# Towards Plausible Cognitive Research in Virtual Environments: The Effect of Audiovisual Cues on Short-Term Memory in Two-Talker Conversations

Jonathan Ehret*[1], Cosima A. Ermert*[2], Chinthusa Mohanathasan*[3], Janina Fels[2], Torsten W. Kuhlen[1], Sabine J. Schlittmeier[3]

[1] *Visual Computing Institute, RWTH Aachen University, Germany*

[2] *Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany*

[3] *Work and Engineering Psychology, RWTH Aachen University, Germany*

[*] *These authors contributed equally to this work.*

## Introduction

When three or more people are involved in a conversation, often one conversational partner listens to what the others are saying and has to remember the conversational content. However, setups in cognitive-psychological experiments often differ substantially from everyday listening situations by neglecting audiovisual cues going beyond the pure speech signals. The presence of speech-related audiovisual cues, such as the spatial position, and the appearance or non-verbal behavior of the conversing talkers, may influence the listener's memory and comprehension of the conversational content. In our project, we provide first insights into the contribution of acoustic and visual cues on verbal short-term memory (STM), and (social) presence. First, we discuss the auditory verbal serial recall (aVSR) task and the heard text recall (HTR) paradigm, which can be used for quantifying STM performance. In several experiments, we investigated how STM performance in conversational settings varies with increasingly more plausible audiovisual characteristics, e.g., spatial auditory cues. Furthermore, we explored the influence of the display device (head-mounted display (HMD) vs. traditional computer monitor) and audiovisual mismatches as multimodal distractors. Adding virtual embodiments (i.e., embodied conversational agents (ECAs)) for the talkers allowed us to conduct experiments on the influence of the fidelity of co-verbal gestures and turn-taking signals (i.e., co-verbal cues to communicate whether a speaker wants to continue speaking or pass the turn on to another speaker). Preliminary results from these experiments will be presented. To conclude this paper, we discuss the insights gained so far and outline further research directions.

## Paradigms

A standard cognitive measure of verbal STM is the well-established aVSR task. In the aVSR task, the digits from 1-9 are presented in random order and need to be recalled in the exact order in which they were presented after a short retention interval [1].

A verbal STM task of undoubtedly higher complexity is the newly developed HTR paradigm. In this novel task, called HTR [2], a coherent text is presented auditorily as a conversation between two talkers followed by questions based on the content of the conversation.

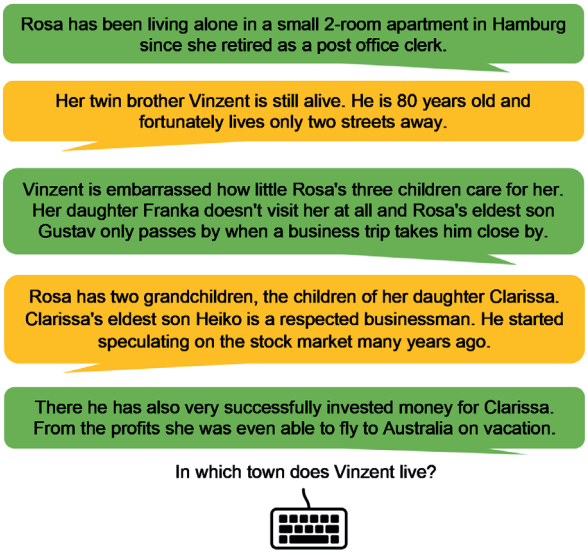


Figure 1: Example of a conversation and one question in the heard text recall task.

The texts describe three generations of a family (grandparents, parents, and children) while also taking into account various aspects of the family members (e.g., place of residence, hobbies, professions). Each text consists of ten sentences and is presented as a conversation between one talker with a female voice and another talker with a male voice (see Figure 1). The turn-taking between the female and the male talker aims to simulate a natural conversation, so sentences linked closely together are spoken by the same conversational partner. For each text, nine corresponding questions have to be answered. The questions ask for names of family members, relations between family members, and other information (profession, hobbies, age, etc.). We recorded speech material and lip movements for the HTR task in the Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm (AuViST) [3]. The HTR is administered either alone as a single task or as a primary listening task in a dual-task paradigm. In such a paradigm, participants are asked to perform two tasks in parallel, a primary task and an unrelated secondary task. With the help of the dual-task paradigm we are able to measure listening effort and thus, respective variations in different experimental conditions. For a more

detailed definition of the dual-task paradigm and listening effort see [2, 4, 5]. As our research showed [6], a suitable secondary task for the HTR is a vibrotactile pattern recognition task (cp. [7]). Here, participants hold a Sony Dualshock 4 controller in their hands, which emanates four vibration patterns (short-short, long-long, short-long, and long-short). Participants have to indicate whether the presented vibration patterns are similar (e.g., short-short) or different (e.g., short-long) by pressing the circle or left arrow button on the game controller. The cognitive tasks were used in a series of experiments where the audiovisual plausibility was systematically increased.

## Auditory Scene Design

So far, STM performance has predominantly been investigated with simplified auralization, e.g., by playing back target and irrelevant signals in a diotic manner. In reality, auditory signals are - with few exceptions - accompanied by spatial information. In acoustic research, there are multiple possibilities on how to obtain spatial sound signals, e.g., via loudspeaker playback or with headphone reproduction using generic or individual head-related transfer functions (HRTFs). HRTFs capture the frequency response function of the human head and can therefore provide plausible binaural signals. In a first study, we investigated the influence of spatially distributed target sources with different spatial sound reproduction techniques on the aVSR paradigm [17]. We could not find an influence of the reproduction technique or the spatial distribution. In a second study, we investigated the effect of spatial separation of the two conversing talkers on the listener's STM and comprehension in the HTR task [4]. Participants performed the HTR as a primary listening task and the vibrotactile pattern recognition task as a secondary task. In the primary listening task participants listened to the HTR conversations in a non-noisy setting. The two alternating talkers' audio signals were presented at a distance of 2.5 m from the listener either spatially separated (+/- 60°) or co-located (0°). Again, we found no variation in memory performance between the two talkers' audio conditions. That means participants did not remember more information when the two talkers were spatially separated than when they were co-located. It is possible, that the investigated settings are too simple. Research has shown, that binaural cues become more relevant in challenging listening settings [8]. In further experiments, we plan to extend the complexity of the acoustic scene, e.g., by adding more distraction sources and room acoustics.

## Visual Conversational Setting

In real life, information is usually perceived multimodally. For example, if we hear a person speak, we see their lips moving at the same time, thus the information is received audiovisually. When aiming to investigate STM performance in realistic settings, information should be presented in multiple modalities. We developed a virtual living room scene with two ECAs telling the HTR stories (see Figure 2), which poses a more plausible setting. We utilized the Unreal Engine 4.27 and MetaHumans [24]. When using embodied virtual representations for the talkers, all modalities (e.g., speech sound quality, visual fidelity, and motion faithfulness) must reach a consistent degree of realism. To this end, we added co-verbal behavior: co-verbal gestures recorded specifically for each sentence of the HTR texts using a self-developed Motion Capture Plugin [25], which utilizes off-the-shelf virtual reality (VR) hardware components; and a realistic gaze model inspired by Pejsa et al. [21]. Furthermore, turn-taking cues [22] (i.e., multimodal signals whether a speaker wants to continue speaking (turn-hold) or wants to pass on the turn to another speaker (turn-yield)) are normally generated in such a conversational setting. We developed a system generating non-verbal turn-taking cues by manipulating the gestures, the gazing, and adding appropriate inhalation sounds [20]. In an evaluation we found that manipulating the gestures out of these modalities was the only effective one to generate convincing turn-taking, however, no changes in the perception of the ECAs was found. With the knowledge gained, we plan to improve this system further to use it in upcoming experiments. Furthermore, we are currently evaluating different manipulations of the co-verbal gestures recorded to (i) improve imperfections due to the recording technology but also (ii) deliberately create misaligned gestures to evaluate their influence on HTR task performance and perceived social presence. Social presence hereby describes the feeling of being in the presence of and interacting with a real person (cf. [23]). We hypothesize that degraded co-verbal gestures will decrease perceived social presence and want to evaluate whether cognitive load during the HTR task can be used as a proxy for social presence. As developing experiments for VR requires additional implementation overhead, we designed a plugin for simpler creation and control of factorial-design studies in the Unreal Engine, which we made publicly available [26]. Furthermore, with this plugin less technical knowledge is required to implement and conduct a

Figure 2: Two embodied conversational agents telling a family story from the heard text recall paradigm, as viewed by the participant. The embodied conversational agents show co-verbal gestures and conversational gaze patterns, which were manipulated to additionally give turn-taking cues [20].

study while the risk of missing crucial flaws is reduced.

## Audiovisual Interaction

As stated before, in real-life we are usually confronted with information from multiple modalities. This does not only apply to target signals, i.e., stimuli we want to focus on, but also irrelevant signals. It has been shown that STM performance can decrease in the presence of irrelevant background sound. For example, the influence of auditory noise on verbal short-term memory performance, i.e., the irrelevant sound effect (ISE), has been investigated to a great extent in cognitive psychology, e.g. [9, 10, 11, 12]. However, there is only very little research on the influence of irrelevant visual information on task performance (e.g., [13, 14]), or irrelevant audiovisual signals. To bridge this gap, we investigated the influence of audiovisual mismatches on STM with audiovisual target presentation. We conducted two listening experiments: one with audiovisual spatial mismatches [17] and one with an audiovisual gender mismatch [18]. To quantify STM performance for heard speech, the aVSR paradigm was employed.

In recent years, using VR scenes presented on HMDs has become increasingly popular in research, as it allows for creating a close-to-real-life, but still controllable, experimental environment. As a step towards more realistic listening experiment settings, we investigated the influence of the display device, namely HMD and computer monitor reproduction, as an additional variable in our experiments.

No effects of the display device or the audiovisual mismatch on STM performance could be detected in either experiment. However, large audiovisual spatial mismatches lead to larger reaction times. We hypothesize that the chosen mismatches did not provide enough distracting potential, as they were static. Future research should focus on whether similar to how the ISE increases with increasing fluctuation strength [12], a possible impact of audiovisual mismatches becomes observable with more dynamic stimuli.

## Conclusion

With our work, we can provide a more plausible paradigm for investigating memory for two-talker conversations. We implemented an audiovisual virtual environment to evaluate comprehension and recall of two-talker conversations by utilizing the HTR paradigm in a close-to-real but fully controllable environment. Our experiments contribute to a growing body of studies that conduct auditory research in more plausible settings both in terms of the listener's task and audiovisual signals. In future research, we plan to use the results regarding co-verbal gestures and turn-taking cues to further investigate whether objective metrics like HTR and dual-task performance (the second to measure cognitive load) can be used as a proxy for social presence (which is currently often evaluated using subject questionnaire-based metrics). To extend the realism of the provided scene even further, we will investigate the influence of complex auditory scenes with regard to sound source distribution and room acoustics. We will also test the potential utility - or even necessity - of an immersive 3D presentation for cognitive psychology research on memory for two-talker conversations. For this purpose, we plan to evaluate further whether immersion is an important factor for these kinds of experiments comparing computer monitor, HMD, and cave automatic virtual environment (CAVE) renderings.

## Acknowledgments

This research was funded by the German Research Foundation (DFG) within the project "Listening to, and remembering conversations between two talkers: Cognitive research using embodied conversational agents in audiovisual virtual environments", which is part of the DFG Priority Program "AUDICTIVE" (SPP 2236). The contribution by Sabine Schlittmeier was supported by a grant from the Head-Genuit-Foundation (P-16/10-W).

## References

[1] Schlittmeier, S.J., Mohanathasan, C. & Fintor, E. (2021) Paradigm for Measuring Verbal Short-Term Memory Capacity for Auditorily or Visually Presented Items: Verbal Serial Recall Task. https://doi.org/10.18154/RWTH-2021-09604

[2] Fintor, E., Aspöck, L., Fels, J., & Schlittmeier, S. J. (2022). The role of spatial separation of two talkers' auditory stimuli in the listener's memory of running speech: listening effort in a non-noisy conversational setting. *International Journal of Audiology, 61*(5), 371-379. https://doi.org/10.1080/14992027.2021.1922765

[3] Ermert, C. A., Mohanathasan, C., Ehret, J., Schlittmeier, S. J., Kuhlen, T. & Fels, J. (2022) AuViST - An Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm. https://doi.org/10.18154/RWTH-2022-10851

[4] Mohanathasan, C., Ehret, J., Ermert, C.A., Fels, J., Kuhlen, T. & Schlittmeier, S.J. (2023). Towards More Realistic Listening Research in Virtual Environments: The Effect of Spatial Separation of Two Talkers in Conversations on Memory and Listening Effort. *Tagungsband DAGA 2023: 49. Jahrestagung für Akustik*, Hamburg (Germany).

[5] Gagné, J. P., Besser, J., & Lemke, U. (2017). Behavioral Assessment of Listening Effort Using a Dual-Task Paradigm. *Trends in hearing*, 21.

[6] Mohanathasan, C., Ermert, C.A., Ehret, J., Fels, J., Kuhlen, T. & Schlittmeier, S. (2022). Measuring Listening Effort in Adverse Listening Conditions: Testing Two Dual Task Paradigms for Upcoming Audiovisual Virtual Reality Experiments. *The 22nd conference of the European Society for Cognitive Psychology (ESCoP 2022)*, Lille, France.

[7] Gosselin, P. A. & Gagné, J. P. (2011). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. *International Journal of Audiology, 50*(11), 786-792.

[8] Oberem, J., Lawo, V., Koch, I. & Fels, J. (2014). Intentional Switching in Auditory Selective Attention: Exploring Different Binaural Reproduction Methods in an Anechoic Chamber. *Acta Acustica united with Acustica, 100*(6), 1139-1148.

[9] Salamé, P., & Baddeley, A. (1989). Effects of Background Music on Phonological Short-Term Memory. *The Quarterly Journal of Experimental Psychology Section A, 41*(1), 107-122. https://doi.org/10.1080/14640748908402355

[10] Surprenant, A.M. (1999), The Effect of Noise on Memory for Spoken Syllables. *International Journal of Psychology, 34*, 328-333. https://doi.org/10.1080/002075999399648

[11] Schlittmeier, S.J. & Hellbrück, J. (2009), Background music as noise abatement in open-plan offices: A laboratory study on performance effects and subjective preferences. *Appl. Cognit. Psychol., 23*, 684-697. https://doi.org/10.1002/acp.149

[12] Schlittmeier, S., Weißgerber, T., Kerber, S., Fastl, H., & Hellbrück, J. (2012). Algorithmic modeling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength. Attention, *Perception & Psychophysics, 74*(1), 194-203. https://doi.org/10.3758/s13414-011-0230-7

[13] Liebl, A., Haller, J., Jödicke, B., Baumgartner, H., Schlittmeier, S., & Hellbrück, J. (2012). Combined effects of acoustic and visual distraction on cognitive performance and well-being. *Applied Ergonomics, 43*(2), 424-434. https://doi.org/10.1016/j.apergo.2011.06.017

[14] Lange, E. (2005). Disruption of attention by irrelevant stimuli in serial recall. *Journal of Memory and Language, 53*(4), 513-531. https://doi.org/10.1016/j.jml.2005.07.002

[15] Li, G., Anguera, J. A., Javed, S., Khan, M. I., Wang, G., & Gazzaley, A. (2020). Enhanced Attention Using Head-mounted Virtual Reality. *Journal of Cognitive Neuroscience, 32*(8), 1438-1454. https://doi.org/10.1162/jocn_a_01560

[16] Wan, B., Wang, Q., Su, K., Dong, C., Song, W., & Pang, M. (2021). Measuring the Impacts of Virtual Reality Games on Cognitive Ability Using EEG Signals and Game Performance Data. *IEEE Access, 9*, 18326-18344. https://doi.org/10.1109/access.2021.3053621

[17] Ermert, C.A., Ehret, J., Kuhlen, T., Mohanathasan, C., Schlittmeier, S.J., Fels, J. (2022). Spatial Audio-Visual Congruency Effects in Virtual Reality Environments. *Proceedings on the 24th International Congress on Acoustics, Gyeongju (South Korea)*, ABS-0227.

[18] Ermert, C.A., Ehret, J., Kuhlen, T., Mohanathasan, C., Schlittmeier, S.J., Fels, J. (2023). Audio-visual Content Mismatches in the Serial Recall Paradigm. *DAGA 2023: 49. Jahrestagung für Akustik, Hamburg (Germany).*

[19] Hurlstone, M. J. (in press). Serial recall. In M. J. Kahana & A. Wagner (Eds.), *The Oxford handbook of human memory*. Oxford University Press

[20] Ehret, J., Bönsch, A., Nossol, P., Ermert, C.A., Mohanathasan, C., Schlittmeier, S.J., Fels, J., & Kuhlen, T.W. (2023). Who's next? Integrating Non-Verbal Turn-Taking Cues for Embodied Conversational Agents. *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (under review).

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

[22] Skantze, G. (2021). Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech and Language 67*, 101178. https://doi.org/10.1016/J.CSL.2020.101178

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

[24] MetaHumans by Epic, URL: https://www.unrealengine.com/metahuman

[25] Full-Body Motion Capture Plugin, URL: https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/MoCapPlugin

[26] Study Framework Plugin, URL: https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/unreal-study-framework