

Beyond words: The impact of static and animated faces as visual cues on memory performance and listening effort during two-talker conversations

Chinthusa Mohanathan^{a,*}, Plamenna B. Koleva^{a,1}, Jonathan Ehret^b, Andrea Bönsch^b, Janina Fels^c, Torsten W. Kuhlen^b, Sabine J. Schlittmeier^a

^a Work and Engineering Psychology, RWTH Aachen University, Jaegerstraße 17-19, 52066 Aachen, Germany

^b Visual Computing Institute, RWTH Aachen University, Kopernikusstraße 6, 52074 Aachen, Germany

^c Institute for Hearing Technology and Acoustics, RWTH Aachen University, Kopernikusstraße 5, 52074 Aachen, Germany

ARTICLE INFO

Keywords:

Short-term memory
Conversations
Listening effort
Static faces
Animated faces
Visual cues
Running speech
Spatial location
Listening

ABSTRACT

Listening to a conversation between two talkers and recalling the information is a common goal in verbal communication. However, cognitive-psychological experiments on short-term memory performance often rely on rather simple stimulus material, such as unrelated word lists or isolated sentences. The present study uniquely incorporated running speech, such as listening to a two-talker conversation, to investigate whether talker-related visual cues enhance short-term memory performance and reduce listening effort in non-noisy listening settings. In two equivalent dual-task experiments, participants listened to interrelated sentences spoken by two alternating talkers from two spatial positions, with talker-related visual cues presented as either static faces (Experiment 1, $n = 30$) or animated faces with lip sync (Experiment 2, $n = 28$). After each conversation, participants answered content-related questions as a measure of short-term memory (via the Heard Text Recall task). In parallel to listening, they performed a vibrotactile pattern recognition task to assess listening effort. Visual cue conditions (static or animated faces) were compared within-subject to a baseline condition without faces. To account for inter-individual variability, we measured and included individual working memory capacity, processing speed, and attentional functions as cognitive covariates. After controlling for these covariates, results indicated that neither static nor animated faces improved short-term memory performance for conversational content. However, static faces reduced listening effort, whereas animated faces increased it, as indicated by secondary task RTs. Participants' subjective ratings mirrored these behavioral results. Furthermore, working memory capacity was associated with short-term memory performance, and processing speed was associated with listening effort, the latter reflected in performance on the vibrotactile secondary task. In conclusion, this study demonstrates that visual cues influence listening effort and that individual differences in working memory and processing speed help explain variability in task performance, even in optimal listening conditions.

1. Introduction

Face-to-face conversations between two or more individuals are among the most common forms of verbal communication. They are rich in both auditory and visual information. In these conversations, the listener has to pay close attention to the speech of the talker(s) and remember the content to participate or to recall the heard information afterward. Listening involves intentionally selecting, arranging, and integrating both verbal and non-verbal information, requiring the processing of information from several modalities, such as auditory and

visual (Imhof, 2010). Visual information, including the mere presence of the talker or the talker with lip sync, likely influences the listener's comprehension and memory of the conversational content. Moreover, the listener's individual cognitive capabilities, such as working memory, attention, and processing speed, may determine how effectively they perceive and process running speech, defined as a continuous flow of speech (e.g., a series of interrelated sentences) (Rönnerberg et al., 2013). The present study investigates whether the visibility of two conversing talkers' faces, either static faces or animated faces with lip sync, improves short-term memory performance for running speech and reduces

* Corresponding author.

E-mail address: chinthusa.mohanathan@psych.rwth-aachen.de (C. Mohanathan).

¹ These authors contributed equally to this work.

listening effort in non-noisy listening settings. Furthermore, we assess listeners' individual working memory capacity, attentional resources, and processing speed in order to examine their roles on listening effort and short-term memory performance for the conversational content.

Listening to and remembering spoken information involves complex cognitive processes. A crucial component in this process is short-term memory performance. Short-term memory is defined as the capacity to retain a limited quantity of information in an active and readily accessible state for a brief period of time (Broadbent, 1958). This study uses the Heard Text Recall (HTR; Schlittmeier, Mohanathasan, Schiller, & Liebl, 2023) task to operationalize short-term memory performance in a conversational listening context. Closely linked to short-term memory performance in listening (auditory) tasks is listening effort. The Framework for Understanding Effortful Listening (FUEL; Pichora-Fuller et al., 2016) describes listening effort as “the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task” (p. S10). Within this framework, listening effort specifically pertains to the mental resources allocated to perceptual and cognitive processes necessary to perceive and encode the auditory input successfully into mental representations.

As a construct, listening effort is to be separated from the so-called task load (or cognitive load), which refers to the inherent cognitive demands a task imposes on the individual (e.g., Lemke & Besser, 2016; Pichora-Fuller et al., 2016). For example, a listening task might impose a high task load due to the task's high burden on executive or memory functions, while at the same time, the listener might still invest little listening effort if the speech signal is clear and easily encoded into mental representations. Conversely, listening effort can be very high in conditions like noisy environments, where the acoustic signal is partially masked. Even though the task load itself might be low, such as when the listener only has to repeat unrelated words or isolated sentences immediately after their presentation, the auditory input may map to multiple memory representations, requiring significant effort to resolve ambiguities and decide on a response.

While FUEL focuses on auditory tasks and the resources required for their successful execution, the integration of visual cues with auditory input, such as lip movements or static faces, introduces multimodal processing demands. In this context, the Ease of Language Understanding Model (ELU model; Rönnerberg et al., 2013) provides a cognitive-psychological theoretical framework to link audiovisual presentation conditions of running speech (“multimodal input” in the ELU) with memory and speech comprehension (referred to as “gist/understanding” in the ELU). The framework outlines how such multimodal inputs are processed through either an implicit route (automatic and requiring fewer cognitive resources) or an explicit route (resource-demanding and effortful). If the implicit route fails (e.g., due to mismatched auditory and visual cues or degraded speech), the explicit route is activated to compensate for the challenge, demanding more cognitive effort and resources.

The measurement of listening effort remains challenging as this cannot be done directly, in contrast to short-term memory performance, which can be assessed readily through content-related questions. While listeners can achieve similar performance results in a listening task across different listening conditions, such as varying signal-to-noise ratios (Howard et al., 2010), the level of listening effort required may vary between these conditions. To assess this variability in listening effort, a dual-task paradigm is used, where participants perform a primary listening task (e.g., recall of conversational content) alongside an unrelated secondary task (Anderson Gosselin & Gagné, 2011; Fintor, Aspöck, Fels, & Schlittmeier, 2022). The dual-task method measures cognitive spare capacity as an indicator of listening effort, based on the assumption that an individual's cognitive resources are limited (Broadbent, 1958; Kahneman, 1973; Navon & Miller, 2002; Pashler, 1994): if one task (the primary task) requires more resources, then fewer resources remain available for the simultaneous performance of the secondary task. For example, if primary task performance is similar in

two experimental conditions, A and B, but secondary task performance is reduced in condition A, this suggests that greater listening effort is required in this experimental condition (Gagné, Besser, & Lemke, 2017; McGarrigle et al., 2014). However, other performance patterns may also indicate listening effort, such as when primary task performance decreases in one condition while secondary task performance remains constant between two conditions (see Mohanathasan et al., 2025).

Listening effort not only depends on the acoustic characteristics of the relevant audio signal or the acoustic conditions in which listening takes place, but also on non-acoustic factors, such as the presence of visual cues related to the auditory speech signal. Previous studies have primarily examined how visual cues, such as seeing the talker's face, including lip sync, impact speech recognition and listening effort in noisy listening conditions (Anderson Gosselin & Gagné, 2011; Brown & Strand, 2019; Fraser, Gagné, Alepins, & Dubois, 2010; Sommers & Phelps, 2016). However, findings from these studies are contradictory. On the one hand, seeing lip movements in addition to hearing a voice might increase listening effort in noisy listening conditions (Fraser et al., 2010). We argue that this may be attributable to the fact that listeners are required to concurrently monitor and integrate visual and auditory information into a unified representation. The increase in effort may stem from the increased cognitive demand associated with simultaneously processing and integrating information from two sensory modalities rather than relying on a single modality (Brown & Strand, 2019). On the other hand, evidence suggests that being able to see and hear a talker simultaneously may instead reduce listening effort in noisy listening conditions. For example, research has shown that younger adults with normal hearing experience lower listening effort when the spoken signal is accompanied by visual cues such as animations of the talker's face with lip sync when listening in noise (Rudner, Mishra, Stenfelt, Lunner, & Rönnerberg, 2016; Sommers & Phelps, 2016). Here, the visual input might have provided complementary information to the auditory input, facilitating the mapping of the input onto phonetic representations in memory, thereby reducing the cognitive demand associated with processing. Yet, even in the absence of background noise, the presence of the talker's face was shown to enhance the free recall of spoken numbers by young adults with normal hearing (Mishra, Lunner, Stenfelt, Rönnerberg, & Rudner, 2013). However, the contradictory results highlight the need for further research to determine the specific conditions under which visual cues either increase or decrease listening effort.

The studies discussed above, particularly those that directly investigate the role of visual information in cognitive processing, provide valuable insights into how visual cues affect listening effort and short-term memory performance. However, they mostly used rather simple auditory stimuli such as single numbers (Mishra et al., 2013; Rudner et al., 2016), words (Sommers & Phelps, 2016), or unrelated sentences (Fraser et al., 2010), which may not capture the complexity and cognitive load of everyday listening tasks. In typical everyday conversations, listeners rarely listen to and recall such basic auditory stimuli. Instead, they need to integrate successively heard words, phrases, and sentences to construct coherent and accurate representations of the conversational content (see the Construction-Integration Model; Kintsch, 1988). Therefore, it remains unclear whether past research findings are directly applicable to more complex everyday listening situations involving two-talker conversations. Thus, future research could focus on more plausible listening scenarios that simulate real-life communication.

In addition, previous studies have primarily investigated the impact of adverse listening conditions on listening effort (e.g., Fraser et al., 2010; Rudner et al., 2016; Sommers & Phelps, 2016), leaving the role of visual cues in listening situations without background noise less explored. To address this research gap, recent studies have begun investigating listeners' short-term memory performance in a listening task that requires the recall of content from running speech by two alternating conversational partners in what are often considered optimal

listening settings (Mohanathan et al., 2025; Fintor et al., 2022; Mohanathan, Fels, & Schlittmeier, 2024).

Although these studies did not include visual cues corresponding to the talkers' speech and focused exclusively on the auditory spatial separation of the talkers, they represent a valuable step towards investigating short-term memory performance and listening effort in plausible listening situations. More specifically, Mohanathan et al. (2024) and Fintor et al. (2022) investigated the impact of spatial separation ($\pm 60^\circ$) versus co-location (0°) of two conversing talkers' audio signals on listeners' ability to temporarily retain and recall heard content (i.e., short-term memory performance) and listening effort in non-noisy listening conditions using a dual-task paradigm. The primary listening task consisted of the HTR task (Schlittmeier et al., 2023), in which participants answered content-related questions immediately after listening to a conversation between alternating female and male talkers. In parallel, participants performed a secondary task unrelated to the primary listening task.

Crucial to the present study, Mohanathan et al. (2024) also demonstrated the relevance of individual cognitive functions - working memory capacity, attention, and processing speed - for listeners' short-term memory performance and listening effort even in quiet settings. Short-term memory performance in listening tasks is influenced not only by the immediate ability to retain information but also by cognitive functions such as working memory capacity, attention, and processing speed. While short-term memory involves the temporary storage of verbal information, working memory refers to the capacity to store and manipulate information simultaneously to complete cognitive tasks (Baddeley, 2020; Daneman & Carpenter, 1980). The ELU model emphasizes the role of working memory across conditions under which individuals with unimpaired or impaired hearing listen to speech (Rönnberg et al., 2013). According to this model, listeners with higher working memory capacity are better able to resolve mismatches between incoming speech and long-term phonological representations, thereby reducing the need for effortful processing. This allows them to employ different types of information and strategies to extract meaning from a spoken message (Rönnberg et al., 2013). For example, Picou, Ricketts, and Hornsby (2011) found that individuals with larger working memory capacities benefited from visual cues, such as seeing the talker's face with lip sync during a recall task in noisy listening conditions. Although speech recognition performance remained constant, the inclusion of visual cues reduced listening effort. Conversely, listeners with lower working memory capacities did not benefit in a similar from visual cues.

Further, Picou, Ricketts, and Hornsby (2013) showed that listeners with low working memory capacity experience greater increases in listening effort when confronted with challenging conditions, such as the presence of background noise. Interestingly, individuals with low working memory capacity but good lip-reading abilities were more likely to benefit from visual cues, like lip sync, resulting in reduced listening effort. These findings suggest that visual cues can help compensate for challenges in language processing by enhancing language representation and reducing listening effort in noise, depending on individuals' cognitive functions (Picou et al., 2011, 2013).

In listening settings, where the auditory signals are clear and not partially masked by background noise, the supporting potential of visual cues may be less pronounced, as the auditory signal alone suffices for understanding speech. Nonetheless, even in quiet settings, visual cues may provide valuable supplementary information aiding speech processing, especially for individuals with lower working memory capacity. This assumption necessitates further research. Specifically, we need to better understand the interplay between auditory and visual information and individual cognitive functions without background noise.

Working memory is closely linked to the efficiency of information processing speed (Homman, Danielsson, & Rönnberg, 2023; Satori, 2021). Processing speed refers to how quickly individuals complete cognitive tasks (Salthouse, 1996), which is crucial for successful and

immediate speech understanding (Ng, Rudner, Lunner, Pedersen, & Rönnberg, 2013; Rönnberg et al., 2013). High processing speed enables working memory to better compensate for adverse listening conditions (Homman et al., 2023). Additionally, Picou et al. (2013) found a relationship between verbal processing speed and the advantages of visual cues, suggesting that listeners with faster verbal processing are more likely to benefit from visual cues.

Moreover, individual differences in working memory capacity are associated with the ability to focus attention and resist distraction during demanding tasks (Engle, 2002; Rönnberg et al., 2013; Satori, 2021). Individuals with high working memory capacity demonstrate enhanced attentional control, enabling them to allocate attentional and cognitive resources more effectively. This ability is particularly relevant in situations requiring divided attention (Rönnberg et al., 2013; Tun, McCoy, & Wingfield, 2009), such as when auditory and visual information must be processed simultaneously, as in the case of visible talker faces.

Building on these considerations, accounting for individual cognitive functions might further elucidate whether speech-related visual cues, such as visible talker faces, also enhance memory for conversational content and/or reduce listening effort in non-noisy listening settings.

1.1. Research intent

The present study investigates whether visual cues in the form of static or animated faces improve short-term memory performance (RQ1) and reduce listening effort (RQ2) during two-talker conversations. A dual-task paradigm was employed (see Fig. 1) (Ehret et al., 2024; Fintor et al., 2022; Mohanathan et al., 2024), where participants completed the HTR task (Schlittmeier et al., 2023), a short-term memory performance task in which participants listen to spoken conversations and then answer content-related questions to assess their ability to retain and recall verbal information over a brief period. We refer to participants' performance in the HTR task as 'short-term memory performance,' which we distinguish from the broader construct of working memory, assessed separately through standardized measures. In the primary listening task, participants answer content-related questions after listening to conversations between a male and a female talker. Static faces (Experiment 1) or animated faces with lip sync (Experiment 2) of a female and a male virtual agent were presented alongside the auditory speech signals. To assess listening effort, participants performed a vibrotactile pattern recognition task as a secondary task (Anderson Gosselin & Gagné, 2011; Mohanathan et al., 2025; Fraser et al., 2010). The utilization of a vibrotactile secondary task is advantageous due to its independence from the visual and auditory modalities. This is particularly useful for studying the effects of visual cues on listening performance, as it avoids the interference that can occur when the primary and secondary tasks rely on the same sensory modality, as discussed in other studies (e.g., Picou & Ricketts, 2014). Working memory, processing speed, and attention were included as covariates to control for individual differences in cognitive capacity that are known to influence task performance (e.g., Homman et al., 2023; Mohanathan et al., 2024; Rönnberg et al., 2013). This approach is supported by findings showing strong effects of processing speed on speech-in-noise performance (Dryden, Allen, Henshaw, & Heinrich, 2017) and by more recent work emphasizing the importance of broader cognitive abilities - such as fluid intelligence, theorized to encompass working memory - in explaining individual differences in listening performance (Eckert, Matthews, Vaden, & Dubno, 2024; Marsja et al., 2025; Marsja, Stenbäck, Moradi, Danielsson, & Rönnberg, 2022). These variables were not central to our hypotheses, and thus were modeled as control variables rather than predictors of theoretical interest. Participants were also queried about their subjective assessment of task demands, their performance, and attention to the visual cues.

As previously stated, speech understanding is improved by the visibility of the talkers providing the auditory signals (Rudner et al., 2016; Sommers & Phelps, 2016). We hypothesized that visual cues - either

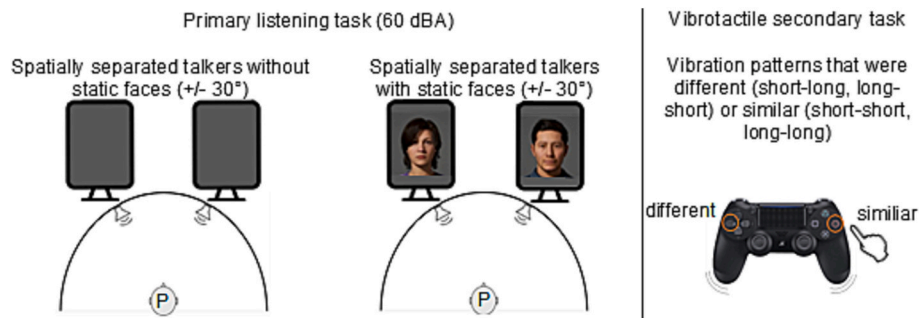


Fig. 1. The dual-task paradigm in Experiment 1. Left side: primary listening task presented at 60 dB(A), which could be performed without or with static faces of the talkers (within-subject). After each text, participants (P) had to answer questions about its content. Right: the vibrotactile secondary task, in which two identical (short-short or long-long) or two different (short-long or long-short) vibration patterns were presented and had to be categorized via a game controller. (Figure adapted from “Listening to two-talker conversations in quiet settings: the role of listeners’ cognitive processing capabilities for memory and listening effort.” by Mohanathan et al., 2024, *Scientific Reports*, 14, 22764).

static or animated talker faces - would improve short-term memory performance in the primary listening task compared to conditions without faces (RQ1). Based on the ELU model (Rönnberg et al., 2013), we argue that visual cues improve memory performance by triggering the implicit processing route, which is a more automated and resource-efficient cognitive processing mechanism. In conditions with static or animated faces, the visual cues either capture and sustain listeners’ attention (static faces) or provide congruent lip movements that align with the auditory speech signal (animated faces). This congruence helps activate the implicit route, allowing listeners to rely on both auditory and visual signals for efficient processing. In contrast, in the absence of visual cues, listeners must rely solely on auditory input, which may trigger the explicit route. This route demands more cognitive resources and focused attention to understand the spoken language, leading to greater listening effort. Furthermore, we hypothesized that the presence of static or animated faces as visual cues would reduce listening effort compared to conditions without faces (RQ2). This is on the basis that the visual input provides complementary information to the auditory input, as previously stated. Specifically, participants were expected to perform better with fewer errors and/or faster RTs in the vibrotactile secondary task during dual-task conditions when static or animated faces were presented compared to when they were not. Finally, we hypothesized that all three individual cognitive functions would be related to the effect of visual cues (compared to no cues) on memory performance and listening effort (RQ3).

2. Methods Experiments 1 and 2

2.1. Experimental design

We combined the data from Experiments 1 and 2. This approach was chosen because both experiments were designed equivalently, with only the type of visual cue (static vs. animated faces) differing. Furthermore, this approach enables a direct comparison between static and animated faces, facilitating a more precise assessment of their impact on listening effort and short-term memory performance, as assessed by the dual-task paradigm. Consequently, the methods for both experiments are presented together.

2.2. Participants

We conducted a Monte Carlo simulation-based power analysis following Arend and Schäfer’s (2019) tutorial, using the R package *simr* (version 4.1.2). The R script for the simulation is included in the Supplement (see R scripts). For the primary listening task, we assumed $\alpha = 0.05$, an intraclass correlation coefficient (ICC) of 0.30, a random slope variance component of 0.09, and a Level 1 sample size of 126 (14 texts $\times$ 9 questions = 126). The ICC represents a conservative medium estimate

appropriate for our repeated-measures design. The random slope variance component was set according to the rule of thumb recommended by Arend and Schäfer (2019). Effect sizes were defined in terms of standardized regression coefficients (β), and we assumed medium effects at Level 1 ($\beta = 0.30$), Level 2 ($\beta = 0.30$), and for cross-level interactions ($\beta = 0.30$), consistent with conventions for power simulation in GLMMs (Maas & Hox, 2005; Aarts et al., 2014; Arend & Schäfer, 2019). This assumption is further supported by effect size estimates reported in recent GLMM-based studies on related topics such as visual cues, listening effort, and speech comprehension, although these studies were not identical in design (Brown & Strand, 2019; Strand, Brown, & Barbour, 2020). We note that this assumption may overestimate the true effect size. We ran 1000 simulations with a Level 2 sample size of 58 participants. The analysis yielded 100% power to detect medium Level 1 effects and 65% power to detect medium Level 2 effects and medium cross-level interactions. For the vibrotactile secondary task, the same assumptions were applied, except that the Level 1 sample size was set to 220 (based on an estimate of the number of completed vibrotactile patterns per participant). Again, 1000 simulations were conducted, yielding 100% power to detect medium Level 1 effects, 66% power for medium Level 2 effects, and 62% power to detect medium cross-level interactions. Given our primary interest in Level 1 effects, we considered a sample size of 58 participants adequate, despite reduced power for Level 2 and cross-level interaction effects.

39 participants voluntarily took part in Experiment 1. However, seven participants were excluded due to technical problems during data collection. Additionally, two other participants were identified as having a hearing impairment through pure tone ascending hearing screening (AURITEC ear 3.0 audiometer and Sennheiser HAD 280 headphones), so they were also excluded from the data analysis. Thus, the final sample consisted of 30 participants (20 females, 9 males, and 1 non-binary) aged between 19 and 44 years ($M = 26.07$ years, $SD = 5.04$).

A new sample of 35 participants took part in Experiment 2. Four participants were excluded from the data analysis due to technical difficulties encountered during data collection. Data from one further participant was removed because they performed the psychometric test to measure attention incorrectly. Two additional participants were excluded due to impaired hearing. The final sample consisted of 28 participants (22 females, 5 males, and 1 non-binary) aged between 19 and 30 years ($M = 22.50$ years, $SD = 2.91$).

All participants in both experiments reported normal or corrected-to-normal vision and had normal hearing sensitivity in both ears (<20 dB hearing level in the frequency range between 250 and 4000 Hz). They also spoke fluent German and were mostly students. The participants were recruited via e-mail and were compensated with course credits or €20 for their participation.

2.3. Stimuli, instruments, and apparatus

Experiment 1 was programmed in PsychoPy 2021.2.3 and Experiment 2 in PsychoPy 2023.2.3 (Python 3.6.6; [Peirce, 2007](#)). Both experiments were conducted on a Dell Latitude 3590 laptop. The dual-task paradigm and the task to measure working memory (the automated Operation Span Task) were displayed on the laptop's 15" non-glare screen, while the visual cues were displayed on two external 24" BenQ BL2400 monitors set up vertically at $\pm 30^\circ$ relative to the participants (see Fig. S1 in the supplementary material, file 1). The auditory stimuli were played via an external Focusrite Scarlett 2i2 2nd Gen sound card and two Genelec 8020 loudspeakers, each placed in front of the corresponding monitor. Participants were seated in the center and at a distance of about 1.5 m from the loudspeakers.

The HTR ([Schlittmeier et al., 2023](#)), which was used as the listening task, consists of German texts presented auditorily using auditory stimuli from a female and male talker from the AuViST database ([Ermert et al., 2023](#)). The auditory stimuli were calibrated to a sound pressure level of 60 dBA at the ear canal entrance of the HMS II.3–33 full-band artificial head. In Experiment 1, we used static faces as visual cues, consisting of a female and a male virtual agent (see Figs. S2 and S3 in the supplementary material, file 1). The virtual agents were created using MetaHuman Creator (Epic Games, v0.5.2). In a pre-test ($n = 14$), in which different virtual agents were presented, the virtual agents that looked most natural and were judged to best match the talkers' voices were selected and used in the current experiment. In Experiment 2, we used the same virtual agents, but here they were animated (see [Video S4](#) in the supplementary material). It should be noted that the slight asynchrony evident in the supplementary video was due to technical limitations in its creation and did not have any impact on the primary experiment, in which the videos were synchronized. The animated videos of the talkers were created using face-tracking data from the AuViST database ([Ermert et al., 2023](#)). The visual representations were dynamically rendered using Unreal Engine 4.27, maintaining identical lighting conditions as those used for the static faces. In addition to the face animation, we also played an idle animation, resulting in a slight body sway. Furthermore, we implemented natural blinking and eye saccades, following [Lee, Badler, and Badler \(2002\)](#), to increase the naturalness of the face animations. For the paper and pencil tests (the Zahlen-Verbindungs-Test and the Frankfurt Attention Inventory 2), a Renkforce RWF-SW-130 digital stopwatch was utilized to accurately measure time.

2.3.1. Dual-task paradigm

2.3.1.1. Primary listening task. The HTR paradigm ([Schlittmeier et al., 2023](#)) was used as the primary listening task to measure short-term memory performance. Here, participants listened to coherent spoken texts in a setting without any background noise. This paradigm has already been used in past studies ([Fintor et al., 2022](#); [Mohanathan et al., 2024](#); [Schiller, Aspöck, & Schlittmeier, 2023](#)). The texts described three generations of families, including grandparents, parents, and children, and provided, for example, information about their occupations, and relationships. Each text contained 5–6 people with different names, hobbies, locations, and ages. Each coherent spoken text in the study consisted of ten sentences and was presented as a conversation between a female and a male talker. For more information, we refer the interested reader to [Schlittmeier et al. \(2023\)](#).

To simulate a natural conversation between a man and a woman, the sentences were reproduced by two talkers who took turns speaking, with a break of 0.6 s between sentences, and were presented at two spatial positions ($\pm 30^\circ$ relative to the listener). The number of sentences each talker spoke was counterbalanced, and talkers never spoke simultaneously, in order to avoid masking speech signals.

It is important to note that this task involved two talkers reading a

prepared text rather than engaging in a real-life conversation. While this structure may influence how participants engage with the task, it offers a controlled environment that minimizes variability and ensures consistency across participants.

For each text, participants answered nine questions about family members' names, relationships, or other information (e.g., occupations, places of residence). Questions were presented visually one after the other on the laptop, and responses were given by typing using an external keyboard. Participants had to answer each question with only one or two words and had unlimited time. Furthermore, they could skip a question by pressing the "enter" key. However, once a question had been skipped, participants could not return to it.

2.3.1.2. Vibrotactile secondary task. The secondary task was a vibrotactile pattern recognition task ([Mohanathan et al., 2025](#)), where participants categorized four vibration patterns (short-short, long-long, short-long, and long-short) using a Sony Dualshock 4 controller held with both hands. The task was to identify whether the presented tactile pattern was the same (i.e., short-short or long-long) or different (i.e., short-long or long-short) by pressing the left arrow or the circle key on the game controller.

In dual-task conditions, the length of the sound files determined the number of presented vibrotactile stimuli. The stimulus interval remained constant at 1.5 s, resulting in a varying response stimulus interval. If the participant did not respond within 1.5 s, the next tactile pattern began.

2.3.2. Self-report questionnaire

We developed a paper-and-pencil German self-report questionnaire with eight closed and two open-ended questions to capture participants' subjective perceptions of their short-term memory performance and listening effort in the two experimental conditions (no faces vs. static faces in Experiment 1 or animated faces in Experiment 2). An overview of all questions can be found in the supplementary material (see file 2).

Participants rated the first seven questions on a five-point Likert scale, ranging from "do not agree at all" (1) to "fully agree" (5). For the last closed question, participants had to choose which condition they found easier.

2.3.3. Covariates

2.3.3.1. Working memory capacity. The automated version of the Operation Span Task (OSPAN; [Turner & Engle, 1989](#)) was used to measure participants' working memory capacity. The automated OSPAN (aOSPAN; [Unsworth, Heitz, Schrock, & Engle, 2005](#)) is mouse-controlled by the participants and automatically records their results. In the aOSPAN task, participants had to maintain an increasing sequence of consonants (e.g., F, H, J), ranging from 3 to 7 letters, which they had to recall after each trial. Each consonant remained on the screen for 800 ms and was preceded by a math problem (e.g., " $3/3 + 2 = ?$ "), followed by a proposed solution to the math problem (e.g., "9"). The task for the participant was to decide whether the given solution was "correct" or "incorrect" by clicking on a button. Subsequently, the letter recall was assessed by requesting participants to choose consonants from a 4×3 consonant matrix. Participants had the option to exclude a consonant by selecting "blank" and to correct a previously entered consonant by selecting "delete". A total of three sets of each set size, ranging from three to seven (e.g., set size three equals three math operations and three consonants), were presented in random order, resulting in a total of 75 letters and 75 math problems. Each participant's overall performance score was calculated by summing the number of consonants and math problems correctly identified.

The aOSPAN reflects the dual-task nature of working memory, as it requires participants to simultaneously process information (e.g., solving math problems) and retain verbal items for later recall. This

combination of storage and processing aligns with theoretical definitions of working memory capacity.

2.3.3.2. Processing speed. Processing speed was measured with the trail-making test (German: Zahlen-Verbindungs-Test, ZVT, Oswald & Roth, 1987). This paper-and-pencil test consists of four forms, each with numbers from 1 to 90 in different positions. The participants' task is to connect all numbers from 1 to 90 in the correct order, starting with the number 1. The overall score (ZVT mean RT) was determined by the average time taken to connect all the numbers across all four test forms.

2.3.3.3. Attentional functions. Finally, attentional functions were measured with the Frankfurt Attention Inventory 2 (FAIR-2) by Moosbrugger and Oehlschlägel (2011). The FAIR-2 is a two-page performance-based attention test in which participants are expected to complete part of each page within the time limit. The focus is on determining which line they can reach while maintaining accuracy. The stimuli consisted of a combination of a shape (circle or square) and dots (two or three dots). Participants had to draw a continuous line under each row of stimuli, indicating each target by drawing an upward line through it.

Performance in FAIR-2 is evaluated using the K variable, which is derived from the German word 'Kontinuität' (meaning 'continuity'). K is calculated by multiplying the number of correctly detected targets by the percentage of correctly inspected items out of the total inspected items. K is the only measure used to indicate the overall consistency of performance.

2.3.4. Sociodemographic data

All participants answered sociodemographic questions such as age, gender, eyesight, handedness, and first language at the beginning of the dual-task experiment.

2.4. Procedure

The procedures for Experiment 1 and Experiment 2 were the same. The experiments took place in a soundproof booth (Studiobox, premium version). Participants received an oral and written briefing on the experiment and its requirements upon arrival at the lab. After signing the consent form, all participants underwent hearing screening to ensure normal hearing according to the World Health Organization grade system (World Health Organization, 1991). For this purpose, pulsed pure tones were played at frequencies ranging from 250 to 4000 Hz. For each frequency, participants were instructed to press a button as soon as they could hear a tone. The complete test protocol was administered in a single session and took approximately 90 min. Written instructions for the dual-task paradigm and the aOSPAN were displayed on the screen, while instructions for the FAIR-2 were provided on a sheet of paper. The experimenter provided oral instructions for the ZVT.

The experimental session started with the dual-task paradigm. A practice block preceded the main task. The practice block began with the primary listening task in a single-task condition without faces, in which one text spoken by two alternating talkers was presented. Then participants familiarized themselves with the vibrotactile patterns of the secondary task through a pre-practice task, where they could press a button on the controller (either PL or O) to activate the corresponding vibrotactile patterns. The combinations of the different vibrotactile patterns were displayed on the laptop screen, and a green color indicated which pattern was currently presented on the controller. For example, when participants pressed the O button of the controller, the controller vibrated in the short-short and long-long vibrotactile patterns in sequence, and the words "short-short" and "long-long" were surrounded by a green rectangle on the laptop screen. This exercise could be repeated as often as necessary. Subsequently, 20 practice trials of the vibrotactile secondary task were conducted as a single-task condition.

After each response, participants received feedback on the accuracy of their responses. If the vibrotactile pattern was correctly identified, the German word for "correct" (richtig) was displayed. Conversely, if the vibrotactile pattern was not correctly identified, the German word for "incorrect" (falsch) was presented. If participants did not respond within 1.5 s, they received automated feedback that they had run out of time. However, this feedback was only provided during the practice block.

Afterward, the participants practiced dual-tasking in two blocks, where they had to perform the listening task, consisting of one text, and the vibrotactile pattern recognition task in parallel. In one dual-task practice block, no faces were displayed. In contrast, in the other block, static (Experiment 1) or animated faces (Experiment 2) of female and male virtual agents were presented. The static or animated faces were displayed separately on two screens while the audio recording of each talker was running. Thus, the images were presented alternately. For example, while the woman's audio recording was playing, only her image was displayed on the screen on the participant's left side. When the audio recording of her speech ended, the image disappeared. Subsequently, the male virtual agent's image was displayed on another screen on the participant's right side as long as that audio recording was running. For half of the participants, the image of the female virtual agent was presented on the left monitor and the image of the male virtual agent on the right monitor (see Fig. 1), while for the other half, the images were presented in the opposite position.

After the practice session, performance measures started. The study consisted of three main blocks: 1) the listening task in a single-task condition, 2) the vibrotactile pattern recognition task in a single-task condition, and 3) both tasks performed simultaneously in the dual-task condition. Four texts were presented in the single primary task condition, two with and two without faces. The secondary task in the single-task condition consisted of 40 trials. Five recordings were played in the dual-task condition with static faces (Experiment 1) or animated faces (Experiment 2) and five recordings without faces. After each text, questions about its content were displayed on the laptop, which the participants answered using an external keyboard. In the dual-task condition, the number of vibrotactile stimuli presented was determined by the length of the audio recording of the primary task, allowing participants to complete as many vibrotactile trials as possible. The condition order (with static faces in Experiment 1 or animated faces in Experiment 2 or without faces) and the main blocks (single primary task, single secondary task, and dual task) were counterbalanced across participants. Furthermore, breaks between blocks were allowed, the lengths of which were not fixed.

After completing the dual-task paradigm, participants were administered the paper-and-pencil self-report questionnaire, followed by the three psychometric tests (aOSPAN, ZVT, and FAIR2) in a randomized order.

The aOSPAN, which measures individuals' working memory capacity, began with a practice session consisting of three parts. Firstly, participants were asked to recall sequences of consonants in ascending order, with set sizes of two and three, each repeated twice. Following the recall, the computer provided fully automated feedback on the number of correctly recalled letters in the current set. Secondly, participants practiced 15 math operations. They had to solve a math operation and click the mouse to proceed with the task. On the following screen, a solution to the math problem was displayed, and participants were asked to indicate whether it was "true" or "false". Feedback on accuracy was provided after each operation. The math practice was used to familiarize participants with the math section of the task and to calculate their mean time to solve the equation. This time ($+2.5 SD$) was used as a time limit for the math portion of the experimental session. In the final practice section, participants were asked to recall sequences of letters and solve math problems, as they would in the experimental session. Each participant completed three practice trials of set size two. If the time taken to solve the math operation exceeded their individual time limit, the program automatically processed and recorded the attempt as

an error. This method was implemented to prevent participants from rehearsing consonants instead of solving math operations. After the practice session, the performance measures began. Participants completed three repetitions of five set sizes, with the order of set sizes randomized. They solved math problems and had to recall a sequence of letters.

Before starting the paper-and-pencil ZVT test to measure processing speed, participants were given a practice sheet with the numbers 1 to 20 in random order, while the actual task consisted of the numbers 1 to 90. Participants completed all four test forms sequentially, with no breaks in between the forms. The time taken to complete each form was recorded using a timer that was started when the participant opened the test sheet.

In the FAIR-2, used to measure individual attentional functions, each participant received a test form and was asked to read the standardized instructions printed on it. After reading the instructions, participants practiced the task with one row of 20 stimuli. Once they successfully completed this practice row, they had to solve two pages containing 16 lines with 20 stimuli per line. The entire test took 6 min, with a three-minute time limit per page.

2.5. Data analysis

The statistical analysis was carried out with R (version 4.1.2; [R Core Team, 2023](#)). All R scripts for statistical analysis are included in the Supplement (see R scripts). The dependent variables comprised performance measures in the HTR task (a binary variable denoting correct responses to nine questions per text, with 1 indicating a correct response and 0 indicating an incorrect response), performance measures in the secondary task (a binary variable indicating correctness in the vibrotactile pattern recognition, with 1 representing the correct responses and 0 representing the incorrect responses), and reaction times (RTs) of correct trials in the secondary task (measuring the duration between the appearance of the vibrotactile pattern and the participant's response).

In this study, we utilized GLMMs to analyze performance and RT data, leveraging the lme4 package (version 1.1.32; see [Bates, Mächler, Bolker, & Walker, 2015](#)) in R. While we initially planned to conduct ANOVA, we transitioned to GLMMs early in the data analysis process. This approach allowed us to appropriately model the repeated-measures structure of the data and better accommodate its distributional properties. However, we additionally report mixed-design ANCOVA in the Supplement (see File 4). While these models yields some differences in significance patterns, as described briefly in the supplement, it relies on assumptions less suitable for our data (e.g., binary, trial-level data). GLMMs offer increased statistical power and are adept at capturing individual-level variability and dependencies among observations, thus surpassing ANOVA ([Jaeger, 2008](#); [Lo & Andrews, 2015](#)). This methodological choice was made to enhance the precision and robustness of the analysis. RT data modeling was exclusively conducted using correct responses, excluding incorrect responses or errors from the analysis. Before analyzing the RT data, RTs under 150 ms ([Whelan, 2008](#)) and those exceeding two standard deviations from the mean were treated as outliers and removed from the analysis, following the procedure of [Berger and Kiefer \(2021\)](#). This resulted in the exclusion of 6.5% of RT data.

Performance was modeled in GLMM using a binomial distribution and a logit link function, while RTs were modeled using a GLMM with a Gamma distribution and a log link function. We opted for a Gamma distribution for RTs because of several considerations. First, RTs are inherently positive and typically right-skewed, making the Gamma distribution an appropriate modeling choice ([Lo & Andrews, 2015](#)), unlike linear regression, which assumes an unbounded range for predictions, including negative values. Second, the log link function allows for modeling multiplicative relationships, which are often more appropriate and interpretable for RT data ([Baayen & Milin, 2010](#)). Second, Gamma GLMMs are well-suited to hierarchical data structures, enabling

us to model both subject- and item-level variability via random effects ([Lo & Andrews, 2015](#)). Finally, using a GLMM framework for both accuracy and RTs ensures consistency in the modeling approach and facilitates coherent comparisons of fixed and random effects across outcomes. While the Ex-Gaussian distribution is frequently used for descriptively characterizing RT distributions, it is not readily implemented within standard GLMM frameworks without custom modifications or Bayesian methods. In contrast, the Gamma distribution is natively supported in common GLMM software, providing a straightforward approach to modeling our data with the desired flexibility and rigor.

Three independent GLMMs were conducted to model the effect of *visual condition* on performance in the primary task (HTR; see [Section 3.1](#)), performance on the secondary task (vibrotactile pattern recognition, see [Section 3.2](#)), and RTs in the secondary task (see [Section 3.2](#)). For each independent GLMM, we started with a full maximal model including all theoretically valid fixed and random (intercept) effects and determined the best-fitting model via backward model selection. The models were compared using Likelihood Ratio tests (LRT). We have provided additional information in the supplementary material (see File 3) on our model selection processes, the alternative models we tested, and the reasons for rejecting them. It is important to note that all fixed factors and covariates were retained, with only the random (intercept) factors and random slopes being optimized. To ensure comparability of effect sizes and improve model stability, all covariates (processing speed, attention, and working memory) were z-standardized (centered and scaled) prior to model fitting across all GLMMs. As a result, the reported coefficients for these predictors represent standardized effects (β). Post hoc pairwise comparisons were conducted using estimated marginal means utilizing the emmeans package (version 1.8.5; [Lenth, 2021](#)). When needed, the significance level was adjusted to account for multiple comparisons by employing Tukey's HSD tests. We computed Variance Inflation Factors (VIFs) from a linear model containing the identical fixed-effects structure, which is the recommended approach because VIFs are not defined for GLMMs with random effects, to assess potential multicollinearity among predictors. To assess model fit, both marginal and conditional R^2 values were calculated using the r2() function from the performance package (version 0.15.0; [Lüdtke et al., 2021](#)). Additionally, semi-partial R^2 values were calculated using the r2beta() function from the r2glmm package (version 0.1.3; [Jaeger, 2017](#)) to provide insights into the contributions of individual predictors.

First, we investigated the effect of *visual condition* on short-term memory performance in the primary listening task (RQ1, see [Section 3.1](#)). It was hypothesized that the inclusion of both static and animated faces as visual cues would enhance participants' short-term memory performance in the primary listening task in comparison to conditions that did not feature faces. For the GLMM modeling of memory performance, we considered *visual condition* (without faces, with faces), *experiment* (Experiment 1, Experiment 2), *number of tasks* (single listening task, dual-tasking), and all their two-way and three-way interactions as fixed factors. Additionally, *processing speed* (ZVT mean RTs), *attention* (K score), and *working memory* (aOSPAN score) were included as covariates (RQ3; all covariates standardized). *Age*, *question number*, and *text* initially functioned as random (intercept) factors, which might contribute to the explanation of differences between groups that are not captured by the fixed effects in the model. *Age* was treated as random intercepts to account for age-related cognitive differences. In order to account for any variability in text difficulty ([Schlittmeier et al., 2023](#)), *question number* and *text* were included as random intercepts. To better capture participant-level variation in how the experimental conditions influenced short-term memory performance, we also tested a random slope for the within-subject predictors *visual condition* and *number of tasks*, as well as their interaction. *Participant ID* served as the grouping factor for this random effects structure. As discussed by [Barr, Levy, Scheepers, and Tily \(2013\)](#), including random slopes is important, as models with only random intercepts can inflate

Type I error rates. In the best-fitting model analyzing performance in the primary listening task, all initially considered factors except for *age* were retained.

Secondly, we investigated the effect of *visual condition* on listening effort, operationalized through performance (errors) and RTs in the secondary task (RQ2), namely the vibrotactile pattern recognition task (see Section 3.2). Participants completed this secondary task under three *task conditions*: (a) in a single-task condition without the listening task and without faces; (b) in a dual-task condition simultaneously with the listening task without faces; or (c) with faces. In Experiment 1, the faces were static, whereas in Experiment 2 they were animated. It was hypothesized that the presence of static or animated faces as visual cues would result in a reduction in listening effort compared to conditions without faces (RQ2). The GLMMs modeled performance and RTs using the three *task conditions* (a)–(c), *experiment*, and their two-way interaction as fixed factors and *processing speed* (mean RTs), *attention* (FAIR-2 K score), and *working memory* (aOSPAN score) as covariates (RQ3; all covariates standardized). Additionally, *age* and *vibration type* were included as random (intercept) factors. *Vibration type* was considered a random intercept to account for individual differences in how participants perceive or respond to different vibration patterns. To account for participant-level variability in response to the within-subject *task conditions*, we considered a random slope for *task condition*, with *participant ID* as the grouping factor. In the best-fitting model analysis of secondary task performance all initially considered factors except for *age* were retained. For the RT analysis, all originally considered factors – except for *age* – were included. However, to enable the computation of conditional R^2 , the random slope was replaced with a random intercept for *participant ID* (see also Supplement File 3 for details).

Finally, we analyzed the self-report questionnaire using descriptive statistics to examine the effect of visual condition on participants' subjective experience (see Section 3.3).

3. Results

3.1. Performance in the primary task (RQ1 & RQ3)

The effect of visual cues on short-term memory performance for conversational content was analyzed based on performance in the primary listening task (RQ1). The descriptive results are displayed in Fig. 2.

Table 1 provides a summary of the best-fitting GLMM that models memory performance. It was hypothesized that the inclusion of both static and animated faces as visual cues would enhance participants' short-term memory performance in the primary listening task, in comparison to conditions that did not feature faces (RQ1). In contrast to our hypothesis, the results demonstrated that there was no significant effect of *visual condition* on memory performance, while considering *experiment*, *number of tasks*, *working memory*, *processing speed*, and *attention* ($\chi^2(1) = 0.04$, $p = .85$). These findings indicate that visual faces, e.g. static or animated faces, did not enhance short-term memory performance in the primary listening task.

However, *experiment* had a significant effect on memory performance ($\chi^2(1) = 28.68$, $p < .001$) while considering *visual condition*, *number of tasks*, *working memory*, *processing speed*, and *attention*. The findings suggest that in Experiment 1 less errors were made than in Experiment 2 (z -ratio = -3.66 , $p < .001$).

Furthermore, a significant effect of *number of tasks* on memory performance was found ($\chi^2(1) = 27.59$, $p < .001$) after controlling for *visual condition*, *experiment*, *working memory*, *processing speed*, and *attention*. This suggests that more memory errors were made in the listening task in the dual-task condition compared to the single-task condition (z -ratio = 3.09 , $p < .001$).

The two-way interactions between *visual condition* and *number of tasks* ($\chi^2(1) = 0.39$, $p = .53$), *visual condition* and *experiment* ($\chi^2(1) = 0.21$, $p = .64$), *number of tasks* and *experiment* ($\chi^2(1) = 0.52$, $p = .47$) was not significant, after adjusting for *working memory*, *processing speed*, and *attention*. The same applies for the three-way interactions between *visual*

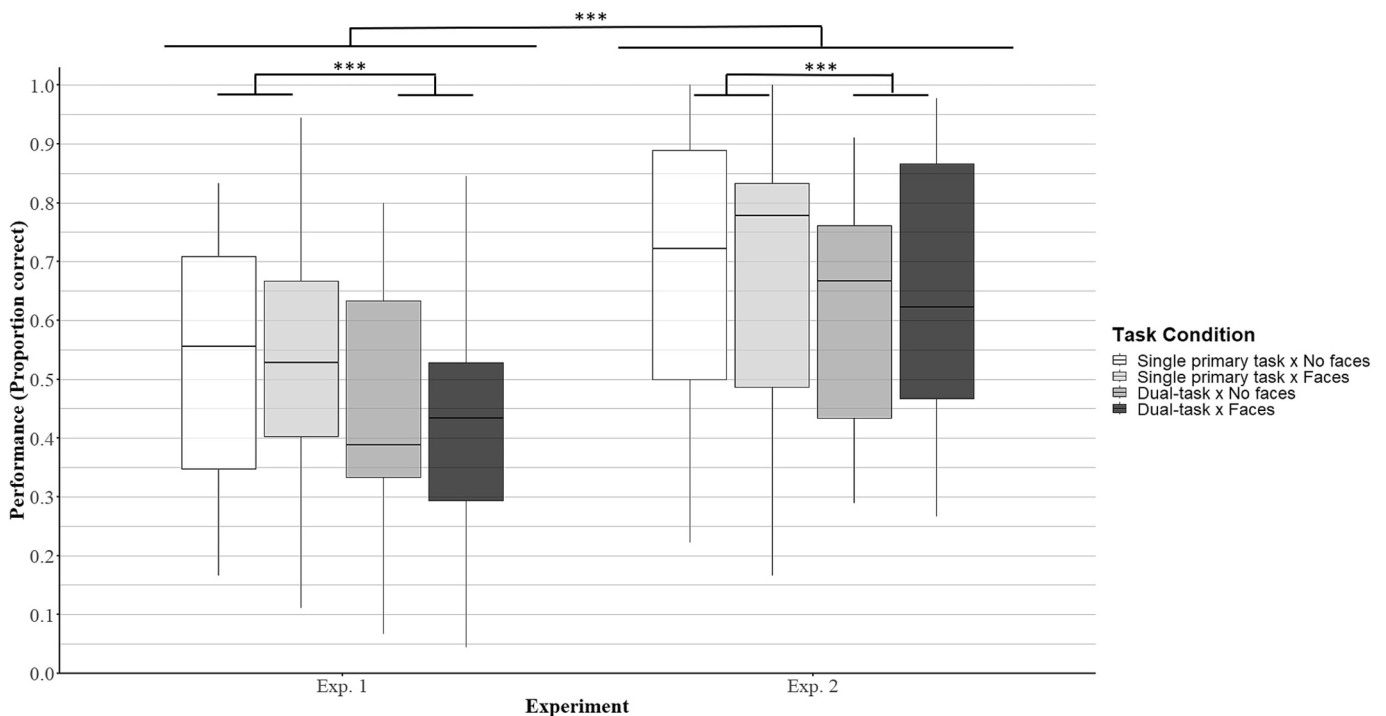


Fig. 2. Performance in the primary listening task. Listening task performance (proportion correct) as a function of *number of tasks* (listening task in single-task condition vs. in dual-task condition) and *visual condition* (without faces vs. with static faces vs. with animated faces). The boxplots show the data distributions for primary task performance (proportion correct). The boxes represent the interquartile ranges, while the lines contained within the boxes represent the medians. The points indicate outliers. *** $p < .001$.

Table 1

Results from the final GLMM modeling performance in the HTR (listening) task as predicted by visual condition and experiment.

Fixed effects and covariates	Estimate	SE	z-ratio	95% CI	p
Intercept	0.08	0.24	0.31	0.67, 1.73	0.75
Visual condition					
No faces	Reference				
Faces	0.02	0.23	0.10	0.65, 1.60	0.92
Experiment					
1	Reference				
2	1.06	0.29	3.66	1.64, 5.09	<0.001
Number of tasks					
Single-task	Reference				
Dual-task	-0.44	0.14	-3.09	0.49, 0.85	<0.001
Visual condition × Experiment					
No faces × Exp.1	Reference				
Faces × Exp. 2	-0.16	0.34	-0.48	0.44, 1.64	0.63
No faces × Exp. 2	Reference				
Faces × Exp. 1	0.16	0.34	0.48	0.61, 2.27	0.63
Visual condition × Number of tasks					
No faces × Single-task	Reference				
No faces × Dual-task	-0.07	0.17	-0.38	0.66, 1.31	0.70
No faces × Dual-task	Reference				
Faces × Single-task	0.07	0.17	-0.38	0.77, 1.50	0.70
Experiment × Number of tasks					
Exp. 1 × Single-task	Reference				
Exp. 2 × Dual-task	-0.06	0.21	-0.29	0.62, 1.42	0.77
Exp. 1 × Dual-task	Reference				
Exp. 2 × Single-task	0.06	0.21	0.29	0.70, 1.61	0.77
Visual condition × Experiment × Number of tasks					
No faces × Exp. 1 × Single-task	Reference				
Faces × Exp. 2 × Dual-task	0.34	0.26	1.29	0.84, 2.35	0.20
No faces × Exp. 1 × Dual-task	Reference				
Faces × Exp. 2 × Single-task	-0.34	0.26	-1.29	0.43, 1.19	0.20
No faces × Exp. 2 × Single-task	Reference				
Faces × Exp. 1 × Dual-task	-0.34	0.26	-1.29	0.43, 1.19	0.20
No faces × Exp. 2 × Dual-task	Reference				
Faces × Exp. 1 × Single-task	0.34	0.26	1.29	0.84, 2.35	0.20
Working memory ^a	0.39	0.12	3.32	1.17, 1.85	0.001
Processing speed ^a	-0.09	0.13	-0.66	0.71, 1.18	0.51
Attention ^a	0.17	0.12	1.36	0.93, 1.49	0.17
Random effects	Variance	SD	Correlations		
Participant ID (Intercept)	0.82	0.90			
Participant ID (Visual condition)	0.99	0.99	-0.49 (Intercept)		
Participant ID (Number of tasks)	0.21	0.46	-0.33 (Intercept), 0.39 (Visual condition)		

Table 1 (continued)

Fixed effects and covariates	Estimate	SE	z-ratio	95% CI	p
Participant ID (Visual condition x Number of tasks)	0.12		0.35	-0.12 (Intercept), -0.54 (Visual condition), -0.57 (Number of tasks)	
Text (Intercept)	0.09		0.29		
Question number (Intercept)	0.13		0.36		

Number of observations: 7308; groups: participant ID = 58, text = 14, question = 9. Confidence intervals were calculated using the Wald method. Model equation: performance ~ visual condition * experiment * number of tasks + working memory_z + processing speed_z + attention_z + (1 + visual condition * number of tasks | participant ID) + (1 | question number) + (1 | text); family = binomial, link function = logit.

^a Estimates are standardized coefficients.

condition and experiment and number of tasks ($\chi^2(1) = 1.66, p = .20$). These results indicate that the effect of visual condition on participants' performance did not vary across the different blocks.

While we hypothesized that *working memory*, *processing speed*, and *attention* would account for variability in the effects of visual cues on memory performance (RQ3), only *working memory* ($\chi^2(1) = 11.04, p < .001$) was found to have a significant effect, whereas *attention* ($\chi^2(1) = 1.85, p = .17$) and *processing speed* ($\chi^2(1) = 0.43, p = .51$) did not.

VIFs for main effects ranged from 1.2 to 7.1, indicating low to moderate collinearity, whereas higher-order interactions had VIFs between 8.7 and 11.2, which is expected in factorial designs. These higher VIFs in interaction terms reflect shared variance but do not compromise the interpretability of main effects, though interactions coefficients should be interpreted with caution.

To evaluate how much variance was explained by the model, we computed marginal and conditional R^2 values. The marginal R^2 was 0.101, indicating that the fixed effects (e.g., visual condition, experiment, number of tasks, and covariates) explained approximately 10.1% of the variance in short-term memory performance. The conditional R^2 was 0.336, showing that the full model, including random effects, explained 33.6% of the variance. Semi-partial R^2 values for individual predictors were very small ($R^2 < 0.001$), highlighting that no single predictor alone accounted for substantial variance, and emphasizing the importance of considering the combined effects of predictors.

3.2. Performance and RTs in the vibrotactile secondary task (RQ2 & RQ3)

The effect of *visual condition* on listening effort was assessed based on participants' performance and RTs in the vibrotactile secondary task (RQ2). The descriptive results for performance based on *visual condition* are depicted in Fig. 3. The GLMM results for performance in the secondary task are shown in Table 2.

The analysis revealed a significant effect of *task condition* on performance in the vibrotactile secondary task ($\chi^2(2) = 21.16, p < .001$), after controlling for *experiment*, *working memory*, *processing speed*, and *attention*. We hypothesized that the presence of static or animated faces as visual cues would result in a reduction in listening effort compared to conditions without faces (RQ2). However, participants made significantly more errors in the vibrotactile secondary task during dual-task conditions with faces compared to no faces (z-ratio = 2.74, $p < .001$).

Additionally, the findings did reveal a significant effect of *experiment* on secondary task performance ($\chi^2(1) = 4.08, p = .04$), after considering *task condition* and the three covariates. Participants made more errors in the vibrotactile secondary task in Experiment 2 than in Experiment 1 (z-ratio = 3.73, $p < .001$).

Furthermore, the two-way interaction of *task condition* and *experiment* was significant ($\chi^2(2) = 78.71, p < .001$). The post hoc analysis showed that participants made significantly more errors in following

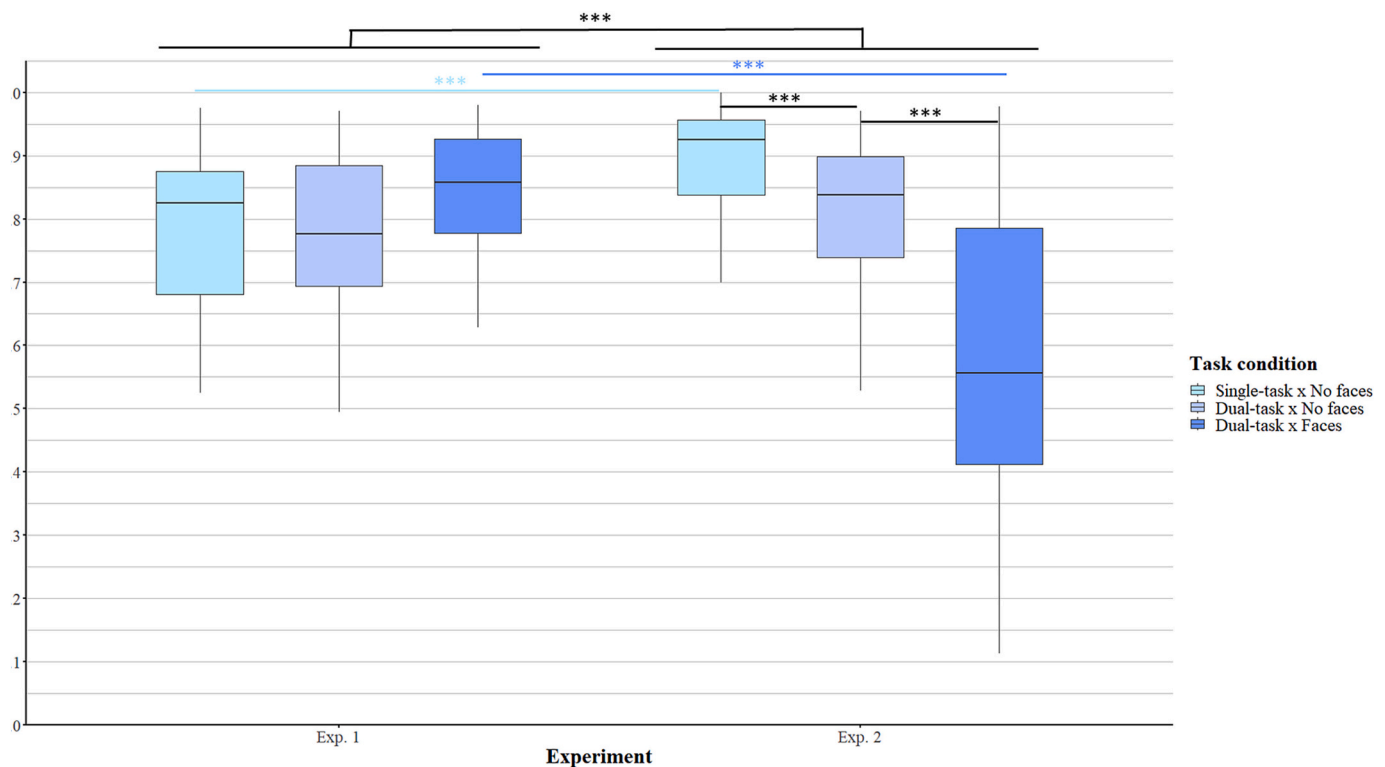


Fig. 3. Performance in the secondary vibrotactile pattern recognition task. Secondary task performance as a function of *task condition* (secondary task in single-task condition without visual cues, in dual-task condition without visual cues, in dual-task condition with static faces, in dual-task condition with animated faces). The boxplots show the data distribution for secondary task performance (proportion correct). The boxes represent the interquartile ranges, while the lines contained within the boxes represent the medians. The points indicate outliers. *** $p < .001$.

conditions: in single-task conditions without faces in Experiment 1 than in Experiment 2 (z -ratio = 3.72, $p = .003$), in dual task conditions without faces than in single-task conditions without faces in Experiment 2 (z -ratio = 4.08, $p < .001$), in dual-task conditions with animated faces than in dual-task conditions without faces (Experiment 2; z -ratio = 7.33, $p < .001$), and in dual-task conditions with animated faces (Experiment 2) than in dual-task conditions with static faces (Experiment 1) (z -ratio = 4.93, $p < .001$).

The present study hypothesized that working memory capacity, processing speed, and attention would account for variability in the effects of visual cues on listening effort (RQ3). However, the analysis indicated that only *attention* ($\chi^2(1) = 7.17$, $p = .007$) was significantly predictive of secondary task performance, while *working memory* ($\chi^2(1) = 1.06$, $p = .30$) and *processing speed* ($\chi^2(1) = 2.42$, $p = .12$) were not.

All covariates showed low VIFs (< 1.7), and the factors and their interaction showed moderate VIFs (GVIF-adjusted values < 2.5). These values indicate that multicollinearity was not a concern for the secondary task performance.

To assess the explanatory power of the model predicting performance in the vibrotactile secondary task, we calculated both marginal and conditional R^2 values. The marginal R^2 was 0.118, indicating that the fixed effects explained approximately 11.8% of the variance in performance. The conditional R^2 was 0.295, suggesting that the full model (including both fixed and random effects) accounted for 29.5% of the variance. Semi-partial R^2 values for each fixed effect were near zero, indicating that no single predictor alone explained a substantial amount of variance in secondary task performance. This suggests that the model's explanatory strength was driven by the combination of predictors and random effects rather than strong individual effects.

The descriptive results for RTs based on *visual conditions* are shown in Fig. 4. The results of the GLMM for RT in the secondary task are presented in Table 3.

Similar to the outcome variable performance, the results indicated a

significant effect of *task condition* on RTs in the vibrotactile secondary task ($\chi^2(2) = 162.20$, $p < .001$), after controlling for *experiment*, *working memory*, *processing speed* and *attention*. A post hoc analysis revealed that RTs were significantly longer in dual-task conditions without faces (z -ratio = 10.58, $p < .001$) compared to single-task conditions without faces. Additionally, RTs were significantly longer in dual-task conditions with faces compared to dual-task conditions without faces (z -ratio = -5.28 , $p < .001$). These results did not align with our hypothesis, as we expected shorter RTs with static or animated faces.

Additionally, similar to the outcome variable secondary task performance, the findings did reveal a significant effect of *experiment* on secondary task RTs ($\chi^2(1) = 0.23$, $p < .001$), after considering *task condition* and the three covariates. Participants reacted significantly slower in Experiment 1 than in Experiment 2 (z -ratio = -3.36 , $p < .001$).

The two-way interaction of *task condition* and *experiment* was significant ($\chi^2(2) = 312.66$, $p < .001$). The post hoc analysis showed that participants RTs were significantly longer in following conditions: in single-task conditions without faces in Experiment 1 than in Experiment 2 (z -ratio = 3.36, $p = .01$), in dual task conditions without faces than in single-task conditions without faces in Experiment 2 (z -ratio = 13.25, $p < .001$), in dual-task conditions without faces than in dual-task conditions with static faces (Experiment 1; z -ratio = 5.40, $p < .001$), in dual-task conditions with animated faces (Experiment 2) than in dual-task conditions without faces in (z -ratio = 11.15, $p < .001$), and in dual-task conditions with animated faces (Experiment 2) than in dual-task conditions with static faces (Experiment 1) (z -ratio = 3.80, $p = .002$).

Finally, the results indicate that *working memory* ($\chi^2(1) = 0.53$, $p = .47$), *attention* ($\chi^2(1) = 0.15$, $p = .69$), and *processing speed* ($\chi^2(1) = 0.03$, $p = .87$) did not have a significant effect on RTs in the secondary task.

All covariates showed low VIFs (< 1.6), and the fixed factors and their interaction showed moderate VIFs (GVIF-adjusted values < 2.4). For secondary task RT, these values indicate that multicollinearity was not an issue.

Table 2

Results from the final GLMM modeling performance in the vibrotactile secondary task as predicted by task condition and experiment.

Fixed effects and covariates	Estimate	SE	z-ratio	95% CI	p
Intercept	1.39	0.22	6.24	2.60, 6.22	0.43
Task condition					
Secondary task in single-task condition	Reference				
Secondary task in dual-task condition without faces	0.03	0.17	0.17	0.74, 1.43	0.86
Secondary task in dual-task condition with faces	0.47	0.20	2.33	1.08, 2.36	0.02
Secondary task in dual-task condition without faces	Reference				
Secondary task in dual-task condition with faces	0.43	0.16	2.74	1.13, 2.11	0.006
Experiment					
1	Reference				
2	1.00	0.27	3.73	1.61, 4.58	<0.001
Task condition x Experiment					
Secondary task in single-task condition x Exp. 1	Reference				
Secondary task in dual-task condition without faces x Exp. 2	-0.82	0.25	-3.22	0.27, 0.72	<0.001
Secondary task in dual-task condition with faces x Exp. 2	-2.44	0.29	-8.29	0.05, 0.16	<0.001
Secondary task in single-task condition x Exp. 2	Reference				
Secondary task in dual-task condition without faces x Exp. 1	0.82	0.25	3.24	1.38, 3.74	<0.001
Secondary task in dual-task condition with faces x Exp. 1	2.43	0.29	8.29	6.41, 20.26	<0.001
Secondary task in dual-task condition without faces x Exp. 1	Reference				
Secondary task in dual-task condition with faces x Exp. 2	-1.61	0.23	-7.15	0.13, 2.11	<0.001
Working memory ^a	0.11	0.10	1.02	0.91, 1.37	0.30
Processing speed ^a	-0.19	0.12	-1.56	0.65, 1.05	0.12
Attention ^a	0.31	0.12	2.67	1.09, 1.71	0.01
Random effects	Variance	SD		Correlations	
Participant ID (Intercept)	0.70	0.83			
Participant ID (Secondary task in dual-task condition without faces)	0.59	0.77		-0.64 (Intercept)	
Participant ID (Secondary task in dual-task condition with faces)	0.89	0.95		-0.33 (Intercept), 0.63 (Task condition)	
Vibration type (Intercept)	0.08	0.28			

Number of observations: 13775; groups: participant ID = 58, vibration type = 4. Confidence intervals were calculated using the Wald method. Model equation: correctness ~ task condition * experiment + working memory_z + processing speed_z + attention_z + (1 + task condition | participant ID) + (1 | vibration type); family = binomial, link function = logit.

^a Estimates are standardized coefficients.

For the secondary task RT model, the marginal R^2 was 0.047, indicating that fixed effects explained approximately 4.7% of the variance in RTs. The conditional R^2 was 0.097, meaning the model including both fixed and random effects accounted for 9.7% of the variance. Despite the significance of task condition and its interactions, semi-partial R^2 values showed that no single fixed effect accounted for a substantial proportion of variance (all $R^2 \approx 0$), suggesting the effects were distributed in

magnitude.

3.3. Effect of visual condition on subjective experience

Participants answered questions about task demands, performance, and attention to visual cues to investigate how the visual condition affected their subjective impressions.

In Experiment 1, which involved static faces, one of the 30 participants (3.3%) reported not having paid attention to the static faces, while two participants (10%) indicated only partial or no attention to the static faces. In contrast, 14 participants (46.7%) rather agreed that they were paying attention, while the other 13 (43.3%) fully agreed that they were paying attention to the static faces. The descriptive data for questions 2 to 7 is shown in Fig. 5. On which condition was easier, 13 participants (43.3%) reported that it was easier with static faces, 10 (33.3%) stated that it was easier without static faces, and seven participants (23.3%) indicated that there was no difference. Regarding the first open-ended question, 24 participants (80.0%) stated that they lacked attention at various times during the experiment. In the second open-ended question, seven participants (23.3%) provided critical suggestions for improving the experiment. One suggestion was to present the vibration on to other parts of the body, such as the arm, to determine further effects. Some participants also recommended adding dynamic faces to improve engagement and to enrich the experience.

In Experiment 2, three out of the 28 participants did not fill out the self-report questionnaire regarding the animated faces. Therefore, the following descriptive results are based on a subsample of 25 participants. Nine participants (36%) reported not having paid attention to the animated faces, and two (8%) outlined neither paying attention nor being attentive to the animated faces. In contrast, nine participants (36%) rather agreed that they were paying attention, while the other five (20%) fully agreed that they were paying attention to the animated faces. Fig. 5 displays the descriptive data for questions 2 to 7. On which condition was easier, 9 participants (36%) reported that it was easier with animated faces, 11 (44%) stated that it was easier without animated faces, and 5 participants (20%) indicated that there was no difference. Regarding the first open-ended question, 24 participants (88%) stated that they lacked attention at various times during the experiment. In the second open-ended question, six participants (24%) provided critical suggestions for improving the experiment. Some suggested using different animated faces for each story to potentially aid memory retention, along with making vibration patterns more distinct. Further suggestions included considering questions related to learning progress and recency effects.

4. Discussion

This study investigated whether static or animated faces improve short-term memory performance and reduce listening effort during two-talker conversations in non-noisy settings. Additionally, we investigated the modulating effect of listeners' individual working memory capacity, attentional functions, and processing speed, which were included as covariates in the analysis.

First, addressing short-term memory performance (RQ1), we investigated the impact of static or animated faces compared to conditions without faces on short-term memory performance, as measured by the HTR task (Schlittmeier et al., 2023), both independently and concurrently with the vibrotactile secondary task. We hypothesized that static or animated faces would improve participants' short-term memory performance for the two-talker conversations in non-noisy listening settings. Contrary to our hypothesis, participants did not recall more information when animated faces or static faces were present in the HTR task than when there were no faces after controlling for working memory capacity, attentional functions, and processing speed.

Second, addressing listening effort (RQ2), we investigated the effect of static or animated faces, compared to listening without faces, on

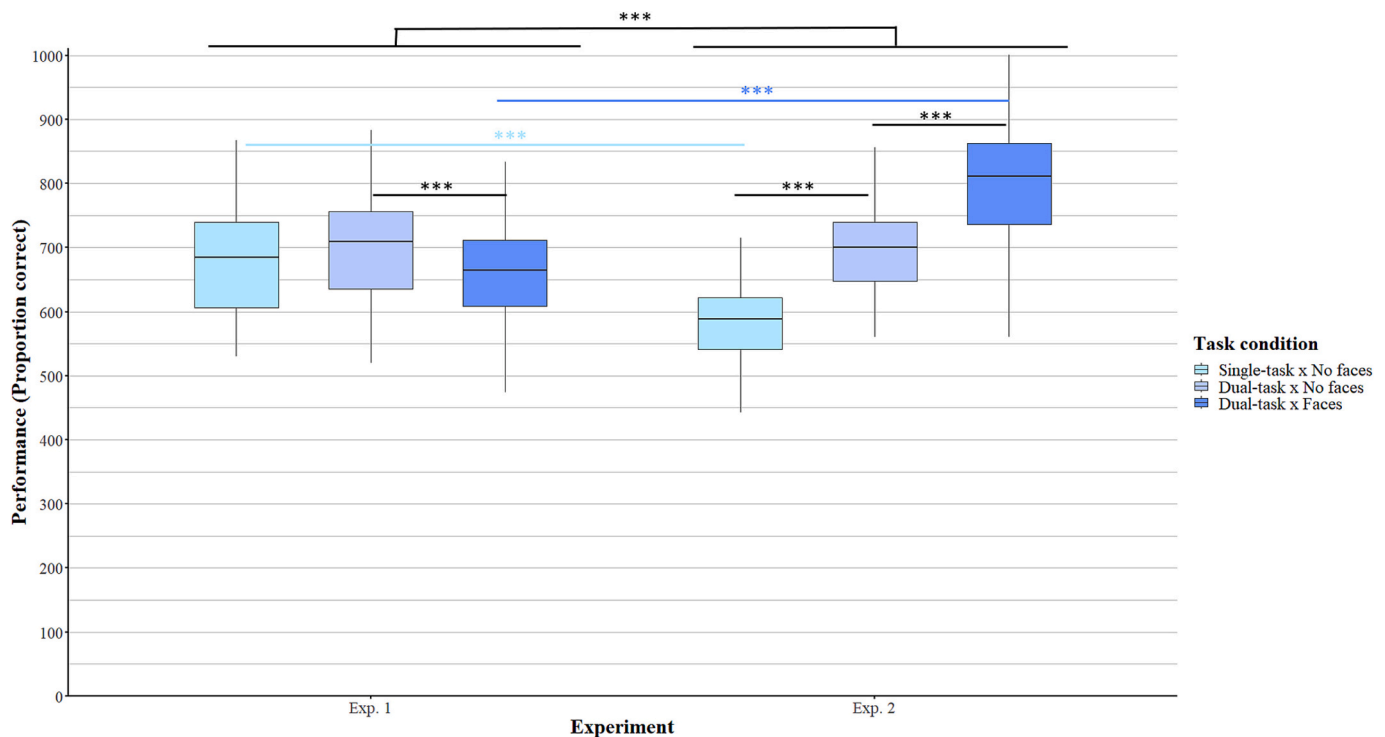


Fig. 4. Reaction times in milliseconds in the secondary vibrotactile pattern recognition task. The boxplots illustrate the data distributions for the secondary task reaction times as a function of *task condition* (secondary task in single-task condition without faces, dual-task condition without faces, dual-task condition with static faces, dual-task condition with animated faces). The boxes represent the interquartile ranges, while the lines contained within the boxes represent the medians. *** $p < .001$.

performance and RTs during the vibrotactile secondary task. Although short-term memory performance was similar across all conditions, differences in listening effort emerged, as reflected in secondary task performance during dual-task conditions when cognitive covariates were considered. For secondary task performance, animated faces led to poorer performance compared to both static faces and no-face conditions, with no significant difference between static faces and no faces. Regarding RTs, responses were the longest with animated faces, followed by conditions without faces, and the shortest with static faces. Furthermore, participants' subjective impressions of the static and animated faces from the post-experimental questionnaire align with the objective RT findings. In Experiment 1, most participants found the task less demanding and easier to perform with static faces than without, whereas in Experiment 2, the majority indicated that the listening task was harder with animated faces than without. In both experiments, only a small proportion of participants reported that the two conditions (with and without visual cues) were equally demanding.

Our hypothesis was that the presence of faces as visual cues would reduce listening effort, as indicated by fewer errors and/or shorter RTs on the vibrotactile secondary task when static or animated faces were present, compared to when no faces were shown. While this held true for static faces, which reduced listening effort compared to the listening condition without faces, animated faces even increased listening effort and represented the visual condition with the highest listening effort in the present study. It is important to note that this effect was observed only for secondary task RTs and not for secondary task performance. Static faces sped up responses but did not change overall accuracy compared to no faces, implying they helped participants process information more efficiently without compromising task success. These findings are consistent with the FUEL framework (Pichora-Fuller et al., 2016), suggesting that animated faces may require greater mental resource allocation, potentially diverting attention from the auditory input.

The latter finding is consistent with Fraser et al. (2010), who in noisy

listening settings found that seeing lip movements in addition to hearing a voice increases listening effort. Fraser et al. (2010) also used a dual-task paradigm, where participants performed speech recognition in noise as the primary task, with a vibrotactile secondary task similar to ours. As discussed in the introduction, we proposed that the increase in listening effort may be due to the listener's task of simultaneously monitoring and integrating information from two modalities - auditory and visual - into one unified percept. Our results support this idea, showing that auditory speech accompanied by animated faces with lip sync results in greater cognitive load (Brown & Strand, 2019), as evidenced by longer RTs and lower performance in the present study. This may even hold true for listening settings without additional background noise.

Our finding that animated faces increase listening effort may appear inconsistent with prior studies (e.g., Rudner et al., 2016; Sommers & Phelps, 2016), which suggest that being able to see and hear a talker reduces listening effort. However, a key difference lies in the complexity of the tasks. Our study involved a focus on memory of conversational content and dual-task demands. In contrast, Rudner et al. (2016) and Sommers and Phelps (2016) used simpler speech stimuli and single-task paradigms, all in noisy listening conditions. This discrepancy highlights the need for further research to explore the role of task complexity in moderating the effects of visual cues. Thus, it is likely that our listening task in dual-task settings involved a higher cognitive load due to its inherent complexity (Kintsch, 1988), making animated faces an additional burden rather than a benefit. The findings suggest that static faces captured and sustained the listener's attention on the current talker, facilitating faster responses without imposing additional cognitive load from face or lip movements.

An alternative explanation for the increased listening effort observed with animated faces may be related to attentional capture by visual motion. Research suggests that animate motion—perceived as self-directed—can draw attention even when irrelevant to the task (Pratt, Radulescu, Guo, & Abrams, 2010), possibly competing with cognitive

Table 3

Results from the final GLMM modeling reaction times in the vibrotactile secondary task as predicted by the task condition and experiment.

Fixed effects and covariates	Estimate	SE	z-ratio	95% CI	p
Intercept	6.51	0.05	124.84	609.39, 747.72	<0.001
Task condition					
Secondary task in single-task condition	Reference				
Secondary task in dual-task condition without faces	0.02	0.01	1.86	0.99, 1.05	0.06
Secondary task in dual-task condition with faces	−0.03	0.01	−2.08	0.95, 0.99	0.04
Secondary task in dual-task condition without faces	Reference				
Secondary task in dual-task condition with faces	−0.05	0.01	−5.39	0.93, 0.97	<0.001
Experiment					
1	Reference				
2	−0.16	0.05	−3.36	0.78, 0.94	<0.001
Task condition x Experiment					
Secondary task in single-task condition x Exp. 1	Reference				
Secondary task in dual-task condition without faces x Exp. 2	0.15	0.02	7.91	1.12, 1.19	<0.001
Secondary task in dual-task condition with faces x Exp. 2	0.33	0.02	17.00	1.34, 1.45	<0.001
Secondary task in single-task condition x Exp. 2	Reference				
Secondary task in dual-task condition without faces x Exp. 1	−0.15	0.02	−7.90	0.83, 0.89	<0.001
Secondary task in dual-task condition with faces x Exp. 1	−0.33	0.02	−16.98	0.69, 0.75	<0.001
Secondary task in dual-task condition without faces x Exp. 1	Reference				
Secondary task in dual-task condition with faces x Exp. 2	0.18	0.02	12.06	1.17, 1.24	<0.001
Working memory ^a	0.02	0.02	0.73	0.97, 1.06	0.47
Processing speed ^a	−0.003	0.02	−0.16	0.95, 1.04	0.87
Attention ^a	−0.01	0.02	−0.39	0.94, 1.04	0.69
Random effects	Variance			SD	
Participant ID	0.01			0.07	
Vibration type	0.001			0.03	

Number of observations: 9794; groups: participant ID = 58, vibration type = 4. Confidence intervals were calculated using the Wald method. Model equation: $rt \sim \text{task condition} * \text{experiment} + \text{working memory}_z + \text{processing speed}_z + \text{attention}_z + (1 | \text{participant ID}) + (1 | \text{vibration type})$; family = Gamma, link function = log.

^a Estimates are standardized coefficients.

resources required for auditory processing. However, motion in itself does not always capture attention unless it signals the appearance of a new perceptual object or disrupts perceptual grouping (Hillstrom & Yantis, 1994). Since the facial animations in our study were continuous and predictable, they may not have triggered attentional capture in the same way abrupt or novel movements would. Still, their persistent motion could have subtly interfered with task focus, particularly in a dual-task context. Applying our findings to the ELU model (Rönnerberg et al., 2013), we suggest that the use of animated faces in the present

study may introduce a mismatch compared to conditions with no faces and static faces. A potential cause of this mismatch could arise from the lip movements, which are close to the spoken words but do not align perfectly on a syllable-by-syllable basis, which we also discuss in our limitations. According to the ELU model, when such a mismatch occurs, it places additional demands on cognitive resources. We argue that this forces the listener to process both auditory and visual streams of information more carefully. As a result, this mismatch may lead to attentional capture, as the lip sync draws the listener's focus, and to increased perceptual load, due to the demands of both the listening task and the vibrotactile secondary task. Together, these factors place a higher cognitive load on the listener, thereby engaging the explicit cognitive processing route. This route is slower, more focused, and more resource-intensive, leaving fewer cognitive resources available for other cognitive tasks performed simultaneously. Consequently, increased cognitive load may reduce performance, as evidenced by poorer performance and higher RTs in the vibrotactile secondary task, as well as a reduced ability to manage dual-task demands. Future research should explore additional hypotheses to further clarify the underlying mechanisms of listening effort, particularly in contexts where multiple sources of information—auditory and visual—are involved.

In line with dual-task theory, our findings show that cognitive processing resources are limited and are allocated across tasks when they are performed in parallel, as stated before (Broadbent, 1958; Kahneman, 1973; Navon & Miller, 2002; Pashler, 1994). Specifically, participants performed better on both the primary listening task and the secondary task, and RTs were shorter in the secondary task when these tasks were carried out separately compared to when they were performed in parallel. This suggests that in single-task conditions, participants were able to allocate all their cognitive resources to one task, resulting in better performance. In contrast, when tasks were performed concurrently, the increased cognitive load likely led to a distribution of resources, which in turn negatively affected performance and RTs on both tasks.

Given that cognitive resources are finite and must be allocated across tasks, individual differences in cognitive capacity – such as working memory, attentional functions, and processing speed – are crucial for understanding variations in listening effort and short-term memory performance. As outlined in the introduction, these cognitive abilities play an important role for short-term memory performance and listening effort. Therefore, they were used as covariates in our analysis to account for their influence on the effects of static and animated faces on memory of conversational content and listening effort.

Third, addressing the role of individual cognitive abilities (RQ3), the present study provides evidence that individual differences in working memory capacity predict performance in short-term memory tasks like the HTR, likely because such tasks require both the temporary storage and concurrent mental processing of conversational content. However, attentional functions and processing speed did not significantly influence short-term memory performance after accounting for other factors in the model. These results contrast with those of Mohanathan et al. (2024), who reported that individual differences in attentional functions and processing speed were related to memory of conversational content, while working memory was not. One possible explanation for this discrepancy may lie in the added visual cues in our study. Specifically, the animated cues introduced dynamic visual information that required participants to track movement or changes in real-time. This additional visual input placed greater demands on working memory, as it must not only hold and update auditory information but also integrate and process the dynamic visual stimuli. In contrast, the study by Mohanathan et al. (2024) did not incorporate any visual cues, which may explain the different patterns of results.

Furthermore, our examination of listening effort in non-noisy listening settings revealed that only attentional functions were significantly associated with performance in the secondary task, although they did not influence RTs. Interestingly, neither working memory capacity nor processing speed emerged as significant covariates for secondary

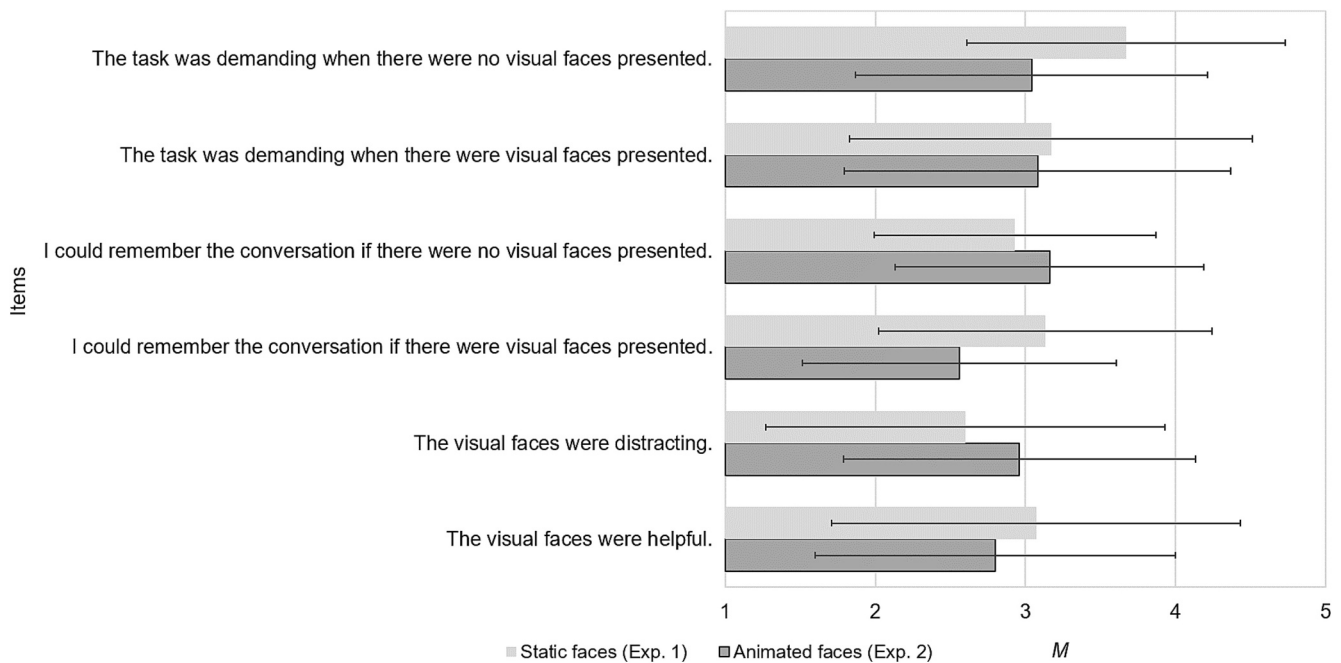


Fig. 5. Descriptive statistics of the self-report questionnaire. Means (with standard deviations) from the self-report questionnaire regarding static faces in Experiment 1 ($n = 30$) and animated faces in Experiment 2 ($n = 25$). A five-point Likert scale was used for subjective impressions of statements, with answers ranging from “do not agree at all” (1) to “fully agree” (5). The questionnaire was presented in German to the participants in the study, but is provided in English in this figure for readability.

task performance and RTs. This suggests that attention specifically affected the accuracy of responses without influencing the speed at which they were made. Again, these findings stand in contrast to those of Mohanathan et al. (2024). Specifically, in their GLMM analysis for performance on the same vibrotactile secondary task, working memory emerged as a significant covariate for secondary task performance, and attentional functions and processing speed were significant covariates for RTs rather than overall task performance (Mohanathan et al., 2024). In addition to the above-mentioned differences in visual cues between our study and Mohanathan et al. (2024), another key distinction lies in our model specification. Specifically, we included participant ID as a random intercept or with random slopes for the within-subject factors in our models. This allowed us to account not only for individual differences in baseline performance but also for variability in how participants responded to different conditions. By modeling both intercepts and slopes, our approach may have captured more nuanced individual response patterns, potentially contributing to differences from the findings of Mohanathan et al. (2024). Importantly, as discussed by Barr et al. (2013), models that include only random intercepts and omit random slopes can lead to inflated Type I error rates. Including both random intercepts and random slopes may improve model precision and validity, enabling a clearer examination of how cognitive functions—such as attention—relate to secondary task performance. These findings underscore the importance of considering the specific task demands and model specifications when interpreting the role of cognitive functions in listening effort and memory performance.

5. Limitations

A central limitation of this study is the limited statistical power for detecting Level 2 and cross-level interaction effects. Based on our design and sample size, the study did not allow sufficient power for detecting medium effects (e.g., $\beta = 0.3$) at these levels. Achieving adequate power would have required a substantially larger Level 2 sample size (potentially exceeding 200 participants) or increased Level 1 data density (e.g., more texts and/or questions, respective vibrotactile patterns; cp. Arend

& Schäfer, 2019; Sommet & Morselli, 2021). However, the study was sufficiently powered to detect Level 1 effects, which were the primary focus of our research question. As such, null findings related to Level 2 and interaction effects should be interpreted cautiously. Importantly, the observed effect sizes in our final models—as reflected in marginal R^2 values—were relatively low across all tasks. Fixed effects explained 10.1% of the variance in primary task performance, 11.8% in the vibrotactile secondary task, and 4.7% in secondary task reaction times. Conditional R^2 values, reflecting the contribution of both fixed and random effects, ranged from 9.7% to 33.6%, highlighting the role of individual variability. Moreover, semi-partial R^2 values for individual predictors were near zero, indicating that no single fixed effect contributed substantially to the explained variance. In GLMMs, semi-partial R^2 values for individual predictors can be small even when predictors are meaningful. This is because variance is often shared among predictors or absorbed by the random-effects structure. As shown in our models, VIFs indicate that multicollinearity among main effects was low to moderate, and higher VIFs for interactions are expected in factorial designs. Thus, the small semi-partial R^2 values most likely reflect shared variance and variance absorbed by the random effects-structure. These findings suggest that the actual effect sizes were smaller and more diffuse than the standardized coefficients ($\beta = 0.30$) assumed in our simulation-based power analysis. While our power assumptions aligned with conventions for multilevel modeling (Arend & Schäfer, 2019; Maas & Hox, 2005), the observed R^2 values underscore the importance of using empirically derived effect size estimates when planning future studies. We recommend that future work consider preliminary estimates (e.g., from pilot data or prior studies) to better calibrate power simulations and expectations.

Another limitation of the current study concerns the interpretation of visual cues. The video and audio were technically synchronized at the level of file playback – that is, the temporal alignment of the two streams was ensured by the presentation software and calibrated replay devices. However, the animated faces' lip movements did not always correspond precisely to the speech on a syllable-by-syllable level. This was a known limitation of the tracking system: an iPhone using ARKit (Apple Inc,

2020) and LiveLinkFace software (Epic Games, 2020) with the depth camera of the iPhone. It operates on a given set of morph-targets to transform the face during animation, which potentially did not capture all intricacies of the movement. This may have resulted in subtle perceptual asynchronies, despite the presence of technical synchrony. Consequently, the animated faces may have introduced variability in audiovisual integration, potentially influencing listening effort in ways not solely attributable to task difficulty or attentional capture.

Furthermore, the external validity of the findings should be considered with caution. The use of animated faces and the controlled laboratory setting differ notably from natural face-to-face interactions and real-world listening environments. Although the study provides valuable insights into audiovisual processing and listening effort, generalizing these results to more ecologically valid contexts requires further investigation.

6. Conclusions

In summary, the present study revealed that animated faces can increase listening effort compared to static faces or conditions without faces. This was evidenced by poorer performance and longer RTs in the secondary task. In contrast, static faces appeared to reduce listening effort, as indicated by shorter RTs in the secondary task.

Previous listening research has primarily focused on less complex cognitive tasks (e.g., single word recognition) in adverse listening conditions. However, it is essential to expand this scope to include more complex and plausible cognitive tasks, involving short-term memory for conversational speech in non-noisy listening conditions. Moreover, further exploration of the nuanced role of visual cues and individual cognitive differences is needed. By broadening the scope of our investigation beyond the examination of simple cognitive tasks in adverse listening conditions to encompass more plausible listening tasks and non-noisy listening conditions, we aim to bridge the gap between controlled laboratory and close-to-real-world settings, contributing to the growing body of research on listening effort.

CRedit authorship contribution statement

Chinthusa Mohanathan: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Formal analysis. **Plamenna B. Koleva**: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis. **Jonathan Ehret**: Writing – review & editing, Resources. **Andrea Bönsch**: Writing – review & editing. **Janina Fels**: Writing – review & editing, Funding acquisition. **Torsten W. Kühlen**: Writing – review & editing, Funding acquisition. **Sabine J. Schlittmeier**: Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Ethics approval and consent to participate

The research has complied with the relevant national regulations, institutional policies, and by the tenets of the Helsinki Declaration. The research was approved by the local ethics committee at the Philosophical Faculty of the RWTH Aachen University (2021_08_FB7_RWTH AACHEN). Furthermore, all participants in this study gave written, informed consent.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (OpenAI) and LanguageTool (LanguageTool GmbH) to improve the clarity and readability of the text and to check for inadvertent errors. ChatGPT was also used to assist in documenting R scripts (e.g., refining code comments). After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the

content of the published article.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): 444724862 - “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” (SPP2236). Plamenna B. Koleva's contribution was supported by a grant from the Head-Genuit Foundation [P-16/10W] as was Sabine Schlittmeier's contribution in securing funding.

Declaration of competing interest

The authors declare that no competing interests exist.

Acknowledgments

The authors thank Chencheng Lei and Jian Pan for programming the experiments, Cosima A. Ermert for helping with the audio reproduction and the valuable project collaboration, Sara C. Neff for her support with data collection and preparation. We are also grateful to Dr. Isabel S. Schiller for her statistical advice, and to Jennifer Klütsch and Luise Haehn for their valuable input and support regarding the simulation-based power analysis.

Author disclosure

The findings of Experiment 1 were presented at the 23rd ESCoP conference held in Porto, Portugal. The theory and methodology of Experiment 2 were presented at the 50th DAGA conference held in Hannover, Germany.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.actpsy.2026.106295>.

Data availability

The raw data supporting the conclusions of this article are available in the supplementary material.

References

- Aarts, E., Verhage, M., Veenvelt, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience*, 17(4), 491–496. <https://doi.org/10.1038/nn.3648>
- Anderson Gosselin, P., & 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. [https://doi.org/10.1044/1092-4388\(2010/10-0069\)](https://doi.org/10.1044/1092-4388(2010/10-0069))
- Apple Inc. (2020). ARKit (version 3.5) [software]. <https://developer.apple.com/augmented-reality/arkit/>.
- Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, 24(1), 1–19. <https://doi.org/10.1037/met0000195>
- Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2), 12–28.
- Baddeley, A. (2020). Working memory. In *Memory* (3. Aufl). Routledge.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.675558>
- Broadbent, D. E. (1958). Perception and communication. Pergamon Press. <https://doi.org/10.1037/10037-000>

- Brown, V. A., & Strand, J. F. (2019). About face: Seeing the talker improves spoken word recognition but increases listening effort. *Journal of Cognition*, 2(1), 44. <https://doi.org/10.5334/joc.89>
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4), 450–466. [https://doi.org/10.1016/S0022-5371\(80\)90312-6](https://doi.org/10.1016/S0022-5371(80)90312-6)
- Dryden, A., Allen, H. A., Henshaw, H., & Heinrich, A. (2017). The association between cognitive performance and speech-in-noise perception for adult listeners: A systematic literature review and Meta-analysis. *Trends in Hearing*, 21, 1–21. <https://doi.org/10.1177/2331216517744675>
- Eckert, M. A., Matthews, L. J., Vaden, K. I., & Dubno, J. R. (2024). Executive function associations with audibility-adjusted speech perception in noise. *Journal of Speech, Language, and Hearing Research*, 67(12), 4811–4828. <https://doi.org/10.1044/2024.JSLHR-24-00333>
- Ehret, J., Bönsch, A., Schiller, I. S., Breuer, C., Aspöck, L., Fels, J., ... Kuhlén, T. W. (2024). Audiovisual coherence: Is embodiment of background noise sources a necessity? 2024 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW). Orlando, FL, USA, 2024, 61–67. <https://doi.org/10.1109/VRW62533.2024.00017>
- Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11(1), 19–23. <https://doi.org/10.1111/1467-8721.00160>
- Epic Games. (2020). Live link face (version 1.0) [Mobile application]. <https://apps.apple.com/us/app/live-link-face/id1495370836>
- Ermert, C. A., Mohanathan, C., Ehret, J., Schlittmeier, S. J., Kuhlén, T., & Fels, J. (2023). AuViST - an audio-visual speech and text database for the heard-text-recall paradigm [dataset]. RWTH Aachen University. <https://doi.org/10.18154/RWTH-2023-05543>
- 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>
- Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research*, 53(1), 18–33. [https://doi.org/10.1044/1092-4388\(2009\)08-0140](https://doi.org/10.1044/1092-4388(2009)08-0140)
- Gagné, J.-P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing*, 21. <https://doi.org/10.1177/2331216516687287>, 233121651668728.
- Hillstrom, A. P., & Yantis, S. (1994). Visual motion and attentional capture. *Perception & Psychophysics*, 55(4), 399–411. <https://doi.org/10.3758/BF03205298>
- Homman, L., Danielsson, H., & Rönnerberg, J. (2023). A structural equation mediation model captures the predictions amongst the parameters of the ease of language understanding model. *Frontiers in Psychology*, 14. <https://doi.org/10.3389/fpsyg.2023.1015227>
- Howard, C. S., Munro, K. J., & Plack, C. J. (2010). Listening effort at signal-to-noise ratios that are typical of the school classroom. *International Journal of Audiology*, 49(12), 928–932. <https://doi.org/10.3109/14992027.2010.520036>
- Imhof, M. (2010). What is going on in the mind of a listener? The cognitive psychology of listening. In A. D. Wolvin (Hrsg.) (Ed.), *Listening and human communication in the 21st century* (1. Aufl, pp. 97–126). Wiley. <https://doi.org/10.1002/9781444314908.ch4>
- Jaeger, B. (2017). r2glmm: Computes R squared for mixed (multilevel) models. <https://CRAN.R-project.org/package=r2glmm>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446. <https://doi.org/10.1016/j.jml.2007.11.007>
- Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163–182. <https://doi.org/10.1037/0033-295X.95.2.163>
- Lee, S. P., Badler, J. B., & Badler, N. I. (2002). *Eyes alive. Proceedings of the 29th annual conference on Computer graphics and interactive techniques* (pp. 637–644). <https://doi.org/10.1145/566570.566629>
- Lemke, U., & Besser, J. (2016). Cognitive load and listening effort: A working model and its application in old age. *Ear and Hearing*, 37, 77S–84S.
- Lenth, R. V. Emmeans: Estimated marginal means. Aka Least-Squares Means. (R package) [Computer software]. <https://cran.r-project.org/package=emmeans>
- Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01171>
- Lüdtke, et al. (2021). Performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1(3), 86–92. <https://doi.org/10.1027/1614-2241.1.3.86>
- Marsja, E., Holmer, E., Stenbäck, V., Micula, A., Tirado, C., Danielsson, H., & Rönnerberg, J. (2025). Fluid intelligence partially mediates the effect of working memory on speech recognition in noise. *Journal of Speech, Language, and Hearing Research*, 68(1), 399–410. <https://doi.org/10.1044/2024.JSLHR-24-00465>
- Marsja, E., Stenbäck, V., Moradi, S., Danielsson, H., & Rönnerberg, J. (2022). Is having hearing loss fundamentally different? Multigroup structural equation modeling of the effect of cognitive functioning on speech identification. *Ear and Hearing*, 43(5), 1437–1446. <https://doi.org/10.1097/AUD.0000000000001196>
- McGarrigle, R., 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. <https://doi.org/10.3109/14992027.2014.890296>
- Mishra, S., Lunner, T., Stenfelt, S., Rönnerberg, J., & Rudner, M. (2013). Visual information can hinder working memory processing of speech. *Journal of speech, language, and hearing research: JSLHR*, 56(4), 1120–1132. [https://doi.org/10.1044/1092-4388\(2012\)12-0033](https://doi.org/10.1044/1092-4388(2012)12-0033)
- Mohanathan, C., Ermert, C. A., Fels, J., Kuhlén, T. W., & Schlittmeier, S. J. (2025). Exploring short-term memory and listening effort in two-talker conversations: The influence of soft and moderate background noise. *PLoS One*, 20(2), Article e0318821. <https://doi.org/10.1371/journal.pone.0318821>
- Mohanathan, C., Fels, J., & Schlittmeier, S. J. (2024). Listening to two-talker conversations in quiet settings: The role of listeners' cognitive processing capabilities for memory and listening effort. *Scientific Reports*, 14, 22764. <https://doi.org/10.1038/s41598-024-74085-1>
- Moosbrugger, H., & Oehlschlägel, J. (2011). *Frankfurter Aufmerksamkeits-Inventar 2 (FAIR-2)*. Bern: Verlag Hans Huber.
- Navon, D., & Miller, J. (2002). Queuing or sharing? A critical evaluation of the single-bottleneck notion. *Cognitive Psychology*, 44(3), 193–251. <https://doi.org/10.1006/cogp.2001.0767>
- Ng, E. H. N., Rudner, M., Lunner, T., Pedersen, M. S., & Rönnerberg, J. (2013). Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. *International Journal of Audiology*, 52(7), 433–441. <https://doi.org/10.3109/14992027.2013.776181>
- Oswald, W. D., & Roth, E. (1987). *Der Zahlen-Verbindungs-Test (ZVT). Ein sprachfreier Intelligenz-Test zur Messung der "kognitiven Leistungsgeschwindigkeit"*. Göttingen: Hogrefe.
- Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, 116(2), 220–244. <https://doi.org/10.1037/0033-2909.116.2.220>
- Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1–2), 8–13. <https://doi.org/10.1016/j.jneumeth.2006.11.017>
- Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., ... Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, 37(1), 5S–27S. <https://doi.org/10.1097/AUD.0000000000000312>
- Picou, E. M., & Ricketts, T. A. (2014). The effect of changing the secondary task in dual-task paradigms for measuring listening effort. *Ear and Hearing*, 35(6), 611–622. <https://doi.org/10.1097/AUD.0000000000000055>
- Picou, E. M., Ricketts, T. A., & Hornsby, B. W. Y. (2011). Visual cues and listening effort: Individual variability. *Journal of Speech, Language, and Hearing Research*, 54(5), 1416–1430. [https://doi.org/10.1044/1092-4388\(2011\)10-0154](https://doi.org/10.1044/1092-4388(2011)10-0154)
- Picou, E. M., Ricketts, T. A., & Hornsby, B. W. Y. (2013). How hearing aids, background noise, and visual cues influence objective listening effort. *Ear and Hearing*, 34(5), e52–e64. <https://doi.org/10.1097/AUD.0b013e31827f0431>
- Pratt, J., Radulescu, P. V., Guo, R. M., & Abrams, R. A. (2010). It's alive! Animate motion captures visual attention. *Psychological Science*, 21(11), 1724–1730. <https://doi.org/10.1177/09567976103874>
- R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rönnerberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., ... Rudner, M. (2013). The ease of language understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, 7. <https://doi.org/10.3389/fnys.2013.00031>
- Rudner, M., Mishra, S., Stenfelt, S., Lunner, T., & Rönnerberg, J. (2016). Seeing the talker's face improves free recall of speech for young adults with Normal hearing but not older adults with hearing loss. *Journal of Speech, Language, and Hearing Research*, 59(3), 590–599. <https://doi.org/10.1044/2015.JSLHR-H-15-0014>
- Salathouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103(3), 403–428. <https://doi.org/10.1037/0033-295X.103.3.403>
- Satori, M. (2021). Effects of working memory on L2 linguistic knowledge and L2 listening comprehension. *Applied Psycholinguistics*, 42(5), 1313–1340. <https://doi.org/10.1017/S0142176421000345>
- Schiller, I. S., Aspöck, L., & Schlittmeier, S. J. (2023). The impact of a speaker's voice quality on auditory perception and cognition: A behavioral and subjective approach. *Frontiers in Psychology*, 14. <https://doi.org/10.3389/fpsyg.2023.1243249>
- Schlittmeier, S. J., Mohanathan, C., Schiller, I. S., & Liebl, A. (2023). 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 [dataset]. RWTH Aachen University. <https://doi.org/10.18154/RWTH-2023-05285>
- Sommers, M. S., & Phelps, D. (2016). Listening effort in younger and older adults: A comparison of auditory-only and auditory-visual presentations. *Ear and Hearing*, 37(1), 62S–68S. <https://doi.org/10.1097/AUD.0000000000000322>
- Sommet, N., & Morselli, D. (2021). Keep calm and learn multilevel linear modeling: A three-step procedure using SPSS, Stata, R, and Mplus. *International Review of Social Psychology*, 34(1), 24. <https://doi.org/10.5334/irsp.555>
- Strand, J. F., Brown, V. A., & Barbour, D. L. (2020). Talking points: A modulating circle increases listening effort without improving speech recognition in young adults. *Psychonomic Bulletin & Review*, 27(3), 536–543. <https://doi.org/10.3758/s13423-020-01713-y>
- Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging*, 24, 761–766. <https://doi.org/10.1037/a0014802>

- Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, 28(2), 127–154. [https://doi.org/10.1016/0749-596X\(89\)90040-5](https://doi.org/10.1016/0749-596X(89)90040-5)
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37(3), 498–505. <https://doi.org/10.3758/BF03192720>
- Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, 58 (3), 475–482. <https://doi.org/10.1007/BF03395630>
- World Health Organization. (1991). *Report of the Informal Working Group on Prevention of Deafness and Hearing Impairment Programme Planning, Geneva, 18-21 June 1991*. WHO/PDH/91.1 (pp. 1–22). World Health (Organization) <https://iris.who.int/handle/10665/58839>.