Listening to, and Remembering Conversations between Two Talkers: Cognitive Research using Embodied Conversational Agents in Audiovisual Virtual Environments

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

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

[2] *Visual Computing Institute, RWTH Aachen University*

[3] *Teaching & Research Area Work and Engineering Psychology, RWTH Aachen University*

Email: janina.fels@akustik.rwth-aachen.de

Introduction

In everyday life, we often encounter conversational situations consisting of three or more people, where one person is mainly a listener while others speak. The listener has to memorize what is said by the talkers in order to join the conversation as a talker himself, or simply remember the content for later use.

Listening to and remembering the content of spoken language is a demanding task from a cognitive-psychological view. Unlike for example reading a text, all the cognitive processes involved in evaluating, remembering, and understanding spoken language must be performed on the basis of a single, time-limited presentation [1]. It can be assumed that the availability of audiovisual information is likely to affect listener's memory and comprehension of the speech content. This includes auditory-perceptual features such as the spatial position of the speakers and the fundamental frequency of the voice (which in many cases indicates gender), visual information such as the appearance of the speaker or co-verbal behavior. For example, past research has shown that certain auditory-perceptive characteristics (e.g., type and level of background noise or masking source positions) can impede listening performance [2, 3, 4]. The availability of plausible auditory cues has been shown to be beneficial for challenging listening situations [5, 6] and visual information like gestures and lip movement, which accompany the spoken word, can improve speech comprehension [7]. Therefore, the hypothesis can be formed, that cognitive performance in conversational situations can be influenced by plausible auditory and visual information.

However, this particular setting where one person listens to running speech from two talkers and has to memorize what is said, has hardly been researched from a cognitive-psychological point of view. In fact, when memory performance for heard speech is studied in experiments, often simplified laboratory settings are employed, where unrelated digits (serial recall task) or isolated sentences (listening span task) are presented as stimulus material instead of running speech, with simplified headphone reproduction and no or only simple visualization (e.g., [8, 9]).

The aim of this project is to gain insights on the influence of audiovisual information on cognitive performance in close-to-realistic virtual environments. To maintain control in the listening experiments while having an almost natural frame, a computer-mediated conversation, e.g., a virtual reality (VR) based setting has proven to have a high ecological validity. Here, embodied virtual

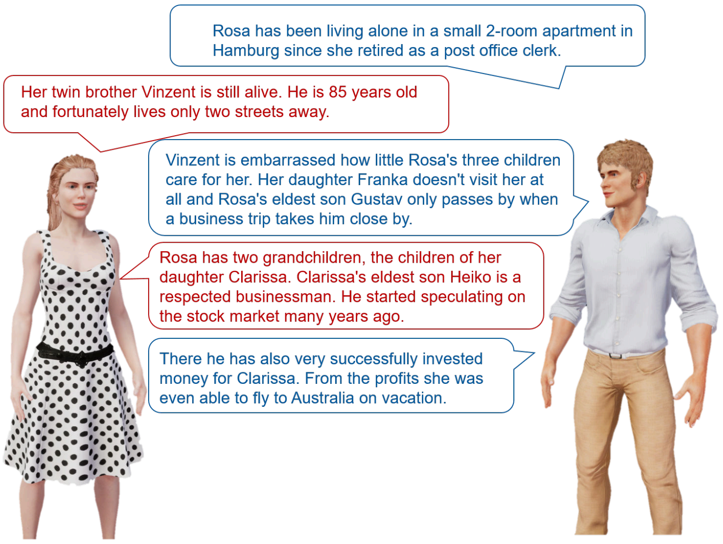
Rosa has been living alone in a small 2-room apartment in Hamburg since she retired as a post office clerk.
Her twin brother Vinzent is still alive. He is 85 years old and fortunately lives only two streets away.
Vinzent is embarrassed how little Rosa's three children care for her. Her daughter Franka doesn't visit her at all and Rosa's eldest son Gustav only passes by when a business trip takes him close by.
Rosa has two grandchildren, the children of her daughter Clarissa. Clarissa's eldest son Heiko is a respected businessman. He started speculating on the stock market many years ago.
There he has also very successfully invested money for Clarissa. From the profits she was even able to fly to Australia on vacation.

Figure 1: Example of a conversational situation in the heard text recall (HTR) paradigm with two virtual agents (VAs) as virtual interlocutors to whom a human addressee listens.

agents (VAs) replace the natural interlocutors, while the users as natural listeners remain as participants.

The heard text recall (HTR) task by Fintor et al. [10] (see Fig. 1) allows for examining listening performance with running speech as a stimulus. Starting from a simple experimental setup (e.g., Fig. 2) the realism of this task will be gradually increased in this project both from an acoustical and a visual point of view, e.g., via incorporating spatial acoustics and creating a VR environment. Nonverbal cues will be provided by VAs. Numerous research was conducted on how to efficiently visually and acoustically simulate and integrate VAs in VR environments, e.g., to spark the feeling of social presence, meaning the feeling of being in the presence and interacting with an actual human being [11, 12, 13].

The results of this project should give insights on which audiovisual characteristics of a more life-like conversational situation in VR environments positively affect subjective experience and cognitive performance, in terms of memory and comprehension for multi-talker speech. Furthermore, the findings could give insights on whether cognitive performance can be used as a quality criterion for virtual scene design, e.g., audiovisual plausibility [14], and social presence. This, however, only applies if the results show, that the degree of realism and the concept of social presence is correlated to memory performance. The underlying hypothesis is that if something in the performance of a virtual agent seems unnatural, this will lower the perceived social presence and also bind cognitive resources, which will lead to a decreased performance. The same applies to aspects of auralization.

This paper gives an overview over the project concept. The first section discusses a well-established recall task, the auditory verbal serial recall (aVSR), as well as the extended HTR paradigm. Afterwards, possibilities of including plausible auralization and visualization are presented. The paper closes with a summary and an outlook.

Recall Task for Cognitive Research

The state of the art for examining listening in every-day situations is the auditory verbal serial recall (aVSR) task. In this task, a list of stimuli is presented and has to be recalled by the participants after a short retention interval. There are multiple possibilities for stimuli as well as input and output modalities. A well-established configuration is using digits from 1-9 as target stimulus which is presented as audio and via a computer screen (e.g., [9]). Such simple laboratory setups differ substantially from everyday listening situations in many aspects: the target signal is not running speech, simple mono- or stereo signals are used neglecting room effects or head-related impulse responses, and the visualization is oftentimes limited to showing only the digits on a computer screen without the talker or the situational context. Thus, it is possible that this simplified task challenges different cognitive functions or processes than a realistic conversational situation.

Using running speech as the target signal, the heard text recall (HTR) paradigm [10] steps closer to realistic listening environments. In this task, a coherent text is presented as a conversation between two speakers in which the speakers take turns and the participant is the listener to this conversation. Content-wise, family constellations (e.g., grandparents, parents, children) are described and further information about the family members (e.g., place of residence, hobbies, occupation) is given. Some information can only be derived by combining several utterances (e.g., Rosa's age in Fig. 1)

In the experiment, the participant's task is to listen attentively to several conversations and to answer content-related questions about them afterwards. As a result, by using coherent speech as a stimulus material and having two talkers who take turns, a more realistic listening scenario is created compared to the aVSR.

Recall and comprehension can be directly measured by asking the listener questions about the content of the conversation and judging the given answer (right/wrong) [10]. Listening effort, which refers to the amount of processing resources a listener allocates to understand running speech in a noisy environment and/or to achieve high performance in a listening-related task [16], cannot be assessed by such direct measures.

When the same task is performed under various conditions, e.g., different background noise level, it is possible to obtain similar performance in the HTR while varying listening effort is required to achieve this level of performance (e.g., [17]). Direct measurements alone are often not sensitive to small experimental variations, i.e., they do not capture whether different listening effort in certain listening conditions was necessary to achieve similar performance in the listening task.

This inaccuracy can be counteracted by using a dual-task design. Here, participants have to perform an unrelated secondary task in parallel to the HTR. Examples for such a dual task can be judging numbers in comparison to a threshold (e.g., [18]) or a tactile stimulation task (e.g., [19]). In general, a listening environment which is more demanding on cognitive resources can be identified if the performance in this second task decreases when changing the experimental conditions, and at the same time the error rate in answering the questions stays the same [9].

Influence of Audiovisual Information on Memory Performance in Conversational Situations

In face-to-face conversations, not only the spoken word is usually heard, but additional audiovisual information is available to accompany or even enrich the conversation, such as mouth movements and gestures. In the scope of this project, these information will be integrated into the HTR through the VAs in the virtual experiemnt environment. That way, the HTR will be extended to an even more realistic tool to evaluate memory performance and speech comprehension.

However, this presents both opportunities and challenges.

On the one hand one could argue that visual cues enhance memory performance, because synchronized lip movements might improve speech comprehension [7], which also reduces subjective listening effort. On the other hand, however, the availability of auditory and visual cues might impede memory performance since, e.g., the audiovisual integration requires more processing capacities in the brain [20]. Thus, it has to be investigated to which extent auditory, perceptual, and visual information support comprehension and recall of spoken content, or whether not at all.
To be able to closely monitor the effect of every change in audiovisual presentation, the realism of the listening experiment will be increased gradually regarding visualization (see Figs. 2-4) and audio reproduction.

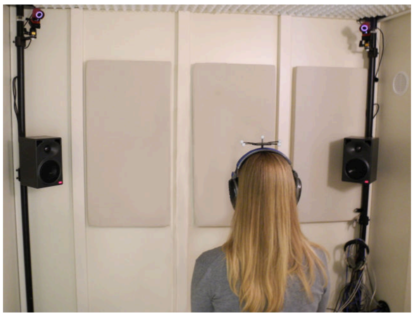

Figure 2: Experimental setup for audio-only condition.

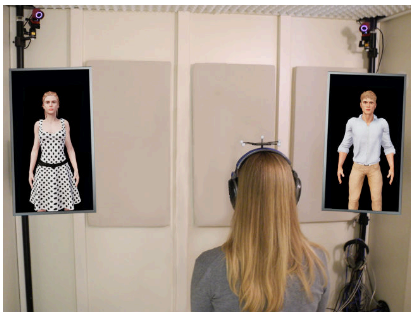

Figure 3: Experimental setup with static visual representations of the virtual interlocutors on two computer screens accompanying the acoustic stimuli of Fig. 2.

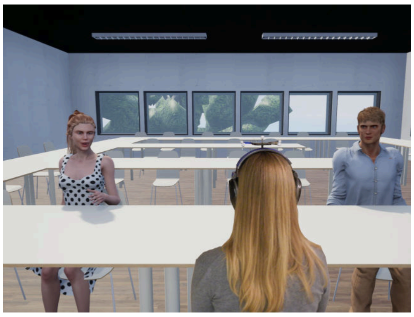

Figure 4: VR-based experimental setup with embodied VAs showing co-verbal cues accompanying the audio stimuli of Fig. 2.

Conclusion and Outlook

The main aim of the proposed project is to investigate the effect of variation in audiovisual characteristics of closer-to-reality listening settings on memory and comprehension of running speech. This will, in turn, allow determining those characteristics of audiovisual virtual reality environments, which are essential for closer-to-reality investigations of memory and comprehension for conversations. The impact of integrating audiovisual information into cognitive-psychological research will be investigated by gradually extending the listening experiment design to a plausible virtual visual and acoustic environment.

The gained insights can then promote theory building about performance-relevant conditions for listening to and processing of conversations while answering the question: do we need a closer-to-real-life audiovisual VR testing environment to measure memory and comprehension for running speech?

Funding Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SPP2236 - 444724862, Listening to, and remembering conversations between two talkers: Cognitive research using embodied conversational agents in audiovisual virtual environments.

References

[1] Imhof, M. (2010): What is Going on in the Mind of a Listener? The Cognitive Psychology of Listening. Listening and Human Communication in the 21st Century, 97–126

[2] Best, V.; Gallun, F.J.; Ihlefeld, A.; Shinn-Cunningham, B.G. (2006): The influence of spatial separation on divided listening. The Journal of the Acoustical Society of America 120, 1506-1516

[3] Arbogast, C.R.M. ; Kidd, G. Jr. (2002): The effect of spatial separation on informational and energetic masking of speech. The Journal of the Acoustical Society of America 112 (5), 2086–2098

[4] Freyman, R.L.; Balakrishnan, U.; Helfer, K.S. (2004): Effect of number of masking talkers and auditory priming on informational masking in speech recognition. The Journal of the Acoustical Society of America 115, 2246-2256

[5] Bronkhorst, A. (2000): The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acustica 86, 117-128

[6] 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

[7] Gonzalez-Franco, M. (2017): Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. Scientific Reports 7 (1)

[8] Craik, F. (1969): Modality effects in short-term storage. Journal of Verbal Learning and Verbal Behavior 8 (5), 658–664

[9] Schlittmeier, S.J.; Hellbrück, J.; Klatte, M. (2008): Does irrelevant music cause an irrelevant sound effect for auditory items? European Journal of Cognitive Psychology 20 (2), 252–271

[10] Fintor, E.; Aspöck, L.; Fels, J.; Schlittmeier, S.J. (2021): 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

[11] Bönsch, A.; Radke, S.; Overath, H.; Asché, L.M.; Wendt, J.; Vierjahn, T.; Habel, U.; Kuhlen, T.W. (2018): Social VR: How Personal Space is Affected by Virtual Agents' Emotions. IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 199–206

[12] Wendt, J.; Weyers, B.; Stienen, J.; Bönsch, A.; Vorländer, M.; Kuhlen, T.W. (2019): Influence of Directivity on the Perception of Embodied Conversational Agents' Speech. Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents

[13] Ehret, J.; Stienen, J.; Brozdowski, C.; Bönsch, A.; Mittelberg, I.; Vorländer, M.; Kuhlen, T.W. (2020): Evaluating the Influence of Phoneme-Dependent Dynamic Speaker Directivity of Embodied Conversational Agents' Speech. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20) 17, 1–8

[14] Blauert, J.; Jekosch, U.(2003): Concepts behind sound quality: Some basic considerations. Proceedings Internoise 2003, 72– 79

[15] Harvey, Alistair J.; Beaman, C. Philip (2007): Input and output modality effects in immediate serial recall. Memory 15 (7), 693–700

[16] Gagné, J.-P., Besser, J.; Lemke, U. (2017): Behavioral assessment of listening effort using a dual-task paradigm. Trends in Hearing

[17] McGarrigle, J.; Munro, K.J.; Dawes, P.; Stewart, A.J.; Moore, D.R.; Barry, J.G.; Amitay, S. (2014): Listening effort and fatigue: What exactly are we measuring? A british society of audiology cognition in hearing special interest group 'white paper'. International Journal of Audiology 53 (7), 433–445

[18] Seeman, S.; Sims, R. (2015): Comparison of psychophysiological and dual-task measures of listening effort. J Speech Lang Hear R 58 (6), 1781–1792

[19] Gosselin, P.A.; Gagné, J.-P. (2011): Older adults expend more listening effort than young adults recognizing speech in noise. Journal of Speech, Language, and Hearing Research 54 (3), 944–958

[20] Mishra, S.; Lunner, T.; Stenfelt, S.; Rönnberg, J. (2013): Visual information can hinder working memory processing of speech. Journal of Speech, Language, and Hearing Research 56 (4), 1120–1132