthesis of Traffic Sound Sources
https://doi.org/10.18154/RWTH-2025-06710

56. Doma, Shaimaa
Numerical and perceptual evaluation of HRTF accuracy
https://doi.org/10.18154/RWTH-2025-07026

57. Kersten, Simon
Modeling and analysis of vibroacoustic mechanisms in hearing
https://doi.org/10.18154/RWTH-2025-07039

58. Seitz, Julia
Measuring listening effort in realistic classroom situations
https://publications.rwth-aachen.de/record/1037287

59. Breuer, Carolin
Evaluating Auditory Selective Attention in Classroom Scenarios using Audiovisual Virtual Reality – From Auditory to Audiovisual Scenarios
https://doi.org/10.18154/RWTH-2026-07159

60. Ermert, Cosima Antonia
Simulating complex listening scenarios in audiovisual virtual reality - The impact of audiovisual cues, spatial noise, and room acoustics on listening performance
https://doi.org/10.18154/RWTH-2026-07781