Influence of speech characteristics on working memory and perceived workload: Findings from experimental listening experiments

Applied Acoustics 255 (2027) 111511

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Influence of speech characteristics on working memory and perceived workload: Findings from experimental listening experiments

Elin Hedlund*, Laura Lindel, Wolfgang Kropp

Division of Applied Acoustics, Department of Architechture and Civil Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

HIGHLIGHTS

- Speech exposure reduced working memory performance compared to silence.
- Semantic content had a negligible impact on working memory performance.
- The impact of speech exposure increased with sound level.
- Subjective ratings increased despite small or absent performance differences.

ARTICLE INFO

Keywords:
Speech exposure
Open office
Working memory
Perceived workload
Listening experiment

ABSTRACT

Exposure to background speech is known to impair working memory performance, a concern particularly relevant in open-plan office environments. This study aimed to examine how specific characteristics of speech, including semantic content and sound pressure level, affect cognitive performance and perceived workload. Controlled laboratory listening experiments were conducted within a common experimental framework, allowing different acoustic characteristics of speech to be compared directly. Participants completed Operation Span tasks (OSPAN) while being exposed to binaural speech signals. Perceived workload was assessed using the NASA Task Load Index (NASA-TLX). In Listening Experiment A, the role of semantic content was investigated using conditions consisting of intelligible speech, locally time-reversed speech, continuous noise, and silence. Listening Experiments B1 and B2 focused on the effects of varying sound pressure levels in speech conditions. Overall, exposure to speech resulted in decreased cognitive performance relative to silence, whereas continuous noise and low-level speech yielded performance outcomes comparable to silent conditions. No significant effect of semantic content was observed in Listening Experiment A. In Listening Experiments B1 and B2, performance declined as sound levels increased, with diminishing effects observed at lower levels. Subjective workload ratings varied across conditions, even where objective performance differences were not detected. These findings suggest that background speech introduces a cognitive cost that may be compensated for in short term through increased mental effort, potentially contributing to long-term fatigue.

1. Introduction

Open-plan offices are now a dominant feature of modern workplaces, driven by expectations of flexibility and efficient space utilization. Although these layouts can facilitate communication and collaboration, they also expose employees to a continuous mix of background sounds, including conversational speech, office equipment, and general environmental noise, as well as constant visual stimuli from surrounding activity. Among these sound sources, background speech has repeatedly been identified as the most intrusive and cognitively disruptive, impairing concentration and task performance [1–3]. Because many office tasks rely on sustained attention and working memory, exposure to speech can interfere with the processes required to maintain and manipulate information during task performance [4,5]. Understanding the mechanisms and extent to which background speech affects cognition is therefore essential for designing work environments that support both well-being and productivity.

* Corresponding author.
Email address: elin.hedlund@chalmers.se (E. Hedlund).
URL: https://www.chalmers.se/en/persons/raelin/ (E. Hedlund).

https://doi.org/10.1016/j.apacoust.2026.111511
Received 7 May 2026; Received in revised form 10 July 2026; Accepted 5 August 2026
Available online 17 August 2026

Speech has been shown to impair short-term memory performance even when it is entirely irrelevant to the task and listeners make deliberate efforts to ignore it [6]. This irrelevant speech effect appears in tasks that require serial recall or maintenance of ordered information, indicating that background speech disrupts cognitive mechanisms responsible for sequencing and retaining information in working memory [5,7,8]. The effect is typically interpreted within theoretical frameworks in which background speech interferes with processes involved in the temporary storage and ordering of information in working memory, including mechanisms associated with temporary storage and serial order processing [9]. Importantly, research also shows that not all auditory input is equally disruptive. The distracting potential of speech depends on properties such as intelligibility, sound level, temporal variability, spectral complexity, and perceptual salience, all of which influence how strongly the signal competes with cognitive processes involved in maintaining ordered representations [10–12].

Speech intelligibility, the degree to which linguistic content of speech can be understood, has been identified as a central determinant of how disruptive speech is to cognitive performance. Research shows that the more intelligible a speech signal is, the greater its potential to interfere with attention and working memory [2,12,13]. However, intelligibility alone does not fully explain the disruptive effect of background speech. Experimental studies have shown that speech-like acoustic structures can impair performance even when linguistic content is not meaningful, and signals such as spectrally rotated or temporally reversed speech can produce disruption comparable to forward speech, indicating that interference can occur even when semantic information is not available [10,12,14].

In addition to intelligibility, properties such as sound level, temporal variation, and spectral complexity have been shown to affect how strongly a signal interferes with cognitive performance [5,10,14]. Although the influence of overall level is not always consistent across studies, increases in level can enhance the perceptual salience of a sound within the auditory scene and thereby increase its potential to capture attention [11,12]. Non-speech sounds, such as steady-state noise or broadband masking signals, typically produce less disruption than speech, highlighting the importance of the dynamic temporal and spectral characteristics of speech for interference with working memory [2,10,13,14]. This shows that the degree of distraction depends not only on the presence of speech but also on how prominent and variable the sound is within the acoustic scene.

Beyond measurable effects on task performance, exposure to background speech has also been associated with increased perceived workload and listening effort [4,12]. However, subjective workload does not always correspond directly to performance outcomes, i.e., listeners can report a higher cognitive load under speech conditions even when performance remains relatively stable [12,15]. This dissociation between subjective effort and performance aligns with capacity-limited models of attention [16] and effortful listening frameworks [17], which propose that people can temporarily compensate for increased task demands by allocating additional cognitive resources [18]. Such compensatory mechanisms may help maintain performance in the short term, but sustained reliance on them can contribute to fatigue and stress and may increase perceived strain during sustained tasks.

These findings show that the disruptive effect of background speech depends on interacting acoustic and cognitive factors, including intelligibility, sound level, speech-like structure, and task demands. Although previous studies have investigated individual acoustic factors such as intelligibility and sound level, these factors have often been examined using different experimental paradigms, cognitive tasks, and evaluation methods. Consequently, a direct comparison of their relative influence on cognitive performance is difficult. A common experimental framework in which individual acoustic characteristics are manipulated systematically could therefore provide a more consistent basis for comparing their relative contributions to speech-related distraction. Another challenge is that acoustic conditions are often described using standardized measures, such as the Speech Transmission Index (STI), which summarize aspects of speech transmission into a single value. While useful within their intended scope, such measures reflect the combined influence of multiple acoustic factors, making it difficult to determine which properties of the signal are responsible for a given rating or how these relate to cognitive disruption.

This motivates controlled experimental studies that examine how specific acoustic conditions influence working memory performance and perceived workload during tasks that require sustained attention.

1.1. Aim

The present study presents results from three separate listening experiments designed to examine how different acoustic characteristics of irrelevant speech affect cognitive performance and perceived workload. More specifically, the study investigates the effects of semantic content, speech-like structure, and sound level under controlled conditions. By systematically manipulating these acoustic characteristics while keeping the experimental framework, cognitive task, and evaluation methods consistent across experiments, the study enables direct comparison of their relative influence on both working memory performance and perceived workload.

During the experiments, participants were exposed to different acoustic stimuli while completing cognitive tasks. Working memory performance was measured using the Operation Span Task (OSPAN) [19], and subjective workload was assessed using the NASA Task Load Index (NASA-TLX) [20].

Listening Experiment A focused on semantic content, intelligibility, and speech-like structure. It was designed to distinguish effects related to linguistic information (intelligible and semantically meaningful speech) from effects caused by the acoustic characteristics of speech, such as temporal modulation patterns or spectral variability. Listening Experiments B1 and B2 focused on sound level in order to examine level-dependent changes in the disruptive effect of speech, where higher levels increase the perceptual salience of the signal within the auditory scene.

2. Listening experiment

All listening experiments were conducted at the Division of Applied Acoustics, Chalmers University of Technology, during June–September 2024, November–December 2024, and February–March 2025 for Experiments A, B1, and B2, respectively. The experimental procedure was identical across all three experiments except that the stimuli were processed differently. Each participant completed the experiment individually in a secluded room to avoid external disturbances. The stimuli were presented through open-back headphones (Sennheiser HD 650) driven by a Focusrite Scarlett Solo audio interface, and all experimental steps were performed digitally using a custom MATLAB application.

During the experimental procedure, participants completed rounds of the Operation Span task (OSPAN) [19], each presented under a different sound condition. The sound conditions were presented in random order for each participant. Following each round, participants rated their perceived workload by completing the NASA Task Load Index (NASA-TLX) questionnaire [20], assessing the subjective demands of the task they had just completed.

2.1. Participants

Listening Experiment A comprised 30 participants aged 20–34 years (M = 25.7, SD = 3.2). One participant reported hearing loss and was excluded from analysis (N = 29, 18 male and 11 female).

Listening Experiment B1 comprised 28 participants aged 21–30 years (M = 25.4, SD = 2.2). Three participants reported hearing loss and were excluded from analysis (N = 25, 12 male, 12 female, and one not specified).

Listening Experiment B2 comprised 30 participants aged 20–32 years (M = 24.6, SD = 3.1). There were no reports of hearing loss (N = 30, 15 male and 15 female).

No formal assessment of English proficiency was performed. Participants were informed prior to recruitment that the experiment would be conducted entirely in English. Consequently, participation assumed a level of English proficiency sufficient to complete the experimental tasks. All participants provided their informed consent prior to participating in the study.

2.2. Stimuli

Impulse responses (IRs) were generated using a hybrid room-acoustic prediction tool [21] in which early reflections were computed with an image-source method using a maximum reflection order of N_{ord} = 16. In the prediction tool, energy is gradually transferred from the image-source model to the acoustic radiosity model which is used to compute the diffuse late part of the impulse response. The contributions from the image-source model and the diffuse radiosity response were merged and distributed into directional components, represented at 5° azimuthal resolution (collapsed to the horizontal plane), producing 72 monaural impulse responses. These directional components were subsequently convolved with the corresponding head-related transfer functions (HRTFs) [22] and combined to obtain binaural room impulse responses (BRIRs), allowing the simulated sound field to be reproduced over headphones while preserving spatial cues.

The simulated room was defined as a shoebox-shaped space with dimensions 32 m × 5.8 m × 3.1 m (XYZ), chosen to represent a large open office. The acoustic properties of the room surfaces were adjusted to obtain a reverberation time of 0.2 s, corresponding to a well-damped office environment. The source and receiver positions were fixed at 5.3 m × 2.9 m × 1.25 m (XYZ) and 7.3 m × 2.9 m × 1.25 m (XYZ), respectively, representing two workstations occupied by seated individuals.

All auditory stimuli were generated by convolving the source signals with the BRIRs, ensuring that all listening conditions differed only in the intended experimental parameters. The stimuli were presented through headphones and the order of the listening conditions was randomized for each participant to avoid order effects.

The playback levels were calibrated by reproducing the stimuli through the experimental headphones mounted on an artificial head [23] in the test environment and adjusting the playback gain to obtain the target A-weighted sound pressure levels. The background noise level was measured using the artificial head and headphones and corresponded to an A-weighted equivalent level of 16 dB.

Listening experiment A

Four acoustical conditions were used in the listening experiment. In addition to silence, denoted C_S, which was used as a baseline anchor, the stimuli included speech, locally time-reversed speech and Brownian noise ($1/|f|^2$), denoted C_{Sp}, C_{RS}, and C_N, respectively. The speech stimuli were derived from recorded utterances of short sentences with low predictability [24] obtained from a speech database [25]. Individual utterances were randomly selected and concatenated into several stimuli of around 10 min each.

The locally time-reversed speech conditions were generated by segmenting the 10-min speech stimuli into random windows of 100–300 ms, reversing each window in time, and concatenating the reversed segments in their original order. This manipulation removes linguistic intelligibility while preserving local amplitude-envelope fluctuations and the rhythmic cadence characteristic of conversational speech.

Brownian noise was included as a non-speech acoustic baseline that provides continuous auditory stimulation without linguistic or speech-like structure. Unlike silence, Brownian noise includes a low-frequency-weighted spectrum characteristic of naturally occurring acoustic backgrounds, allowing the separation of the effects of mere auditory presence from the effects of speech-like processing. All stimuli were convolved with the BRIR to produce binaural renderings for headphone playback.

Speech and locally time-reversed speech were adjusted to have comparable SPLs, yielding an A-weighted equivalent level of 47 dB. The level was selected to represent typical conversational speech levels encountered in open-plan office environments [26], consistent with the speaker–listener geometry used in the acoustic model [21]. This was intended to reflect realistic open-office conditions and support ecological validity for the cognitive tasks.

The noise stimulus was adjusted to achieve a loudness perceptually comparable to the speech stimuli using the ISO 532–1 (Zwicker) loudness model [27], corresponding to an A-weighted equivalent level of 44 dB. The lower A-weighted level, despite comparable loudness, reflects the predominantly low-frequency content of the Brownian noise, as A-weighting applies stronger attenuation at low frequencies than loudness estimates.

Listening experiment B1 & B2

Each of these experiments used four acoustical conditions. In addition to silence, which served as a baseline anchor, the remaining conditions consisted of speech presented at different sound levels. The stimuli were generated using the same speech material and BRIRs as in Listening Experiment A, and were rendered for headphone playback using the same procedure.

In Listening Experiment B1, three acoustical conditions with different presentation levels were used, together with silence C_S. The A-weighted equivalent speech levels in the conditions were low (31 dB), medium (51 dB), and high (61 dB), denoted C_{31}, C_{51}, and C_{61}, respectively. The selected levels were chosen to span a wide range of conversational speech levels that may occur in open-plan office environments [26], from low-level background speech to clearly audible nearby speech.

In Listening Experiment B2, the A-weighted equivalent speech levels were low (31 dB), medium (37 dB) and high (43 dB), denoted C_{31}, C_{37}, and C_{43}, respectively. Compared to Listening Experiment B1, a narrower range at the lower-level end was used, allowing for a more detailed characterization of level-dependent effects under quieter conditions.

2.3. Operation span task

The Operation Span Task [19] is a measure of working memory capacity. Unlike simple serial recall tasks, the OSPAN task is designed to assess the simultaneous processing and storage of information, making it a measure of complex working memory capacity. In the present study, the task was implemented as an automated computer-based version developed in MATLAB. Participants were presented with a sequence of mathematical operations, such as 3 x 2 − 1 = 7, to determine true or false. Each operation remained on the screen for up to 10 s and the program advanced automatically if no response was given within that time. The mathematical operations were interleaved with words to be remembered, each displayed for a fixed duration of 2 s. The sequence length increased across trials, ranging from three to eight operation-word pairs, and the participants were informed that recall would occur only after the full sequence. After the final operation-word pair in each sequence, participants had up to 30 s to recall and enter the words in the order they believed they had appeared. All time limits were communicated to the participants before the test began. Each OSPAN round had a variable duration determined by the response times of the arithmetic problems and recall tasks, with a maximum possible duration of 9 min and 36 s. Individual condition durations ranged from approximately 3.7 to 7.1 min (M = 5.6) in Listening Experiment A, 3.7 to 7.2 min (M = 5.7) in Listening Experiment B1, and 4.0 to 7.6 min (M = 5.8) in Listening Experiment B2.

Fig. 1 provides an overview of how the operation, word, and recall components were presented to participants in the MATLAB-based version of the OSPAN task. Each listening experiment consisted of four rounds of the OSPAN task, where each round contained one trial of each

Fig. 1. Interface illustrations from the OSPAN task, recreated for readability. Left: Mathematical operation presented to the participant. Middle: Word presented immediately after the mathematical operation. Right: Recall screen where participants entered the recalled words.

sequence length (3-8 operation-word pairs), resulting in six trials per round and 24 trials in total. Performance was scored based on the number of words correctly recalled, considering both order-sensitive and order-free accuracy. Several iterations of the same word contributed as one correctly recalled word in an incorrect position, regardless of whether it was placed in the correct position. For each participant, the order of the mathematical operations and the order of the words were randomized independently.

Processing accuracy was monitored to ensure engagement. Although an 85% accuracy threshold in the processing component is commonly used as an exclusion criterion in OSPAN tasks [28], participants in the present study were not excluded solely on the basis of low math accuracy, as the study focused on within-participant differences across conditions. Instead of applying the 85% threshold automatically, participants were excluded only when low equation accuracy co-occurred with additional indicators of disengagement, such as unusually fast response times, excessive timeouts, or disproportionately high word-recall performance relative to equation accuracy. No participants were excluded from any of the listening experiments due to non-engagement.

2.4. NASA task load index

The NASA Task Load Index (NASA-TLX) [20] is a subjective workload assessment tool that evaluates dimensions of task experience. The six sub-scales are mental demand, physical demand, temporal demand, performance, effort, and frustration, which are rated on a 21-point scale originating from a 7-point Likert structure. All scales range from *very low* to *very high*, except for performance, which ranges from *perfect* to *failure*. The NASA-TLX produces an overall workload score by combining the ratings with a weighting procedure. The weighting reflects how strongly each workload factor contributes to the participant's perceived workload. To obtain these weights, participants compare the six dimensions in pairs and choose the one that contributed more to their workload. Consequently, each sub-scale appears in five comparisons, and the number of times it is selected becomes its weight, a value between zero and five. The final workload score is calculated by multiplying each rating by its weight, summing the six weighted values and scaling the result to produce a score between 0, indicating minimal workload, and 100, indicating maximal workload.

In this study, participants were asked to rate the six dimensions after each round of the OSPAN task to reflect the workload associated with the sound condition for that round, where the rating interface required participants to position a slider along each scale. The weighting procedure was performed only after the fourth and final round, consistent with the NASA-TLX methodology, which treats the weights as a stable, participant-specific representation of the relative importance of the workload sources rather than condition-specific values [29]. Fig. 2 illustrates the rating and weighting interface. No time limits were imposed for either the ratings or the weighting procedure. Following each NASA-TLX questionnaire, participants were allowed a short break before proceeding to the next OSPAN round.

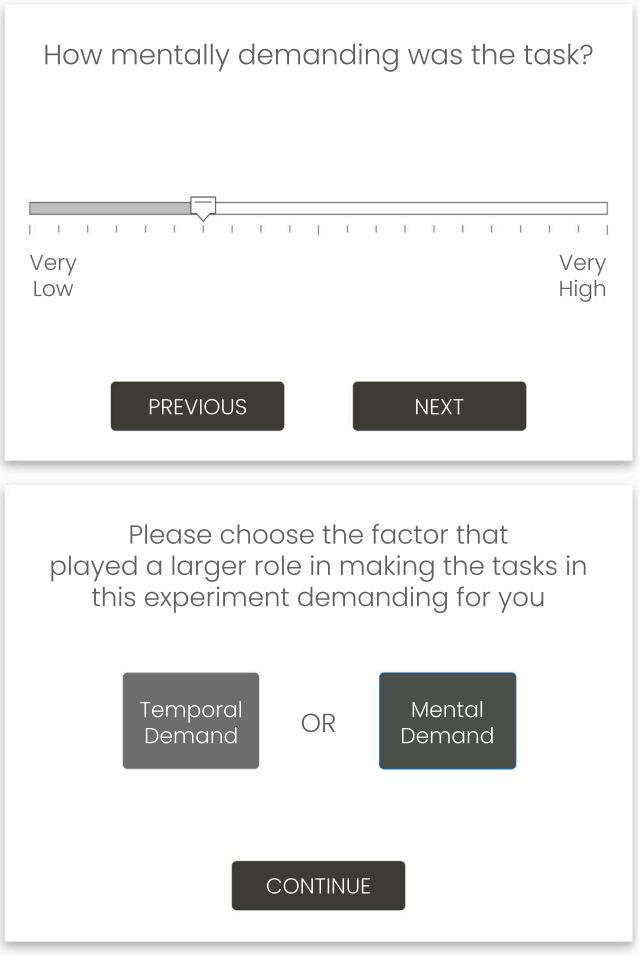

Fig. 2. Interface illustrations from the NASA-TLX questionnaire, recreated for readability. Top: Example of the sub-scale rating. Bottom: Example of the pairwise comparison.

3. Results

Results from the OSPAN task and NASA-TLX questionnaire are presented in this section for each listening experiment, acoustic condition, and participant. The results from the Operation Span Task (OSPAN) are presented with separate results for correct number of words with order sensitive accuracy as well as order-free accuracy.

For each listening experiment, Friedman tests were conducted to assess whether ratings differed across listening conditions. To avoid assumptions about normality, non-parametric tests were used. Statistical significance was evaluated at $\alpha = .05$. Post-hoc pairwise comparisons were performed using Bonferroni-corrected Wilcoxon signed-rank tests (6 comparisons, $\alpha = .0083$).

Figs. 3(a), 4(a) and 5(a) show OSPAN task results, where left-facing violins correspond to order-sensitive accuracy (correct words in correct order) and right-facing violins correspond to order-free accuracy (correct words, regardless of position). Figs. 3(b), 4(b) and 5(b) show NASA-TLX workload ratings. Boxes indicate the interquartile range (IQR), with the median shown as a horizontal line. Medians were used to account for possible skewness. Whiskers extend to the most extreme values within $1.5 \times$IQR, and points beyond this range are plotted as outliers. All statistics are calculated using the full dataset (including outliers).

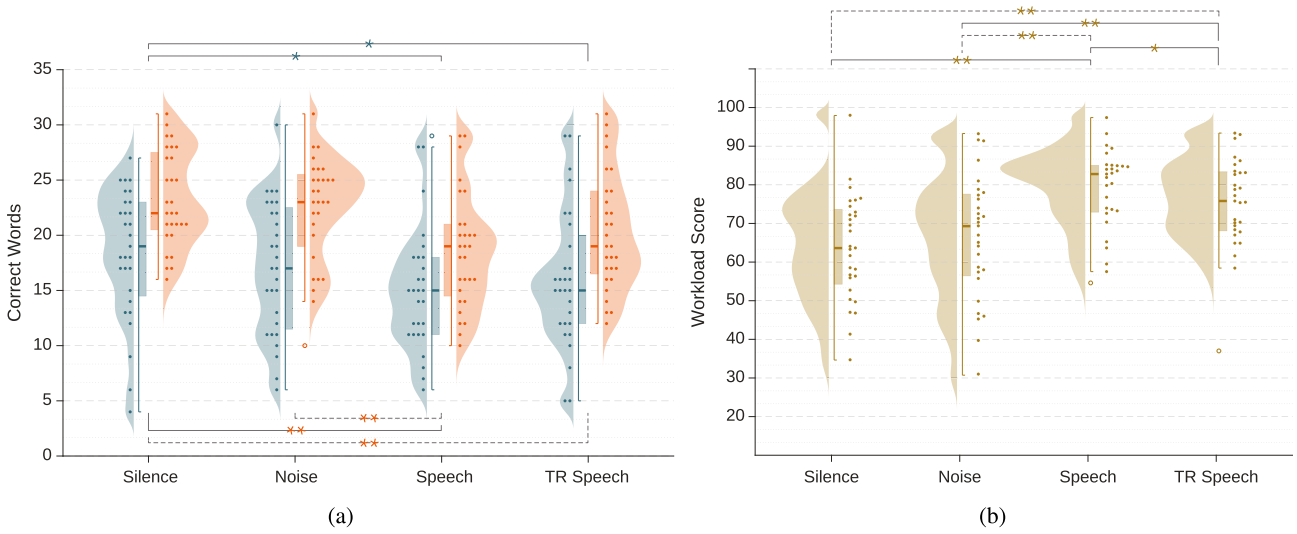

Fig. 3. (a) OSPAN task results for Listening Experiment A. Left-facing violins correspond to order-sensitive accuracy and right-facing violins correspond to order-free accuracy. (b) Subjective workload score from the NASA-TLX questionnaire for Listening Experiment A. Asterisks specify level of significance (* for $p < .0083$ and ** for $p \leq .001$).

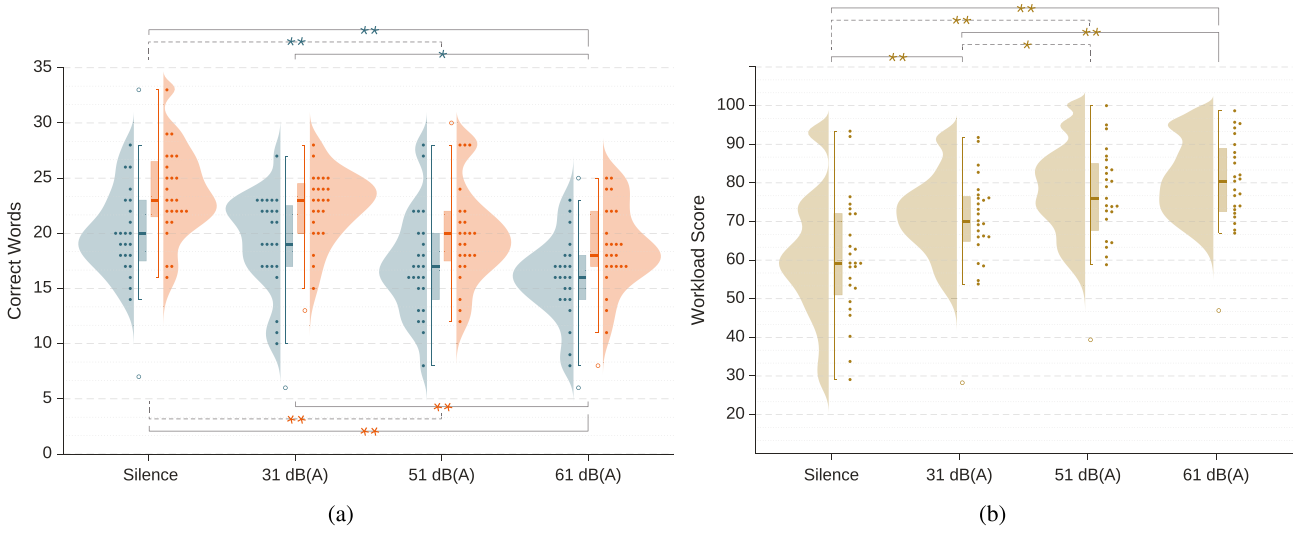

Fig. 4. (a) OSPAN task results for Listening Experiment B1. Left-facing violins correspond to order-sensitive accuracy and right-facing violins correspond to order-free accuracy. (b) Subjective workload score from the NASA-TLX questionnaire for Listening Experiment B1. Asterisks specify level of significance (* for $p < .0083$ and ** for $p \leq .001$).

3.1. Listening experiment A

Operation span task

A significant main effect of condition was observed when recall was scored in correct serial order, $\chi^2(3) = 9.044, p = .029$. When recall accuracy was scored irrespective of serial order, the effect of condition was more pronounced, $\chi^2(3) = 26.967, p < .001$.

Fig. 3(a) shows OSPAN task performance across the four listening conditions. Performance was highest in the silent condition (C_S) under both scoring approaches. The speech (C_{Sp}) and locally time-reversed speech (C_{RS}) conditions yielded comparable recall and lower performance relative to silence. By contrast, the noise condition (C_N) was associated with recall levels closer to those observed in silence, particularly when serial order was not considered.

Post-hoc tests showed that, for order-sensitive accuracy, recall in the silent condition was significantly higher than in C_{Sp} ($p = .003$) and C_{RS} ($p = .008$). When serial order was disregarded, these differences became even more pronounced (both $p < .001$). No significant difference was observed between C_{Sp} and C_{RS} under either scoring method. Likewise, C_N did not significantly differ from C_S. However, when recall was scored irrespective of order, performance in C_N was significantly higher than in C_{Sp} ($p < .001$). Statistical test results are summarized in Table 1, with complete results reported in Appendix A, Table 7.

NASA-TLX

A significant main effect of condition on perceived workload was found, $\chi^2(3) = 43.055, p < 0.001$.

Fig. 3(b) presents NASA-TLX scores for Listening Experiment A. Workload ratings were highest in the speech condition (C_{Sp}), followed by locally time-reversed speech (C_{RS}). In contrast, the noise condition (C_N) and the silent condition (C_S) were associated with lower perceived workload, indicating a clear separation between speech-based conditions and noise or silence.

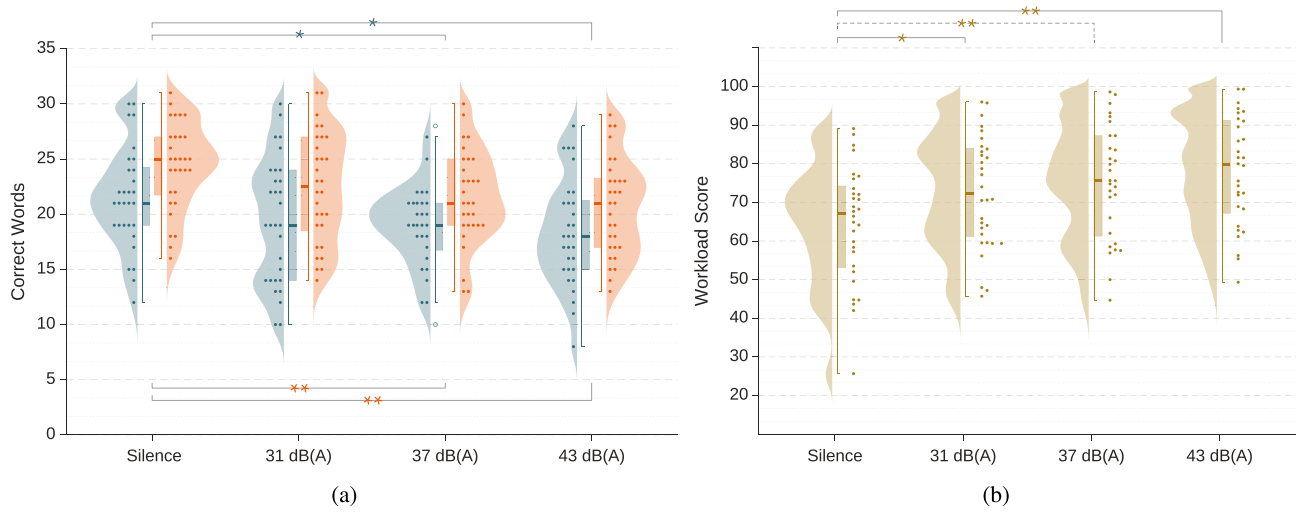

Fig. 5. (a) OSPAN task results for Listening Experiment B2. Left-facing violins correspond to order-sensitive accuracy and right-facing violins correspond to order-free accuracy. (b) Subjective workload score from the NASA-TLX questionnaire for Listening Experiment B2. Asterisks specify level of significance (* for $p < .0083$ and ** for $p \leq .001$).

Post-hoc comparisons showed that C_{Sp} was rated as significantly more demanding than C_{RS} ($p = 0.008$), C_N ($p < 0.001$) and C_S ($p < 0.001$). Similarly, C_{RS} was associated with significantly higher workload ratings than both C_N ($p = 0.001$) and C_S ($p < 0.001$). No significant difference was observed between C_N and C_S. Statistical test results are summarized in Table 2, with complete results provided in Appendix A, Table 10.

3.2. Listening experiment B1

Operation span task

A significant main effect of listening condition was observed when recall was scored in correct serial order, $\chi^2(3) = 24.063, p < .001$. A similarly strong effect was found when recall was scored with order-free accuracy, $\chi^2(3) = 34.021, p < .001$.

Fig. 4(a) illustrates OSPAN performance for the four sound-level conditions. As in Listening Experiment A, recall was highest in the silent condition (C_S), whereas recall declined at higher sound levels, with the medium level condition (C_{51}) and the high level condition (C_{61}) showing progressively lower scores. The low level condition (C_{31}) however, yielded similar recall scores as C_S. When serial order was disregarded, performance in C_S and C_{31} appeared even more similar, while the decrease at higher sound levels remained evident.

Post-hoc comparisons confirmed that recall in C_S was significantly higher than in both C_{51} and C_{61} for order-sensitive and order free scoring (both $p < .001$). In addition, recall in C_{31} exceeded that in C_{61} when serial order was considered ($p = .002$), with an even stronger effect when order was ignored ($p < .001$). Statistical test results are summarized in Table 3, with complete results provided in Appendix A, Table 8.

NASA-TLX

A significant main effect of condition on perceived workload was found, $\chi^2(3) = 45.720, p < .001$.

Fig. 4(b) illustrates NASA-TLX scores for Listening Experiment B1. Perceived workload was lowest in the silent condition (C_S). Introducing speech increased reported workload, and ratings showed a gradual rise with increasing sound level.

Post-hoc tests supported this pattern where workload ratings in C_{31} were significantly higher than in the silent condition ($p < .001$) but significantly lower than in both C_{51} ($p = .002$) and C_{61} ($p < .001$). In turn, both C_{51} and C_{61} were associated with significantly higher workload ratings than silence (both $p < .001$). No significant difference was observed between C_{51} and C_{61}. Statistical test results are summarized in Table 4, with complete results reported in Appendix A, Table 11.

Table 1
Summary of statistical test results for the OSPAN task across conditions in Listening Experiment a (significant comparisons in **bold**).

		Order-sensitive	Order-free
Friedman test		$\chi^2(3) = 9.044$ $p = 0.029$	$\chi^2(3) = 26.967$ $p < 0.001$
Wilcoxon signed-rank tests*	C_{Sp} - C_{RS}	$p = 0.499$	$p = 0.029$
	C_{Sp} - C_N	$p = 0.088$	**$p < 0.001$**
	C_{Sp} - C_S	**$p = 0.003$**	**$p < 0.001$**
	C_{RS} - C_N	$p = 0.202$	$p = 0.073$
	C_{RS} - C_S	**$p = 0.008$**	**$p < 0.001$**
	C_N - C_S	$p = 0.236$	$p = 0.333$

* Bonferroni-corrected significance level: $\alpha = 0.0083$

Table 2
Summary of statistical results for NASA-TLX scores across conditions in Listening Experiment a (significant comparisons in **bold**).

Friedman test	$\chi^2(3) = 43.055$ $p < 0.001$	
Wilcoxon signed-rank tests*	C_{Sp} - C_{RS}	**$p = 0.008$**
	C_{Sp} - C_N	**$p < 0.001$**
	C_{Sp} - C_S	**$p < 0.001$**
	C_{RS} - C_N	**$p = 0.001$**
	C_{RS} - C_S	**$p < 0.001$**
	C_N - C_S	$p = 0.304$

3.3. Listening experiment B2

Operation span task

A significant main effect of condition was observed when recall was scored in correct serial order, $\chi^2(3) = 13.968, p = .003$, and as for Listening Experiments A, a stronger effect was observed when recall accuracy was evaluated irrespective of serial order, $\chi^2(3) = 24.228, p < .001$.

Fig. 5(a) presents OSPAN task performance for the sound-level conditions for Listening Experiment B2. The silent condition (C_S) again yielded the highest recall for both scoring approaches. Performance

Table 3
Summary of statistical test results for the OSPAN task across conditions in Listening Experiment B1 (significant comparisons in **bold**).

		Order-sensitive	Order-free
Friedman test		$\chi^2(3) = 24.063$ $p < 0.001$	$\chi^2(3) = 34.021$ $p < 0.001$
Wilcoxon signed-rank tests*	C_{31} - C_{51}	$p = 0.077$	$p = 0.043$
	C_{31} - C_{61}	**p = 0.002**	**p < 0.001**
	C_{31} - C_S	$p = 0.190$	$p = 0.026$
	C_{51} - C_{61}	$p = 0.081$	$p = 0.038$
	C_{51} - C_S	**p < 0.001**	**p < 0.001**
	C_{61} - C_S	**p < 0.001**	**p < 0.001**

* Bonferroni-corrected significance level: $\alpha = 0.0083$

Table 4
Summary of statistical results for NASA-TLX scores across conditions in Listening Experiment B1 (significant comparisons in **bold**).

Friedman test	$\chi^2(3) = 45.720$ $p < 0.001$	
Wilcoxon signed-rank tests*	C_{31} - C_{51}	**p = 0.002**
	C_{31} - C_{61}	**p < 0.001**
	C_{31} - C_S	**p < 0.001**
	C_{51} - C_{61}	$p = 0.065$
	C_{51} - C_S	**p < 0.001**
	C_{61} - C_S	**p < 0.001**

Table 5
Summary of statistical test results for the OSPAN task across conditions in Listening Experiment B2 (significant comparisons in **bold**).

		Order-sensitive	Order-free
Friedman test		$\chi^2(3) = 13.968$ $p = 0.003$	$\chi^2(3) = 24.228$ $p < 0.001$
Wilcoxon signed-rank tests*	C_{31} - C_{37}	$p = 0.693$	$p = 0.193$
	C_{31} - C_{43}	$p = 0.385$	$p = 0.030$
	C_{31} - C_S	$p = 0.049$	$p = 0.015$
	C_{37} - C_{43}	$p = 0.256$	$p = 0.152$
	C_{37} - C_S	**p = 0.002**	**p < 0.001**
	C_{43} - C_S	**p = 0.002**	**p < 0.001**

* Bonferroni-corrected significance level: $\alpha = 0.0083$

Table 6
Summary of statistical results for NASA-TLX scores across conditions in Listening Experiment B2 (significant comparisons in **bold**).

Friedman test	$\chi^2(3) = 22.280$ $p < 0.001$	
Wilcoxon signed-rank tests*	C_{31} - C_{37}	$p = 0.417$
	C_{31} - C_{43}	$p = 0.047$
	C_{31} - C_S	**p = 0.006**
	C_{37} - C_{43}	$p = 0.019$
	C_{37} - C_S	**p < 0.001**
	C_{43} - C_S	**p < 0.001**

in the acoustical conditions (C_{31}, C_{37} and C_{41}) was lower overall and showed broadly similar levels of recall.

Post-hoc analyses indicated that recall in both C_{37} and C_{43} was significantly lower than in the silent condition when serial order was considered (both $p = .002$). The same pattern was observed when order was disregarded, with both conditions again differing significantly from silence (both $p < .001$). In contrast, recall in C_{31} did not differ significantly from the silent condition. No significant differences were detected among the acoustical conditions themselves. Statistical test results are summarized in Table 5, with complete results reported in Appendix A, Table 9.

NASA-TLX

A significant main effect of condition on perceived workload was found, $\chi^2(3) = 22.280, p < .001$.

Fig. 5(b) illustrates NASA-TLX scores for Listening Experiment B2. Perceived workload was lowest in the silent condition (C_S) and increased once speech was introduced. The three sound level conditions yielded progressively higher ratings overall, although the differences between C_{31} and C_{37} were relatively small.

Post-hoc tests showed that workload ratings in C_{31}, C_{37} and C_{43} were all significantly higher than in the silent condition ($p = .006$, $p < .001$ and $p < .001$, respectively). No significant differences were observed between the acoustical conditions themselves. Statistical test results are summarized in Table 6, with complete results reported in Appendix A, Table 12.

4. Discussion

The results indicate that exposure to speech reduces working memory performance, as reflected by a decrease in correctly recalled words in conditions containing speech. This effect was observed for both order-sensitive and order-free scoring of the OSPAN task, indicating that speech interferes with working memory processes irrespective of whether serial order is taken into account. Across the experiments, the silent condition consistently yielded the highest recall performance, suggesting that the absence of competing auditory information provides the most favorable environment for maintaining the task items.

The results further suggest the presence of a level-related threshold below which speech does not measurably disrupt performance. In Listening Experiment B1, recall in the lowest speech level condition (C_{31}) was comparable to the silent baseline, whereas higher sound levels were associated with progressively lower performance. A similar pattern was observed in Listening Experiment B2, where the higher speech levels differed significantly from the silent condition, while recall in C_{31} did not. Although performance in C_{31} in B2 was somewhat lower than in the silent condition, the relatively large spread of scores in this condition indicates considerable variability between participants in susceptibility to low-level speech distraction. Consequently, the results do not allow a clear distinction between performance under this condition and the silent baseline. Together, these findings suggest a level-dependent pattern, where disruption appears more pronounced at higher levels.

Beyond the lowest speech level, both Listening Experiments B1 and B2 indicate that the magnitude of disruption increases with sound level. In both experiments, the middle and higher speech levels were associated with progressively lower recall. This pattern suggests that the disruptive influence of speech is not only determined by its presence but also by its relative level, with stronger speech signals producing greater interference with the cognitive processes required for the task.

The comparison between stimulus types in Listening Experiment A indicates that the disruptive effect cannot be attributed simply to the presence of any auditory stimulus. Performance was comparable between the silent condition and the Brownian noise condition, while the speech and locally time-reversed speech conditions produced similar results. In other words, an acoustic background alone did not produce measurable disruption. Instead, the effect appears to be associated with speech-like signals.

Previous studies have reported stronger distraction with higher intelligibility. However, this does not necessarily imply that reducing intelligibility eliminates the disruptive effect of speech. In the present study, intelligible speech and locally time-reversed speech produced comparable working memory performance, suggesting that reducing intelligibility alone was insufficient to substantially reduce disruption. Especially considering that the Brownian noise condition produced performance comparable to silence, the results suggest that the disruption in this study was more strongly linked to the continuously changing temporal and spectral characteristics of speech than to the mere presence of sound or linguistic meaning. Therefore, the present findings should not be interpreted as evidence that intelligibility is irrelevant. Rather, they

suggest that, under the tested conditions, the preserved temporal and spectral characteristics of speech contributed more strongly to working memory disruption than semantic content.

The stronger omnibus effects observed when recall was scored irrespective of serial order indicate that removing the order requirement made the differences between conditions more pronounced. This does not necessarily imply that speech affects a specific component of recall. Instead, the pattern suggests that speech-related distraction may influence multiple aspects of task performance, including both the retention of items and the processes involved in organizing them.

A different perspective emerges when considering the subjective workload ratings. Even in conditions where recall performance remained comparable to the silent baseline, participants reported increased workload when exposed to speech. This suggests that participants may have exerted additional effort to maintain their performance in the presence of distracting speech. In other words, performance levels remained stable but maintaining that level appeared to require greater mental effort. This pattern indicates that increases in perceived effort can occur even in the absence of measurable performance decrements.

Such compensation effects are particularly relevant when considering real-world work environments. In settings such as open-plan offices, exposure to speech often occurs continuously over extended periods. While short-term cognitive tasks may still be performed successfully, sustained increases in listening effort could contribute to fatigue, reduced well-being, or decreased productivity over time. The present results therefore highlight the importance of considering both objective performance measures and subjective workload when evaluating the impact of acoustic environments on cognitive functioning.

4.1. Limitations

Investigating working memory performance in controlled listening experiments inevitably involves a trade-off between experimental control and ecological validity. Artificially generated acoustic environments allow for precise manipulation of acoustic parameters that would be difficult to isolate in real-world settings. However, this level of control also introduces limitations when interpreting how the results generalize to everyday situations. In practice, speech exposure typically occurs over extended periods and in combination with other factors that could influence cognitive performance, such as varying work tasks, interactions with colleagues, and concurrent auditory and visual distractions. Furthermore, the present study focused on a single cognitive paradigm designed to assess working memory. Although the OSPAN task captures certain cognitive processes, workplace tasks involve a broader range of cognitive functions. Consequently, the present findings should be interpreted as reflecting working memory performance under controlled experimental conditions rather than the full complexity of real workplace environments.

The participant sample was relatively homogeneous with respect to age, which might limit the extent to which the findings generalize to more diverse workplace population, where cognitive abilities and responses to background speech may differ across individuals. Furthermore, no a priori power analysis was performed to determine a target sample size. Future studies involving larger participant samples would improve the statistical robustness of the findings and allow for a more reliable evaluation of subtle differences between acoustic conditions.

The relatively large spread of scores observed in some conditions indicates inter-individual variability in susceptibility to speech-related distraction. The present study was designed to compare acoustic conditions at the group level rather than identify the participant-related factors underlying this variability. Such factors could include differences in noise sensitivity, cognitive capacity, or language proficiency. Because English proficiency was not objectively assessed, its potential contribution to the observed variability cannot be determined. Participants with lower English proficiency may have extracted less semantic information from intelligible speech, potentially attenuating effects specifically related to semantic processing.

Some methodological considerations should also be acknowledged when interpreting the subjective workload measures, specifically the weighting procedure included in the NASA-TLX where participants compare workload dimensions in pairs. Previous studies have noted that such pairwise comparisons can be conceptually complex and may lead to inconsistencies in the resulting importance rankings, particularly when participants perceive several workload dimensions as similarly relevant [30,31]. However, the weighting procedure can still be useful in cognitive tasks because it allows participants to reduce the influence of dimensions that may be less relevant.

In the present study, room acoustics were simulated using an image source model combined with a radiosity-based model for late reverberation. Although the simulation allowed reflections up to a high order, the relatively short reverberation time meant that most acoustic energy was contained in the direct sound and early reflections. Under these conditions, the amount of energy available for redistribution by the radiosity model becomes limited. This effect may also have been amplified by the relatively large dimensions of the simulated space, where the short reverberation time further reduces the contribution of late reflections. Since the reverberation time was selected independently of the dimensions of the reference office, the simulated environment should be regarded as a controlled acoustic condition rather than a physically representative office environment.

This illustrates a broader limitation of parameter-based acoustic simulation. Under certain acoustic conditions, model parameters such as reflection order, room dimensions, or reverberation time do not always translate directly into acoustically representative environments. The resulting sound field may therefore influence how participants perceive and process the auditory scene, which should be considered when interpreting the applicability of the findings to real office environments.

These considerations, however, do not invalidate the experimental paradigm but clarify the scope within which the results should be interpreted. The findings reflect cognitive performance under controlled, model-defined acoustic conditions, which provide valuable insight into the mechanisms by which speech influences working memory. At the same time, translating these findings to real-world environments requires careful consideration of how acoustic modeling assumptions shape the perceptual characteristics of the stimuli.

Future work should investigate more diverse participant groups, longer exposure durations, varying task demands, and larger sample sizes in order to further clarify how individual variability and acoustic characteristics jointly influence speech-related distraction.

5. Conclusion

The present study examined how different acoustic characteristics of background speech influence working memory performance and perceived workload across three controlled listening experiments. The results show that speech disrupts cognitive performance relative to silence, with stronger effects at higher sound levels, while lower-level speech produces more variable and less consistent interference. At the same time, subjective workload increased systematically with speech exposure, even in conditions where objective performance remained relatively stable.

The results are consistent with previous research that demonstrates the disruptive effects of irrelevant speech while suggesting that reducing intelligibility alone might not be sufficient to eliminate cognitive interference. Instead, the preserved temporal and spectral characteristics of speech appear to contribute more strongly to working memory disruption than semantic content under the tested conditions. In addition, the observed dissociation between performance and perceived workload aligns with theoretical accounts suggesting that individuals can compensate for increased task demands through increased cognitive effort.

Beyond confirming established effects, the present study refines the understanding of how sound level modulates speech-related distraction. The results indicate that the impact of speech is level-dependent, with clearer performance decrements emerging at higher levels, while lower-level speech does not consistently impair performance despite increased perceived effort. This pattern suggests the presence of a range of sound levels in which compensatory mechanisms may offset disruption, thereby masking measurable performance effects, although substantial inter-individual variability remains.

The present study complements previous research by examining multiple acoustic characteristics within a common experimental framework, allowing their effects on both cognitive performance and perceived workload to be compared directly. Rather than considering semantic content, speech-like structure, and sound level in isolation, the combined findings indicate that these characteristics influence speech-related distraction to different extents under comparable experimental conditions. In particular, the results suggest that sound level and the preserved temporal and spectral characteristics of speech contribute more strongly to working memory disruption than semantic content alone. This has implications for the design of workspaces, suggesting that reducing the level and prominence of background speech may be critical not only for maintaining performance but also for limiting cognitive strain during sustained tasks.

CRediT authorship contribution statement

Elin Hedlund: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Laura Lindel:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wolfgang Kropp:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

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

During the preparation of this work, generative artificial intelligence tools were used to refine the language. The authors reviewed and edited the content as needed and take full responsibility for the content of the published article. No artificial intelligence tools were used to create, alter, or manipulate figures or images, and all scientific content, interpretations, data, figures, and conclusions were developed by the authors.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Wolfgang Kropp reports that financial support was provided by AFA Insurance. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The project was funded by Afa Försäkring, an organization owned by Sweden's labor market parties, project number 190278. The listening experiments have been approved by the Swedish Ethics Review Authority (Etikprövningsmyndigheten) with number 2023-03339-01.

Appendix A. Statistical details

Tables 7–9 contain statistical details for correct word scores per condition for Listening experiments A, B1 and B2, respectively. Results include order-sensitive and order-free accuracy as well as Friedman tests and post-hoc Wilcoxon signed-rank comparisons (Bonferroni-corrected $\alpha = 0.0083$) with corresponding effect sizes.

Tables 10–12 contain statistical details for subjective workload scores in the NASA-TLX questionnaire in Listening Experiments A, B1, and B2, respectively, including Friedman test and post-hoc Wilcoxon signed-rank comparisons (Bonferroni-corrected $\alpha = 0.0083$) with corresponding effect sizes.

Table 7
Statistical details for OSPAN task for Listening Experiment A. Significant comparisons after correction are in **bold**.

	Order-sensitive		Order-free	
Friedman test	$\chi^2(3) = 9.044$, $p = 0.029$		$\chi^2(3) = 26.967$, $p < 0.001$	
Stimuli	Med	*(IQR)*	Med	*(IQR)*
C_S	19	*(14.50–23.00)*	22	*(20.50-27.50)*
C_N	17	*(11.50–22.50)*	23	*(19.00-25.50)*
C_{RS}	15	*(12.00–20.00)*	19	*(16.50-24.00)*
C_{Sp}	15	*(11.00–18.00)*	19	*(14.50-21.00)*
Wilcoxon signed-rank tests ($\alpha = 0.0083$)				
C_{Sp} - C_{RS}	$Z = -0.676$, $p = 0.499$, $r = 0.126$		$Z = -2.189$, $p = 0.029$, $r = 0.407$	
C_{Sp} - C_N	$Z = -1.704$, $p = 0.088$ $r = 0.316$		**$Z = -3.806$, $p < 0.001$** $r = 0.707$	
C_{Sp} - C_S	**$Z = -2.991$, $p = 0.003$** $r = 0.555$		**$Z = -4.309$, $p < 0.001$** $r = 0.800$	
C_{RS} - C_N	$Z = -1.277$, $p = 0.202$ $r = 0.237$		$Z = -1.792$, $p = 0.073$ $r = 0.333$	
C_{RS} - C_S	**$Z = -2.638$, $p = 0.008$** $r = 0.490$		**$Z = -3.496$, $p < 0.001$** $r = 0.649$	
C_N - C_S	$Z = -1.184$, $p = 0.236$ $r = 0.220$		$Z = -0.968$, $p = 0.333$ $r = 0.180$	

Table 8
Statistical details for OSPAN task for Listening Experiment B1. Significant comparisons after correction are in **bold**.

	Order-sensitive		Order-free	
Friedman test	$\chi^2(3) = 24.063$, $p < 0.001$		$\chi^2(3) = 34.021$, $p < 0.001$	
Stimuli	Med	*(IQR)*	Med	*(IQR)*
C_S	20	(17.50–23.00)	23	(21.50–26.50)
C_{31}	19	(17.00–22.50)	23	(20.00–24.50)
C_{51}	17	(14.00–20.00)	20	(17.50–22.00)
C_{61}	16	(14.00–18.00)	18	(17.00–22.00)
Wilcoxon signed-rank tests ($\alpha = 0.0083$)				
C_{31} - C_{51}	$Z = -1.770$, $p = 0.077$ $r = 0.335$		$Z = -2.020$, $p = 0.043$ $r = 0.382$	
C_{31} - C_{61}	**$Z = -3.170$, $p = 0.002$** $r = 0.600$		**$Z = -3.674$, $p < 0.001$** $r = 0.694$	
C_{31} - C_S	$Z = -1.311$, $p = 0.190$ $r = 0.248$		$Z = -2.228$, $p = 0.026$ $r = 0.421$	
C_{51} - C_{61}	$Z = -1.745$, $p = 0.081$ $r = 0.330$		$Z = -2.079$, $p = 0.038$ $r = 0.393$	
C_{51} - C_S	**$Z = -3.497$, $p < 0.001$** $r = 0.661$		**$Z = -3.491$, $p < 0.001$** $r = 0.660$	
C_{61} - C_S	**$Z = -4.036$, $p < 0.001$** $r = 0.763$		**$Z = -4.271$, $p < 0.001$** $r = 0.807$	

Table 9
Statistical details for OSPAN task for Listening Experiment B2. Significant comparisons after correction are in **bold**.

	Order-sensitive		Order-free	
Friedman test	$\chi^2(3) = 13.968$ $p = 0.003$		$\chi^2(3) = 24.228$ $p < 0.001$	
Stimuli	Med	*(IQR)*	Med	*(IQR)*
C_S	21	*(19.00-24.25)*	25	*(21.75-27.00)*
C_{31}	19	*(14.00-24.00)*	22.5	*(18.50-27.00)*
C_{37}	19	*(16.75-21.00)*	21	*(19.00-25.00)*
C_{43}	18	*(15.00-21.25)*	21	*(17.00-23.25)*
Wilcoxon signed-rank tests ($\alpha = 0.0083$)				
C_{31} - C_{37}	$Z = -0.395, p = 0.693$ $r = 0.072$		$(Z = -1.301, p = 0.193)$ $r = 0.238$	
C_{31} - C_{43}	$Z = -0.868, p = 0.385$ $r = 0.159$		$(Z = -2.173, p = 0.030)$ $r = 0.367$	
C_{31} - C_S	$Z = -1.973, p = 0.049$ $r = 0.360$		$(Z = -2.433, p = 0.015)$ $r = 0.444$	
C_{37} - C_{43}	$Z = -1.136, p = 0.256$ $r = 0.207$		$(Z = -1.434, p = 0.152)$ $r = 0.262$	
C_{37} - C_S	$\mathbf{Z = -3.109, p = 0.002}$ $r = 0.568$		$\mathbf{(Z = -3.806, p < 0.001)}$ $r = 0.695$	
C_{43} - C_S	$\mathbf{Z = -3.148, p = 0.002}$ $r = 0.575$		$\mathbf{(Z = -3.910, p < 0.001)}$ $r = 0.714$	

Table 10
Statistical details for NASA-TLX ratings for Listening Experiment A. Significant comparisons after correction are in **bold**.

Friedman test	$\chi^2(3) = 42.269, p < 0.001$	
Stimuli	Median	*(IQR)*
C_S	63.62	*(54.30-73.56)*
C_N	69.30	*(56.69-77.04)*
C_{RS}	75.81	*(68.10-83.26)*
C_{Sp}	82.79	*(72.95-85.01)*
Wilcoxon signed-rank tests ($\alpha = 0.0083$)		
C_{Sp} - C_{RS}	$\mathbf{Z = -2.649,\ p = 0.008,\ r = 0.4919}$	
C_{Sp} - C_N	$\mathbf{Z = -4.444,\ p < 0.001,\ r = 0.8252}$	
C_{Sp} - C_S	$\mathbf{Z = -4.638,\ p < 0.001,\ r = 0.8613}$	
C_{RS} - C_N	$\mathbf{Z = -3.211,\ p = 0.001,\ r = 0.5963}$	
C_{RS} - C_S	$\mathbf{Z = -4.227,\ p < 0.001,\ r = 0.7849}$	
C_N - C_S	$Z = -1.027,\ p = 0.304,\ r = 0.1907$	

Table 11
Statistical details for NASA-TLX ratings for Listening Experiment B1. Significant comparisons after correction are in **bold**.

Friedman	$\chi^2(3) = 45.720, p < 0.001$	
Stimuli	Median	*(IQR)*
C_S	59.20	*(51.01-72.04)*
C_{31}	70.11	*(63.97-76.42)*
C_{51}	76.07	*(67.70-85.22)*
C_{61}	80.26	*(71.59-88.89)*
Wilcoxon signed-rank tests ($\alpha = 0.0083$)		
C_{31} - C_{51}	$\mathbf{Z = -3.027,\ p = 0.002,\ r = 0.5720}$	
C_{31} - C_{61}	$\mathbf{Z = -4.157,\ p < 0.001,\ r = 0.7856}$	
C_{31} - C_S	$\mathbf{Z = -3.538,\ p < 0.001,\ r = 0.6686}$	
C_{51} - C_{61}	$Z = -1.843,\ p = 0.065,\ r = 0.3483$	
C_{51} - C_S	$\mathbf{Z = -4.292,\ p < 0.001,\ r = 0.8111}$	
C_{61} - C_S	$\mathbf{Z = -4.372,\ p < 0.001,\ r = 0.8262}$	

Table 12
Statistical details for NASA-TLX ratings for Listening Experiment B2. Significant comparisons after correction are in **bold**.

Friedman test	$\chi^2(3) = 22.280, p < 0.001$	
Stimuli	Median	*(IQR)*
C_S	67.04	*(53.16–74.22)*
C_{31}	72.83	*(61.21–84.01)*
C_{37}	75.63	*(61.31–87.26)*
C_{43}	79.77	*(64.27–91.21)*
Wilcoxon signed-rank tests ($\alpha = 0.0083$)		
C_{31} - C_{37}	$Z = -0.812,\ p = 0.417,\ r = 0.1483$	
C_{31} - C_{43}	$Z = -1.985,\ p = 0.047,\ r = 0.3624$	
C_{31} - C_S	$\mathbf{Z = -2.725,\ p = 0.006,\ r = 0.4975}$	
C_{37} - C_{43}	$Z = -2.355,\ p = 0.019,\ r = 0.4300$	
C_{37} - C_S	$\mathbf{Z = -3.857,\ p < 0.001,\ r = 0.7042}$	
C_{43} - C_S	$\mathbf{Z = -4.124,\ p < 0.001,\ r = 0.7529}$	

Data availability

Data will be made available on request.

References

[1] Banbury SP, Berry DC. Office noise and employee concentration: identifying causes of disruption and potential improvements. Ergonomics 2005;48:25–37. https://doi.org/10.1080/00140130412331311390

[2] Hongisto V. A model predicting the effect of speech of varying intelligibility on work performance. Indoor Air 2005;15:458–68. https://doi.org/10.1111/j.1600-0668.2005.00391.x

[3] Haapakangas A, Hongisto V, Eerola M, Kuusisto T. Distraction distance and perceived disturbance by noise - an analysis of 21 open-plan offices. J Acoust Soc Am 2017;141:127–36. https://doi.org/10.1121/1.4973690

[4] Jahncke H, Hygge S, Halin N, Green AM, Dimberg K. Open-plan office noise: cogntive performance and restoration. J Environ Psychol 2011;31:373–82. https://doi.org/10.1016/j.jenvp.2011.07.002

[5] Beaman CP. Auditory distraction from low-intensity noise: a review of the consequences for learning and workplace environments. Appl Cogn Psychol 2005;19:1041–64. https://doi.org/10.1002/acp.1134

[6] Colle HA, Welsh A. Acoustic masking in primary memory. J Verbal Learn Verbal Behav 1976;15:17–31. https://doi.org/10.1016/S0022-5371(76)90003-7

[7] Colle HA. Auditory encoding in visual short-term recall: effects of noise intensity and spatial location. J Verbal Learn Verbal Behav 1980;19:722–35. https://doi.org/10.1016/S0022-5371(80)90403-X

[8] Jones DM, Macken WJ. Irrelevant speech and serial recall. Q J Exp Psychol 1992;44:673–88. https://doi.org/10.1111/j.1467-9450.1992.tb00911.x

[9] Baddeley A. The episodic buffer: a new component of working memory? Trends Cogn Sci 2000;4:417–23. https://doi.org/10.1016/s1364-6613(00)01538-2

[10] Jones D, Madden C, Miles C. Privileged access by irrelevant speech to short-term memory: the role of changing state. Q J Exp Psychol 1992;44A:645–69. https://doi.org/10.1080/14640749208401304

[11] Ellermeier W, Hellbrück J. Is level irrelevant in "irrelevant speech"? Effects of loudness, signal-to-noise ratio, and binaural unmasking. J Exp Psychol Hum Percept Perform 1998;24:1406–14. https://doi.org/10.1037//0096-1523.24.5.1406

[12] Schlittmeier SJ, Hellbrück J, Thaden R, Vorländer M. The impact of background speech varying in intelligibility: effects on cognitive performance and perceived disturbance. Ergonomics 2008;51:719–36. https://doi.org/10.1080/00140130701745925

[13] Venetjoki N, Kaarlela-Tuomaala A, Keskinen E, Hongisto V. The effect of speech and speech intelligibility on task performance. Ergonomics 2006;49:1068–91. https://doi.org/10.1080/00140130600679142

[14] Jones DM, Alford D, Macken WJ, Banbury SP, Tremblay S. Interference from degraded auditory stimuli: linear effects of changing-state in the irrelevant sequence. J Acoust Soc Am 2000;108:1082–8. https://doi.org/10.1121/1.1288412

[15] Muckler FA, Seven SA. Selecting performance measures: "objective" versus "subjective" measurement. Hum Factors 1992;34(4):441–55. https://doi.org/10.1177/001872089203400406

[16] Kahneman D. Attention and effort. Prentice-Hall; 1973.

[17] Pichora-Fuller MK, Kramer SE, Eckert MA, Edwards B, Hornsby BWY, et al. LEH. Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). Ear Hear 2016;37: 5S–27S, https://doi.org/10.1097/AUD.0000000000000312

[18] Kjellberg A. Noise. In: Waldron HA, Edling C, editors. Occupational health practice, 4th ed. Butterworth Heinenmann; 1997. p. 241–56.

[19] Turner ML, Engle RW. Is working memory capacity task dependent? J Mem Lang 1989;28:127–54. https://doi.org/10.1016/0749-596X(89)90040-5

[20] Hart SG, Staveland LE. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Human mental workload, volume 52 of advances in psychology. North-Holland; 1988. p. 139–83. https://doi.org/10.1016/S0166-4115(08)62386-9
[21] Forssén J, Müller L, Hedlund E, Kropp W. Hybrid prediction tool implemented for acoustic design studies of open plan office spaces. Appl Acoust 2026;248. https://doi.org/10.1016/j.apacoust.2026.111284
[22] Braren HS, Fels J. A high-resolution head-related transfer function data set and 3D-scan of KEMAR. Technical report Institute and Chair for Hearing Technology and Acoustics, RWTH Aachen University; 2020.
[23] HMS V 1502 digital artificial head measurement system with high dynamic range (dual ADC). Head Acoustics GmbH. 2026. https://cdn.head-acoustics.com/fileadmin/data/global/Datasheets/Artificial_Heads_Binaural_Recording/HEAD-acoustics-Data-Sheet-HMS-V-1502-Digital-Artificial-Head-Measurement-System-with-High-Dynamic-Range_Dual-ADC-EN.pdf.
[24] IEEE Subcommittee on Subjective Measurements. IEEE recommended practice for speech quality measurements. IEEE Trans Audio Electroacoust 1969;AU-17(3):225–46. https://doi.org/10.1109/TAU.1969.1162058. IEEE Standards Publication No. 297-1969.
[25] Kabal P. TSP speech database. Technical Report Version 2 Department of Electrical & Computer Engineering, McGill University; 2018.
[26] Warnock ACC, Chu WT. Voice and background noise levels measured in open offices. Technical Report IRC-IR-837 Institute for Research in Construction. National Research Council Canada.; 2002.
[27] for Standardization IO. Acoustics - methods for calculating loudness - Part 1: Zwicker method. Technical Report ISO 532-1:2017(E), Geneva, Switzerland 2017.
[28] Unsworth N, Heitz RP, Schrock JC, Engle RW. An automated version of the operation span task. Behav Res Methods 2005;37(3):498–505. https://doi.org/10.3758/BF03192720
[29] Hart SG. NASA task load index (TLX): paper-and-pencil package - volume 1.0. NASA Technical Report. CA United States: NASA Ames Research Center Moffett Field; 1986. https://ntrs.nasa.gov/citations/20000021488.
[30] Virtanen K, Mansikka H, Kontio H, Harris D. Weight watchers: NASA-TLX weights revisited. Theor Issues Ergon Sci 2022;23:725–48. https://doi.org/10.1080/1463922X.2021.2000667
[31] Hart SG. NASA-task load index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 2006;50(9):904–8. https://doi.org/10.1177/154193120605000909