Journal of Memory and Language 131 (2023) 104427

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Conceptualising acoustic and cognitive contributions to divided-attention listening within a data-limit versus resource-limit framework

Sarah Knight *, Lyndon Rakusen, Sven Mattys *

Department of Psychology, University of York, UK

ARTICLE INFO

Keywords:
Speech-in-noise
Masking
Divided attention
Working memory
Individual differences
Cognitive listening

ABSTRACT

An understanding of how listeners divide their attention between two simultaneous talkers requires modelling the interaction between acoustic factors (energetic masking) and cognitive processes (control of auditory attention). The impact of spatial separation between the two talkers on this interaction is unclear, since separation is likely to create both acoustic benefits (release from energetic masking) and cognitive costs (increased demands on spatial attentional control). To explore this question, we manipulated the degree of energetic masking (high vs. low) and spatial separation (collocated to dichotic) between two simultaneous talkers. When energetic masking was high (Experiment 1, unmanipulated talker voices), transcription performance improved monotonically from collocated to dichotic, owing to a gradual release from energetic masking. When energetic masking was low (Experiment 2, bandpass-filtered talker voices), the benefit of spatial separation disappeared; performance even worsened in the dichotic condition. Additionally, across both experiments, individual differences in working memory best predicted transcription performance in conditions where energetic masking was low. These results suggest that energetic masking is the dominant challenge during divided-attention listening, but that the contribution of cognitive control and working memory can be observed when energetic masking is reduced, at least in the context of the current paradigm. The findings are discussed in light of Norman and Bobrow's (1975) concept of data-limited vs. resource-limited tasks, which we propose is a promising framework for reinterpreting existing results from speech-in-noise perception research.

Introduction

Speech perception in multi-talker environments is often challenging. Difficulties may be due to spectrotemporal overlap between a target voice and competing (masker) voices, resulting in direct competition at the cochlear level – a phenomenon referred to as *energetic masking* (Brungart, 2001). However, even when energetic masking is minimal, difficulties can arise from an inability to successfully parse the auditory scene into separate streams (segregation), particularly when the target and maskers share phonological, prosodic, or semantic properties. Additionally, even after successful segregation, listeners may struggle to allocate attention to the target stream and inhibit the masker (Shinn-Cunningham, 2008). Such failures in segregation and attention allocation are often referred to as *informational masking* and may manifest as, for example, misallocations of portions of masker speech to the target (Cooke et al., 2008).

Performance can be improved by separating targets and maskers spatially (Culling & Stone, 2017; Kidd & Colburn, 2017). For example, Arbogast et al. (2002) found an improvement of up to 7 dB in the target-to-masker ratio required for accurate target recognition when target and masker were separated by 90° on the azimuth plane compared to when they were collocated. This benefit, which is referred to as *spatial release from masking* (Litovsky, 2012), is primarily due to a reduction in energetic masking: as a masker is moved away from a target, there is an increase in the number and duration of spectrotemporal regions in which the energy in the target exceeds that in the masker for a given ear, which in turn increases target intelligibility (Edmonds & Culling, 2006). Spatial release from masking can lead to significant improvements in performance. For instance, Freyman et al. (1999) reported an advantage of at least 12 dB SNR when target and masker were separated by 60° as opposed to collocated.

The benefit of spatial separation is clear for situations in which listeners are required to track only one auditory stream (i.e., selective attention). The situation is less straightforward when two streams must be tracked simultaneously (i.e., divided attention). Understanding the role of spatial separation during divided-attention listening raises

* Corresponding authors at: Department of Psychology, University of York, York YO10 5DD, UK.
E-mail addresses: sarah.knight3@york.ac.uk (S. Knight), sven.mattys@york.ac.uk (S. Mattys).

https://doi.org/10.1016/j.jml.2023.104427
Received 11 August 2022; Received in revised form 19 April 2023; Accepted 21 April 2023
Available online 21 May 2023

theoretical questions about the interaction between energetic masking and the cognitive mechanisms involved in auditory attention control. Of particular interest is whether the benefit of spatial separation observed during selective-attention listening extends to divided-attention listening. Although spatial separation between two simultaneous streams should lead to reduced mutual energetic masking, existing studies point to possible costs associated with tracking multiple auditory stimuli when they are perceived to be in different spatial locations. Such costs have been observed when participants are required to track stimuli alternating between the ears (left–right-left–right, etc.). For instance, Treisman (1971) showed that alternating dichotic presentation of digits led to poorer recall than binaural presentation, implying a cost of shifting attention between the two locations. Similarly, Axelrod et al. (1968) found that, when presented with either alternating dichotic clicks or successive monaural clicks, participants needed a faster dichotic rate to match perceived rates across the two conditions. This implies that switching auditory attention between sound sources at different spatial locations causes listeners to "lose" a certain amount of information. Furthermore, Axelrod and Powazek (1972) showed that the perceived rate of clicks coming from alternating spatial locations varied monotonically with angular separation, which indicates that the cost of attention shifts may increase with spatial separation (although see Mondor & Zatorre, 1995). Taken together, these studies suggest that spatial separation between two talkers during divided-attention listening may be detrimental to performance compared to when the talkers are collocated, and this cost may increase as spatial separation increases.

How, then, might listeners cope with the potential costs of divided-attention listening? One possibility is that they recruit additional cognitive resources, such as working memory (WM), to facilitate attentional switches between spatial locations. Research has emphasised the importance of cognition for speech perception in adverse conditions (e.g., Arlinger et al., 2009; Mattys et al., 2012; Rönnberg et al., 2021), and WM is thought to be particularly vital because of its role in the maintenance of degraded speech fragments for delayed integration (Gatehouse et al., 2003). The fact that attentional control is often seen as a key component of WM (e.g., Baddeley, 2000; Engle & Kane, 2004; Unsworth & Spillers, 2010) suggests close connections between WM and wider processes of selective attention and attention switching. Indeed, good WM is associated with better inhibition of irrelevant sound during selective attention (Conway et al., 2001) and with increased flexibility in attention switching during divided attention (Colflesh & Conway, 2007). If listeners recruit WM to counteract the costs of spatial separation between simultaneous talkers, then the relationship between an individual's WM ability and their performance on a divided-attention listening task should be stronger when that task involves spatially-separated than collocated streams.

Lin and Carlile (2015) provided some supporting evidence for this claim. In their study, listeners were asked to track target sentences presented in a background of other talkers, and the target talker's location either stayed the same within a trial or changed. They found that switch costs (i.e., the difference in performance between fixed- and changed-location trials) were correlated with scores on a WM task, with poorer WM associated with larger switch costs. This suggests that WM is recruited more heavily when the to-be-tracked streams are spatially separated.

Data-limited vs. resource-limited processes

To investigate the benefits and costs of spatial separation for divided-attention listening, we considered the distinction made by Norman and Bobrow (1975) between data-limited and resource-limited processes. According to this framework, a task is characterised by the extent to which performance on that task depends on the allocation of processing resources. A task is said to be resource-limited if performance is determined primarily by the amount of processing resources allocated to that task (e.g., a digit span task). However, if performance is largely independent of the resources allocated, and instead primarily determined by the quality of the data available, then the task is said to be data-limited (e.g., an audiometric test). Crucially, in the case of data-limited tasks, the allocation of additional processing resources is unlikely to lead to further improvements. For instance, recruiting additional resources during an audiometric test will not improve performance if the tones are below the listener's detection thresholds.

The application of Norman and Bobrow's framework to selective- and divided-attention listening requires, first, a definition of the concept of "data limit" in this context. The quality of a speech signal (the data) may be impaired in various ways, e.g., accented speech, disordered speech, energetic masking, hearing impairment (see Mattys et al., 2012). However, much work on speech perception in adverse conditions has focused on degradations arising from energetic masking. Energetic masking is a useful characterisation of challenging listening conditions not only because it is relevant to almost every real-world context, but also because it is an objective attribute of the overlap between competing auditory streams, whatever those streams are (unlike, for example, the highly idiosyncratic signal degradation arising from an individual's particular pattern of hearing impairment). Energetic masking is therefore the definition of data limit used in the current study. With this in mind, the question of whether divided-attention listening in a multi-talker environment is data-limited or resource-limited can be said to depend critically on the level of energetic masking: when energetic masking is high, portions of the target speech will simply be unavailable to the listener, thus creating a data limit.

Applying Norman and Bobrow's (1975) framework to divided-attention listening, we first anticipate that the reduction in energetic masking arising from spatial separation should improve data availability and allow additional processing resources to positively impact performance (a *spatial benefit*). At the same time, tracking streams across spatial locations should put high demands on these processing resources and possibly affect performance negatively (a *spatial cost*). Thus, we expect that benefits and costs will trade off as spatial separation increases. Furthermore, since spatial separation should allow additional processing resources to impact performance – in contrast to data-limited situations, in which additional resources should have only a small impact – we also expect a stronger relationship between individual differences in cognitive abilities and performance when divided-attention listening involves a high degree of spatial separation (resource limit) than when it involves a high degree of energetic masking (data limit).

The trade-off described above, if confirmed, could explain the mixed findings reported in the literature regarding the role of cognitive abilities – and specifically WM – during speech perception in challenging listening environments. Indeed, the proposed link between good WM and successful speech perception in adverse conditions remains debated when considering younger adults with audiometrically normal hearing (Besser et al., 2013; Bianco & Chait, 2022; Füllgrabe & Rosen, 2016). One potential explanation for this ambiguity is that many studies have focused on adverse conditions involving energetic masking, making it likely that the tasks in those studies were data-limited. As discussed above, a data limit would substantially reduce the likelihood of observing a link between performance and cognitive resources and, by extension, a link between performance and individual differences in WM. Only studies where the speech perception tasks were resource-limited would be able to demonstrate such a link, hence the discrepancies in reported findings.

It is worth noting that the data-limit explanation is in contrast to Rönnberg et al.'s (2010, 2019, 2021) Ease of Language Understanding (ELU) model. The ELU model postulates that increased background noise should lead to greater reliance on cognitive resources during speech processing, thus giving rise to a *stronger* observable link between WM and speech perception during conditions with more energetic masking. However, support for the ELU model comes largely from

selective attention studies involving older and/or hearing-impaired listeners (e.g., Arehart et al., 2013; Ng & Rönnberg, 2020; Rudner et al., 2011; see Rönnberg et al., 2022, for a review). As discussed above, hearing impairment introduces an additional and idiosyncratic type of source degradation beyond energetic masking alone, and it is conceivable that different types of degradation may interact with WM in different ways. For example, it is plausible that the recruitment of WM is more effective in counteracting degradation caused by listener-based factors (e.g., hearing impairment) than degradation caused by external factors (e.g., background noise). This notwithstanding, there is some empirical support for a stronger (as opposed to weaker) relationship between WM and speech perception with increased energetic masking for young, normal-hearing listeners (e.g., Michalek et al., 2018). Ultimately, whether the Norman and Bobrow (1975) framework or the ELU model best accounts for the relationship between cognition and speech perception in a divided-attention listening scenario for younger, normal-hearing listeners remains an empirical question.

Few studies have directly explored the trade-off between spatial benefits and costs during divided-attention listening, and the findings are mixed. Moreover, the methods used in those studies differed significantly. Ihlefeld and Shinn-Cunningham (2008) used pairs of sentences from the CRM corpus (Bolia et al., 2000) and presented them as collocated (either at 0° or 90° azimuth) or spatially separated (one voice at each location). Participants reported the content of both sentences. Pinto et al. (2020) used word lists produced by four talkers, presented either as collocated or in different spatial locations (−80°, −40°, +40°, +80° azimuth). Participants were asked to detect a target word, which was either produced by one designated talker (selective attention) or could be produced by any of the talkers (divided attention). Both studies reported a benefit of spatial separation on the ability to track multiple voices simultaneously. Moreover, Pinto et al. (2020) showed that the magnitude of the spatial benefit was comparable across the selective and divided attention conditions, implying that divided-attention listening is dominated by the benefit arising from release from energetic masking, with little or no spatial cost. However, this contrasts with Best et al. (2006), who showed that the cost of tracking two voices compared to one grew as spatial separation between the voices increased, suggesting that increased demands on spatial attention allocation can negatively impact performance.

The above studies used either a task featuring closed response sets (CRM sentences) or a task involving online monitoring of word lists (Pinto et al., 2020). These tasks are far removed from real-world communication and may not fully engage the cognitive processes associated with the perception of connected speech, e.g., syntactic parsing and semantic integration, which are known to recruit broader cognitive processes such as WM (Best et al., 2018; MacDonald & Christiansen, 2002; Waters & Caplan, 1996). As a result, they may be less sensitive to factors affecting top-down processing, such as spatial separation, than tasks using more natural stimuli. Sample sizes in these studies were also generally small, with both Best et al. (2006) and Ihlefeld and Shinn-Cunningham (2008) testing fewer than ten participants per experiment.

The present study

In the present study, we attempted to address the trade-off between spatial benefits (i.e., release from energetic masking) and spatial costs (i.e., exercising attentional control) using a "split-listening" paradigm similar to that used by Best et al. (2006). This paradigm measures listeners' ability to track two simultaneous talkers varying in their perceived spatial location. In each trial, participants heard a male speaker and a female speaker simultaneously over headphones. Each speaker said a meaningful but low-predictability sentence (e.g., M: *The box was thrown beside the parked lorry*; F: *Glue the paper to the dark blue background*). Participants were asked to pay attention to both speakers. Immediately after stimulus offset, one of the two voices was cued (e.g., male) and participants had to report the content of the sentence spoken by the cued speaker. The voice to report was specified at the end of stimulus presentation to avoid strategic listening.[1] The relative intensity of the two voices was manipulated to reflect different degrees of perceived spatial separation, from collocated (diotic presentation) to maximally separated (dichotic presentation, i.e., each voice presented in a separate stereo channel).

As described earlier, spatial separation sets up a trade-off between factors related to energetic masking, on the one hand, and factors related to attentional control on the other. In the collocated condition, energetic masking is high, but the burden on spatial attentional control is low, since the two voices originate from the same location. By contrast, in the maximally separated (dichotic) condition, energetic masking should be minimal, but the burden on spatial attentional control is high, since tracking both voices requires dividing one's attention between widely-spaced locations. Intermediate conditions (near, far) offer intermediate proportions of energetic masking and attentional demands. As a result, the paradigm allows the relative contributions of spatial benefits and costs during divided-attention listening to be decoupled. Fig. 1 presents a schematic illustration of the relative contributions of energetic masking and cognitive demands across the different spatial conditions.

In Experiment 1, the stimuli were created from natural speech, thus implementing real-world energetic masking. Experiment 2 was similar to Experiment 1, except that the voices were bandpass filtered such that each speaker was in a separate frequency region. This ensured the absence of spectral overlap between the two voices (i.e., no energetic masking). We hypothesised that the energetic masking present in Experiment 1 would create a strong data limit in the collocated condition, and that transcription performance would therefore be dominated by release from energetic masking as the two voices became more spatially separated. In other words, we predicted a clear spatial benefit, with performance improving as perceived distance between targets increased, thus replicating the results from previous studies (Ihlefeld & Shinn-Cunningham, 2008; Pinto et al., 2020). Although a spatial cost may also be present in this context, we expected it to be smaller in size than the benefit arising from spatial release from energetic masking, leading to a net improvement in performance with increased spatial separation. In contrast, with energetic masking (and hence the data limit) removed, Experiment 2 should primarily be a test of processes related to attentional control. We therefore expected a smaller improvement in performance with increased spatial separation compared to Experiment 1 – if any at all – with the additional prediction that performance might even drop for conditions with the largest spatial separation.

To address the contribution of individual differences in WM to divided-attention listening, participants' WM capacity was measured using the Letter-Number Sequencing task (LNS), which is designed to tap into both verbal short-term memory capacity and executive control (Wechsler, 1997). On the whole, we expected a positive relationship between LNS scores and transcription accuracy, with higher LNS scores associated with higher accuracy. Critically, however, we also expected that this relationship would emerge most strongly when the transcription task was not data-limited, that is, in the spatially separated conditions of Experiment 1 and in all the conditions of Experiment 2. In both cases, we hypothesised that the reduction in, or absence of, energetic masking would increase the potential for individual differences in cognitive abilities (WM) to impact performance.

[1] The fact that participants were required to report the content of only one voice minimised unnecessary task demands. Although the split-listening task is likely to tap into WM for basic maintenance and recall of the materials, the load should be comparable across all spatial conditions. Any variability in the contribution of WM to performance across the spatial conditions is therefore likely to reflect the role of WM in identification – the question of interest – rather than in maintenance and recall.

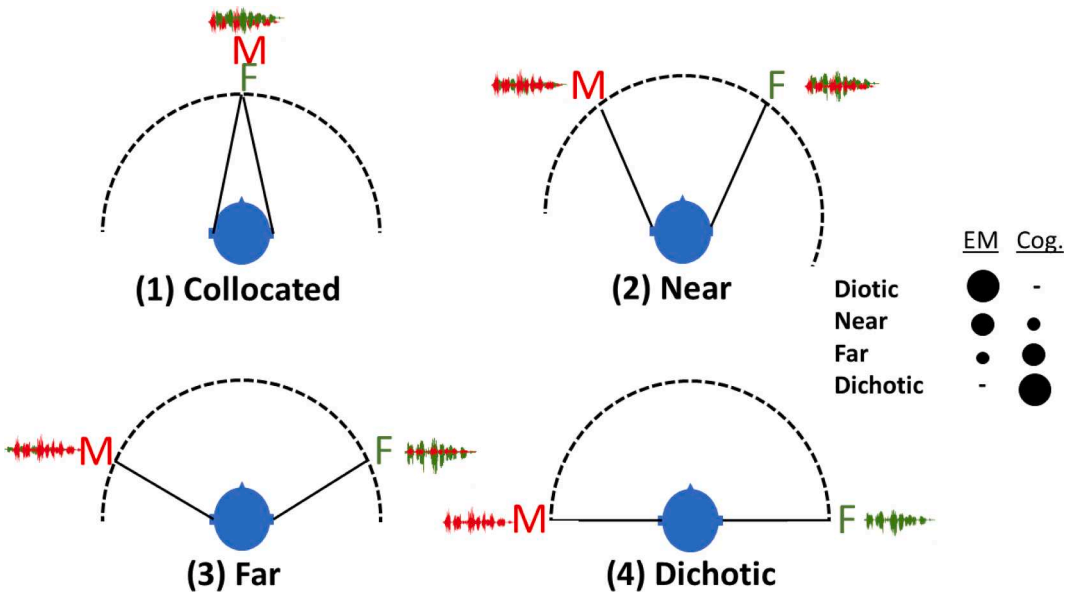


Fig. 1. Schematic of the split-listening paradigm used in Experiments 1 and 2. Participants are asked to track two simultaneous voices, male (M) and female (F). The relative spatial location of the two voices is manipulated (collocated, near, far, dichotic), leading to listening conditions with different proportions of energetic masking (EM) and cognitive (attentional) demands. Right-side inset: The size of the black circles represents the relative contributions of energetic vs. cognitive demands.

Experiment 1

Data availability and ethics

All stimuli, data and analysis code for both experiments are available at https://osf.io/vznhu/. Ethical approval for the experiments was granted by the local departmental ethics committee (ref. 757), all participants provided informed consent before taking part, and all procedures were performed in compliance with relevant laws and institutional guidelines.

Methods

Participants

Participants were recruited via the Prolific recruitment platform (www.prolific.co) and testing was carried out via the online testing platform Gorilla (www.gorilla.sc; Anwyl-Irvine et al., 2020). The sample size was chosen on the basis of the planned analyses. We chose to use generalised linear mixed models (GLMMs) to analyse our data (see Analyses below), and it has been suggested as a rule-of-thumb that a minimum of 1600 observations per condition is necessary to adequately power a mixed effects model (Brysbaert & Stevens, 2018). Given 20 trials in each condition (see Procedure below), a total of 160 participants were recruited in the first instance to ensure double this number of observations per condition. However, data from 7 participants were discarded due to poor performance on the LNS task (see Analyses). This left a final sample of 153 participants (mean age = 27.8 [SD = 6.3], female/male = 79/74). All participants initially completed a headphone check based on a task designed and validated by Woods et al. (2017). In each trial of this task, participants were asked to identify the quietest of three tones. The use of antiphase audio for some of the tones meant that this task could only be successfully completed with stereo headphones. Participants were required to correctly respond to at least 5/6 trials in order to continue with the experiment. Recruitment filters were used on Prolific to ensure that all participants spoke English as their first language, had self-reported normal hearing, and had a Prolific approval rating of at least 90%. Approval ratings reflect the number of Prolific studies for which a participant has not been rejected by the researcher (e.g., because of attention check failures), and therefore indicates level of compliance and engagement in previous studies.

Materials

The stimuli for the split-listening task were drawn from a set of 750 meaningful sentences (IEEE sentences; Rothauser, 1969) modified to fit modern British English. Each sentence contained 5 keywords which were used to score transcription accuracy. All sentences were spoken by a male talker and a female talker, both native British English speakers. The male and female speakers had average f0 values of 164 Hz (SD = 50 Hz, range = 80–332 Hz) and 213 Hz (SD = 73 Hz; range = 117–494 Hz), respectively. The male and female sentences were pseudo-randomly paired, and a subset of 100 pairs with the closest male–female matching durations were selected. Some pairs were manipulated using Adobe Audition to ensure that the largest difference in duration was less than 50 ms. The average duration of the pairs was 2.7 s (SD = 0.05 s; range = 2.6–2.8 s). Of the 100 pairs, 80 were chosen for the main task and a further four were used for practice trials.

The relative intensity of the male and female sentences within a pair was manipulated to create four spatial conditions: diotic (both speakers presented in both channels), 33% dichotic, 66% dichotic, and 100% dichotic audio (each speaker presented in a separate stereo channel). For simplicity, these conditions will be referred to as collocated, near, far, and dichotic, respectively. This was achieved using the panning function from the audio manipulation module Pydub in Python (https://pydub.com/). To create dichotic stimuli, the male and female sentences were played in separate channels. To create the far condition, the panning function was applied to the dichotic stimuli to create versions in which, when wearing headphones, one talker appeared to be approximately 60° to the left of the listener and the other talker approximately 60° to the right. To create stimuli for the near condition, the process was repeated to create apparent locations of approximately 30° to the left and right. To create the collocated condition, the sentence pairs were combined to make one-channel mono audio. All individual sentences were root-mean-square (RMS) equalised to the same level before mixing and applying the panning function. After applying the panning function, all stimuli were then RMS-equalised again to the same overall level.

Procedure

During the initial audio checks, participants first heard a brief

segment of white noise which had been RMS-equalised to the same level as the stimuli in the main task, and were asked to adjust their volume to an audible and comfortable listening level. Participants then completed the headphone check task described above.

Participants who passed the headphone check proceeded to the main split-listening task. For this task, participants were assigned to one of two designs, in which the spatial location of the stimuli either varied randomly from trial-to-trial or was blocked (*mixed* or *blocked,* respectively). This dual-design set-up was intended to rule out any order or practice effects related to specific spatial configurations – in other words, to test the robustness of any spatial effects given trial-to-trial (as opposed to block-by-block) changes in perceived speaker location. Within the *mixed* design, participants were assigned to one of four groups. All participants heard the same sentence pairs, but the spatial condition (collocated, near, far, dichotic) for each unique pair was counterbalanced between the groups. For the *blocked* design, the order of presentation of blocks was also partly counterbalanced, leading to eight groups in total. For the near, far, and dichotic trials, the assignment of talker (male vs. female) to channel (left vs. right) was randomised. For all participants, the task started with four practice trials, one in each spatial condition. The main task consisted of 80 trials with the opportunity for a short break every 20 trials. For participants in the mixed design, all 80 sentence pairs were presented in a random order, divided into 4 blocks of 20 trials. For participants in the blocked design, each block of 20 trials corresponded to one of the four spatial conditions, with the order of sentence pairs randomised within that block and the order of block presentation partly counterbalanced across participants.

On each trial, participants were presented with a fixation cross while the audio stimulus played. After listening to each stimulus, participants were given a visual prompt ("Male" or "Female") as to which talker to transcribe. The male voice was prompted in 40 of the 80 sentence pairs and the female voice in the other 40 pairs. Participants were asked to type their responses into a response box and press the enter key to start the next trial. They were given a maximum of 60 s to respond before the next trial started.

After completing the split-listening task, participants performed the Letter-Number Sequencing (LNS) task, a measure of working memory. On each trial, participants heard a sequence of letters and numbers spoken by a male talker. Immediately afterwards, they were asked to type out the letters in alphabetical order followed by the numbers in ascending order. Sequences began with a length of two and increased in length by one item every three trials, with a maximum length of eight. The LNS is relatively simple compared to some WM tasks. However, it taps into the core features of WM (i.e., storage and manipulation), and has been shown to relate to performance on a range of speech-in-noise perception tasks (e.g., Heinrich & Knight, 2016) as well as to other, more complex WM tasks (e.g., the size-comparison span; Heinrich et al., 2016). It is also straightforward to implement online and has good test–retest reliability (Heinrich & Knight, 2020).

At the end of the experiment, participants were given the opportunity to leave comments and mention if they used any particular strategies while participating.

Analyses

Each target sentence contained 5 keywords. For each trial, performance transcription was averaged across the five keywords. The outcome variable was therefore per-trial scores for each participant expressed as a proportion of keywords correctly transcribed.

For the LNS task, trials were scored as incorrect if there was any deviation from a fully correct response. If participants got all three trials of a particular length incorrect, the task ended; otherwise there was a maximum number of trials (and hence maximum score) of 21. Of the 160 participants initially recruited for the study, 7 participants scored less than 3 in the LNS task, indicating that they were not able to successfully complete the task even with a list length of 2 (i.e., one letter and one number). Because such low performance likely reflects misunderstood instructions and/or inattentiveness rather than genuinely poor WM, these participants were removed from all subsequent analyses. The final dataset was therefore slightly unbalanced, since there were unequal numbers of participants in each group. However, our primary statistical analyses involve the use of GLMMs, which are typically robust to unbalanced data except in extreme cases (e.g., Schielzeth et al., 2020).

Data were analysed in R (v. 4.1.1), using RStudio (v. 1.4.1717) and the packages *dplyr*, *tidyr*, *lme4*, *emmeans* (for posthoc pairwise comparisons) and *rstatix* (to generate summary statistics). Generalised linear mixed-effects models (GLMMs) with a binomial distribution and logit link were used to model the proportion correct scores. Since participants heard all four spatial conditions and all items (sentence pairs) were presented in all four conditions, we were able to fit a maximal random effects structure (Barr et al., 2013) with random intercepts for participants and items, and, for each intercept, a correlated random slope for spatial condition. The fixed effects structure included spatial condition (dichotic, far, near, collocated), design (mixed, blocked), LNS score (centred), and their interactions. Spatial condition was treatment coded with the collocated condition as the baseline. Likelihood ratio tests (LRTs) were used to determine whether the fixed effects contributed significantly to the model. Specifically, the full model as described above was compared to a reduced model that did not include the effect of interest. All models used the BOBYQA optimiser (Powell, 2009) and a maximum of 10^9 iterations. The full model specification was as follows:

score ~ condition + design + LNS + condition:design + condition:LNS + design:LNS + condition:design:LNS + (1 + condition|participant) + (1 + condition|item)

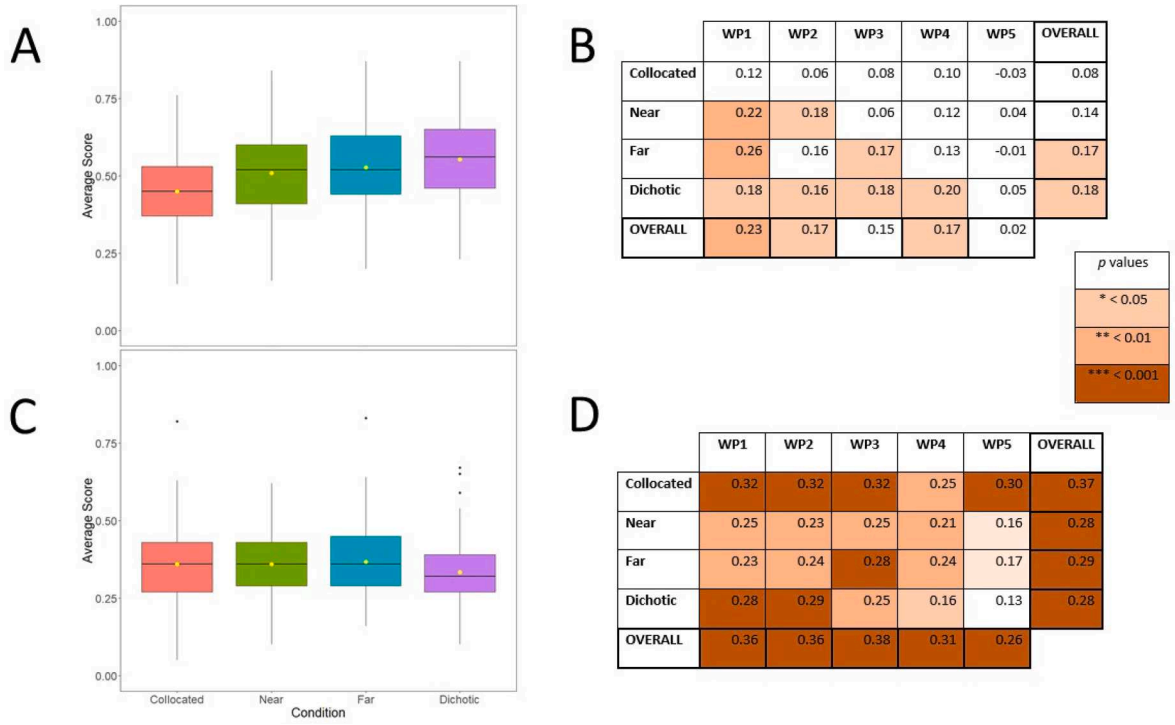
The full analysis code is available in the open materials.

Posthoc pairwise comparisons for the GLMMs were conducted using the *pairwise* argument from the *emmeans* function, which conducts z-tests on the model data and produces Tukey-corrected p-values. We also explored intrusion errors in participants' responses, that is, words reported from the non-target voice. The absolute number of intrusions was low, however, so we were not able to model these in a similar way to the proportion correct scores.

Results

Overall performance was around 50% correct across all four spatial conditions (see Fig. 2, panel A). This is somewhat better than the performance levels reported for a comparable paradigm by Best et al. (2006), where scores were typically below 35% correct. Average performance on the LNS task was 12.6 (SD = 3.7). Results from the initial GLMM modelling showed no significant effects or interactions involving the design factor (mixed, blocked) (see Appendix). Since the effect of design was not our primary research question, and to maximise statistical power, we removed the design term from the fixed effects structure. This created a new model with fixed effects of condition, LNS, and their interaction. LRTs on this model indicated no significant interaction between condition and LNS, $X^2(3) = 3.41$, $p = .33$, and no main effect of LNS, $X^2(1) = 1.18$, $p = .28$. However, there was a significant main effect of condition, $X^2(3) = 70.16$, $p < .001$, reflecting better performance as spatial separation increased. Posthoc pairwise comparisons based on estimated marginal means from the GLMM indicated that performance in the collocated condition was significantly worse than in all other conditions (all $p < .001$). Performance in the near condition was significantly worse than in the dichotic condition ($p < .001$). Performance in the far condition was significantly worse than in the dichotic condition ($p = .03$). The near and far conditions were not significantly different to each other ($p = .26$). Raw means for the four conditions are displayed in Fig. 2, panel A. Estimated model means were 44.3%, 50.9%, 53.0% and 56.1% for the collocated, near, far, and dichotic conditions respectively.

Intrusion errors (i.e., words reported from the non-cued target voice)



	WP1	WP2	WP3	WP4	WP5	OVERALL
Collocated	0.12	0.06	0.08	0.10	-0.03	0.08
Near	0.22	0.18	0.06	0.12	0.04	0.14
Far	0.26	0.16	0.17	0.13	-0.01	0.17
Dichotic	0.18	0.16	0.18	0.20	0.05	0.18
OVERALL	0.23	0.17	0.15	0.17	0.02	

	WP1	WP2	WP3	WP4	WP5	OVERALL
Collocated	0.32	0.32	0.32	0.25	0.30	0.37
Near	0.25	0.23	0.25	0.21	0.16	0.28
Far	0.23	0.24	0.28	0.24	0.17	0.29
Dichotic	0.28	0.29	0.25	0.16	0.13	0.28
OVERALL	0.36	0.36	0.38	0.31	0.26	

Fig. 2. A: Average performance (proportion of correct keywords) for each spatial condition in Experiment 1. B: Correlation coefficients (Pearson's *r*) between LNS and transcription accuracy for each spatial condition and word position (WP) in Experiment 1. C: Average performance (proportion of correct keywords) for each spatial condition in Experiment 2. D: Correlation coefficients (Pearson's *r*) between LNS and transcription accuracy for each spatial condition and word position (WP) in Experiment 2. For panels A and C, lower and upper hinges of boxes correspond to the first and third quartiles. Whiskers extend to the largest and smallest values no further than 1.5*IQR from the mean. Horizontal bars indicate condition medians and yellow dots indicate condition means. For panels B and D, significance levels are indicated by orange shading, with darker colours indicating smaller *p* values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

were rare, constituting only 3.9% of responses overall, and this rate was consistent across the four spatial conditions (3.9%, 3.7%, 3.7%, and 3.8% for the collocated, near, far, and dichotic conditions, respectively).

Sentence recall is a complex process involving semantic, lexical, and phonological cues. Nevertheless, there is evidence for something approximating a serial recall curve (Ebbinghaus, 1913) for sentence materials, in which performance is better (i.e., fewer errors) for early and late items – albeit with local fluctuations reflecting linguistic categories and structures (e.g., Mandler & Mandler, 1964; Wearing, 1971). Thus, memory processes are likely to differ according to the position of a to-be-remembered word within a sentence. With this in mind, we decided to conduct an exploratory analysis of the LNS scores according to word position. Although there was no significant main effect of LNS on performance in the GLMM analysis, it is plausible that WM might contribute to performance more strongly at some points within a sentence than others. We therefore examined the correlations between LNS scores and transcription accuracy for each word position (1 to 5) in each spatial condition (Fig. 2, panel B). Because this analysis was not part of our original design, there was not sufficient power to include word position as a predictor in an additional GLMM, and all reported correlational analyses are therefore exploratory. These caveats notwithstanding, the results highlight two broad patterns. First, the relationship between LNS and transcription accuracy was stronger for words occurring towards the beginning than towards the end of sentences. This likely reflects the greater involvement of WM in maintaining sentence-initial words than more recently-heard words. Second, the relationship between LNS and transcription accuracy was weakest for the collocated condition and strongest for the dichotic condition. This may reflect the data limit caused by energetic masking in the collocated condition and the absence of energetic masking in the dichotic condition. As discussed earlier, the auditory inaccessibility of portions of the target in the collocated condition may have reduced the extent to which cognitive resources, such as WM, were able to contribute to transcription performance, thus weakening the relationship between LNS and transcription accuracy in this condition.

Discussion

The main effect of spatial condition on transcription accuracy suggests a benefit arising from spatial release from masking. Thus, listeners' ability to track two voices simultaneously seems to have been driven primarily by factors related to the acoustic integrity of the speech signal, with no evidence of a spatial cost arising from attentional demands, even when spatial separation between target voices was large, as in the dichotic condition. We also observed a low rate of intrusion errors (i.e., words reported from the non-cued target voice), suggesting that participants were successfully able to segregate the two voices. Improved performance with spatial separation can be interpreted in two ways. On the one hand, it is possible that, contrary to our initial assumption, spatial separation does not lead to an increased demand on cognitive resources compared to when talkers are collocated. On the other hand, it is possible that the size of the spatial benefit created by release from masking was large compared to any increased spatial-attentional demands, thus producing a net improvement in performance even under conditions of extreme spatial separation.

Furthermore, energetic masking may have created an intrinsic data limit that made it difficult to observe effects related to individual differences in WM. This possibility was supported by our analyses of the LNS data. Although exploratory, they show that the relationship between LNS scores and transcription accuracy was weakest in the spatial condition with the greatest amount of energetic masking (collocated condition). In order to isolate the cognitive costs of divided-attention

listening and observe relationships with individual cognitive differences, it is therefore necessary to investigate it in a context in which energetic masking has been removed. This was the aim of Experiment 2.

Experiment 2

Methods

Participants

Participants were recruited and tested as for Experiment 1, using the same recruitment filters on Prolific. A total of 160 participants were recruited in the first instance. However, data from 3 participants were discarded due to poor performance on the LNS task (see Analyses for Experiment 1 above). This left a final sample of 157 participants (mean age = 27.5 [SD = 6.0]; female/male = 96/61).

Materials

The stimuli were those of Experiment 1, except that the male and female sentences were filtered into non-overlapping frequency bands before being combined to create the different spatial conditions. Specifically, the sentences produced by the male talker were lowpass-filtered below 1400 Hz, while the sentences produced by the female talker were bandpass-filtered between 1500 Hz and 6000 Hz, in both cases with a smoothing window of 50 Hz. These values were chosen because pilot data comparing various other combinations indicated that these specific filters had a similar effect on intelligibility for both the male and female sentences. Fig. 3 shows the spectrograms for a sentence pair as it appeared in Experiment 1 (panel A) vs. Experiment 2 (panel B), with the male talker in blue and the female talker in orange. The two talkers' overlapping spectra (i.e., energetic masking) are visible in the former, as is their clear spectral separation (i.e., no energetic masking) in the latter.

Procedure and analyses

Experimental procedure and statistical analyses were as for Experiment 1.

Results

Overall performance was around 35% correct across all four spatial conditions (Fig. 2, panel C), which is lower than in Experiment 1 and likely a consequence of reduced naturalness caused by the filtering procedure. However, this level of performance is comparable to that reported in Best et al. (2006). As anticipated, the impact of the filtering was similar for the male and female talkers, with average performance for male vs. female sentences of 35.3% and 35.8% respectively. Average performance on the LNS task was 13.0 (SD = 3.4). As before, the initial GLMM model included fixed effects of spatial condition (dichotic, far, near, collocated), design (mixed, blocked), LNS score (centred), and their interactions. This model revealed no significant effects or interactions involving the design factor (see Appendix); we therefore removed the design factor from the model, as for Experiment 1. This created a new model with fixed effects of condition, LNS, and their interaction. LRTs on this model revealed a significant main effect of condition, $X^2(3) = 12.42$, $p < .01$. Posthoc pairwise comparisons based on model means indicated that this effect was driven by poorer performance in the dichotic condition: scores in this condition were significantly worse than in the far condition ($p < 0.01$) and marginally worse than in the near condition ($p = 0.058$). There were no significant differences between the collocated, near, and far conditions (all $p > .70$). Raw means for the four conditions are displayed in Fig. 2, panel C. Estimated model means were 34.7%, 34.9%, 35.7%, and 32.2% for the collocated, near, far, and dichotic conditions respectively. LRTs also revealed a significant main effect of LNS scores, $X^2(1) = 24.41$, $p < .001$, indicating that participants with higher LNS scores also had higher transcription scores. There was no significant interaction between condition and LNS, $X^2(3) = 4.90$, $p = .18$.

Intrusion errors were even less frequent in absolute terms than in Experiment 1, constituting only 1.8% of responses overall. This may suggest an enhanced ability to segregate the target voices due to the spectral separation between them. However, the lower intrusion rate is also consistent with the generally poorer accuracy in this experiment, which suggests that perception of both target voices was vulnerable to the reduced naturalness caused by the filtering process. The rate of intrusions was consistent across the different spatial conditions (1.8%, 1.7%, 2.0%, and 1.9% for the collocated, near, far, and dichotic conditions, respectively).

Finally, we explored the correlation between LNS scores and transcription accuracy across word positions, as reported in Fig. 2, panel D. Significant correlations were found for every combination of word position and spatial condition, with the exception of positions 4 and 5 in the dichotic condition. In other words, good WM capacity was associated with good transcription performance across spatial locations and word positions, but that relationship was somewhat weaker for words occurring towards the end of sentences and only when spatial separation between the target talkers was large. Note too, that, although significant correlations were found for all spatial conditions, the consistently strongest correlations occurred in the collocated condition. This is in sharp contrast to Experiment 1, in which the collocated condition produced the weakest correlations between LNS and transcription accuracy.

Comparison between Experiment 1 and Experiment 2

To make a direct comparison between the patterns of performance and contribution of WM across Experiments 1 and 2, we ran a GLMM combining the data of both experiments. The random effects structure included random intercepts for participants and items, with correlated random slopes for spatial condition in both cases. The fixed effects included spatial condition (collocated, near, far, dichotic), LNS score (centred), experiment (1, 2), and their interactions. The effects of interest were those involving an interaction with the experiment factor. Results from LRTs indicated a significant interaction between condition and experiment, $X^2(3) = 115.00$, $p < .001$, confirming the contrasting effect of spatial separation across the two experiments. Specifically, Experiment 1 showed a benefit of spatial separation which grew with increased spatial separation. This benefit was not observed in

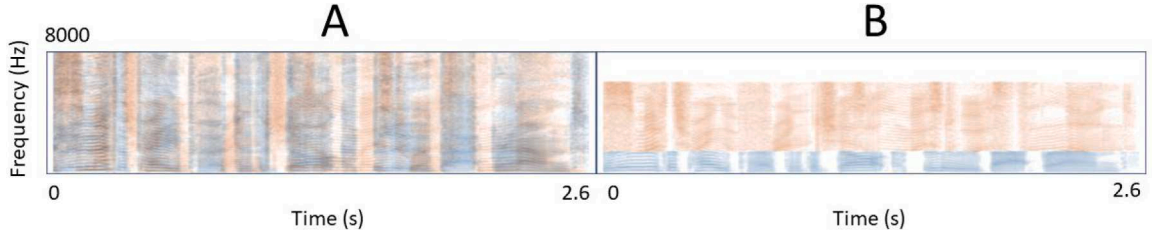


Fig. 3. Spectrograms for the same sentence pair presented in Experiment 1 (panel A) and Experiment 2 (panel B), with the male talker in blue and the female talker in orange. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Experiment 2, and a small but significant detriment to performance was observed in the dichotic condition. LRTs also revealed a significant interaction of LNS × experiment, $X^2(1) = 6.77$, $p < .01$, showing that LNS scores were a better predictor of transcription accuracy in Experiment 2 than Experiment 1. Finally, there was a significant 3-way interaction of condition × LNS × experiment, $X^2(3) = 8.18$, $p = .04$. The relationship between LNS scores and transcription accuracy was not only stronger for Experiment 2 than for Experiment 1, it was particularly strong for the collocated condition in Experiment 2.

Discussion

Experiment 2 shows that the benefit of spatial separation observed in Experiment 1 was largely due to release from energetic masking. Indeed, when energetic masking was removed through bandpass filtering, that benefit disappeared, and a cost of spatial separation was observed in the dichotic condition. These results suggest that energetic masking is dominant during divided-attention listening (thus resulting in a spatial benefit), but also that the challenges of cognitive control due to spatial separation (i.e., a spatial cost) can be observed when energetic masking is removed. This pattern in turn suggests that the recruitment of cognitive processes is more likely to support successful listening in the absence of a data limit. The LNS data were also consistent with this claim: high LNS scores (good WM) were associated with high transcription accuracy, but only when energetic masking was removed (Experiment 2). An exploratory break-down by serial word position suggested that the link between LNS and performance held across the majority of spatial conditions and word positions when energetic masking was removed through bandpass filtering. In Experiment 1, however, that link was only reliably apparent in the dichotic condition, i.e., when energetic masking was removed through spatial separation. Although statistical power was too low to systematically test this difference, the overall pattern suggests that the contribution of cognitive abilities – and specifically WM – to speech perception is most easily observed in the absence of energetic masking, consistent with the data-limit claim.

General discussion

Speech perception in multi-talker environments is challenging, due to both energetic masking (i.e., spectrotemporal overlap between target and masker voices) and informational masking (i.e., interference with stream segregation and attention allocation). Spatial separation between target and masker voices has been shown to improve performance during selective-attention listening by reducing energetic masking and thus increasing target intelligibility. However, it is unclear how spatial separation between two voices may impact performance during divided-attention listening, when both voices need to be tracked simultaneously. Specifically, although spatial separation during divided-attention listening should provide the same benefits as during selective-attention listening in terms of reductions in energetic masking, it also introduces the need to exercise additional cognitive control in order to divide auditory attention between spatial locations and compensate for information loss, thus potentially creating a cost. We have argued that these spatial benefits and costs can usefully be conceptualised in terms of Norman and Bobrow's theory of data- vs. resource-limited tasks (Norman & Bobrow, 1975). According to this framework, any given task can be characterised by the extent to which performance depends on the allocation of processing resources. A task is resource-limited if performance is contingent on the resources allocated to that task. In contrast, a task is data-limited if performance is determined primarily by the amount and quality of the data available, and the allocation of additional processing resources does not lead to further improvements.

In the case of speech perception in multi-talker environments, energetic masking is a critical determinant of the relationship between performance and resources. When energetic masking is high, portions of the target speech are unavailable to the listener, thus rendering the task data-limited. Considered from this perspective, spatial separation during divided-attention listening should alleviate the data limit by reducing energetic masking, and hence, improve performance. However, spatial separation may also increase the processing resources needed to control attention across spatial locations and compensate for lost information, effectively introducing a cognitive cost. The alleviation of the data limit though spatial separation should also move the task into a resource-limited mode, whereby the allocation of additional cognitive resources (if available) has the potential to improve performance. As a result, spatial separation should increase our ability to observe differences in performance related to individual differences in cognitive abilities.

In the current experiments, we addressed these hypotheses using a split-listening paradigm, in which participants heard meaningful sentences spoken by two talkers simultaneously over varying degrees of spatial separation. Participants had to attend to both speakers and then report the sentence spoken by one of the two. The voice to report was specified at the end of stimulus presentation to avoid strategic listening. In Experiment 1, the stimuli were created from natural speech and thus contained real-world-like energetic masking. In Experiment 2, the voices were spectrally filtered so that each speaker was in a separate frequency region, thus removing energetic masking.

Overall performance

Average split-listening performance was around 50% correct in Experiment 1 and 35% correct in Experiment 2. In both cases, this was somewhat better than the average performance level reported by Best et al. (2006), where scores in a similar paradigm were typically below 35% correct. The poorer performance in the Best et al. studies may be due to the fact that the same talker was used for both target voices on any given trial, thus increasing the difficulty of segregating the two streams, or to other differences in stimuli, set-up, and task demands.

The poorer performance in Experiment 2 compared to Experiment 1 is likely to reflect the reduced naturalness of the sentences after the filtering process. Although performance still remained within an acceptable range for comparable multi-talker experiments, future studies should explore alternative filtering procedures to attempt to preserve naturalness and intelligibility in the targets as far as possible (for example, by using multiple non-overlapping frequency bands across the frequency range; see Experiment 1 from Best et al., 2006; Ihlefeld & Shinn-Cunningham, 2008).

Intrusions were rare in both experiments across all conditions. This suggests that participants were able to successfully segregate the two target voices even when they were collocated and energetic masking was high. This is likely due to the distinct spectral characteristics of the two voices (i.e., male vs. female) and to the semantic and syntactic coherence of each sentence acting as an additional streaming device.

Effect of spatial separation

Spatial separation between target voices had a clear beneficial effect when energetic masking was present (Experiment 1) – performance improved monotonically as spatial separation increased. However, there was no benefit of spatial separation when energetic masking was removed (Experiment 2). In fact, there was evidence of a decrease in performance when the target voices were maximally separated. These results suggest that the benefits of spatial separation are largely due to release from energetic masking, and hint at a cost of spatial separation once energetic masking is removed. In terms of the data- vs. resource-limit framework, these results indicate that energetic masking was imposing a strong data limit in the collocated condition of Experiment 1. The reduction in energetic masking caused by spatial separation alleviated this limit, allowing for a considerable improvement in performance – an improvement that dominated the pattern of results, making any costs of spatial separation difficult to detect. In Experiment 2, by

contrast, the task was largely resource-limited, thus allowing costs associated with the increased need for cognitive control to become apparent.

It is worth noting that posthoc pairwise comparisons for Experiment 2 indicated a significant difference only between the dichotic (maximally separated) and far conditions, with a marginally significant difference between the dichotic and near conditions. One interpretation of this pattern is that an intermediate degree of spatial separation was ideal for split-listening performance under these conditions, affording enhanced streaming based on spatial separation without imposing too large a demand on spatial attentional control. This explanation can be seen as an extension of findings in the domain of selective attention, where it has been shown that listeners benefit from turning their heads slightly away from a target talker when masker noise is presented from a different spatial location, thus maximising spatial release from masking and optimising conditions for better-ear listening (Grange & Culling, 2016a; 2016b). In a divided-attention listening situation, however, the ideal spatial locations of the target talkers relative to the listener's head position would need to balance spatial release from masking and binaural segregation cues with spatial attentional control demands. As a result, an intermediate separation (akin to our "far" condition) may produce the best performance. This explanation is speculative, however, and therefore requires further exploration.

The role of working memory (WM)

WM is hypothesised to be important for speech perception in challenging listening environments (e.g., Rönnberg et al., 2021). However, the Norman and Bobrow (1975) framework suggests that the influence of WM may not be evident in listening contexts where energetic masking is imposing a significant data limit. As discussed above, tasks need to be resource-limited for variations in processing resources – such as WM – to have an observable effect on performance. Supporting this hypothesis is Neher et al.'s (2009) finding that hearing-aid users' WM and attention abilities can better predict their speech-in-noise performance if the target and competing talkers are spatially separated than if they are collocated. Additional support comes from Janse and Andringa (2021), who found that the association between WM and word-in-noise recognition for older adults was weaker under more acoustically-challenging listening conditions. Similarly, the review conducted by Humes (2007) suggests that the importance of cognitive factors becomes most apparent for older listeners once audibility passes a certain threshold.

Bearing in mind the data- vs. resource-limit framework, we therefore made two predictions. First, we expected to see a stronger relationship between WM (as measured by the LNS task) and split-listening performance when energetic masking was minimal. Second, we expected to see a stronger relationship between WM and split-listening performance when the highest levels of cognitive control were required (i.e., larger as opposed to smaller spatial separation).

The first prediction was confirmed: we observed a relationship between LNS and split-listening scores only for Experiment 2. This finding is more closely aligned with Neher et al.'s (2009) results than with the predictions of the ELU model (e.g., Rönnberg et al., 2021). However, the current findings and the ELU model are not necessarily in direct conflict. First, as noted above, a speech signal may be "data limited" in a variety of ways (c.f., Mattys et al., 2012). These may include degradation of the source itself (e.g., through filtering), the presence of external distractors (e.g., competing talkers), or listener-based factors which result in an imperfectly encoded signal (e.g., aging, hearing impairment). The stimuli in Experiment 2 could be argued to be "data limited" due to the filtering procedure, which resulted in a drop in overall intelligibility. In other words, Experiment 2 was a more challenging listening situation than Experiment 1, and may thus have recruited relevant cognitive processes more strongly (Akeroyd, 2008; Rönnberg et al., 2019). Also, and as noted above, it is conceivable that different types of degradation may interact with WM in different ways. For example, it is possible that degradation caused by listener-based factors can be ameliorated to a greater degree by the recruitment of WM than degradation caused by energetic masking. Indeed, our results from Experiment 2 already suggest this type of dissociation, with WM coming to the fore when energetic masking (an external degradation) is removed but filtering (a source-based degradation) is introduced.

A second point to consider when interpreting our results relative to the ELU model is that studies supporting a stronger role for WM when the speech signal is impoverished have typically used selective, not divided, attention tasks. It may therefore be the case that the role of WM, and its relationship to speech perception under conditions of degradation, changes when listeners are required to track more than one stream simultaneously. Future work should address this possibility by using the split-listening set-up within the context of a selective-attention task – that is, indicating to participants before each trial which of the two target voices to track.

Finally, we used the LNS task to assess WM, whereas many studies based on the ELU model have used more complex WM tasks such as the Reading Span Task (RST) or Size-Comparison Span (SICspan). Our results may therefore reflect the recruitment of a subset of WM processes as indexed by the LNS, whereas results obtained using the RST or SICspan may reflect a broader range of cognitive processes with different relationships to speech perception. More generally, it is important to bear in mind that any results obtained using a single measure of WM will be dependent on the specific task demands of the chosen WM measure.

Our second prediction – that the relationship between WM and split-listening performance would be stronger for larger as opposed to smaller spatial separation – was partially confirmed. On the one hand, we observed no significant interactions between LNS scores and spatial condition in the main analyses. Thus, these analyses did not indicate a stronger relationship between WM and split-listening performance for larger than smaller degrees of spatial separation. On the other hand, subsequent analyses looking at this relationship across word positions in the target sentences lend some support to our hypothesis. As noted above, these analyses were exploratory, and no statistical comparisons between correlation coefficients were carried out. However, two relevant patterns emerge. First, the relationship between WM and split-listening scores was weakest for later word positions. This is perhaps unsurprising, since words occurring later in the target sentence require a shorter period of maintenance in WM before recall. Second, the WM/split-listening relationship was weakest for the collocated condition in Experiment 1, whereas the relationship was fairly strong across all spatial conditions in Experiment 2. This again suggests the presence of a data limit: where energetic masking was highest (the collocated condition of Experiment 1), there was only a limited role for cognitive abilities, whereas in any situation where the data limit was eased – either through spatial release from masking (dichotic condition, Experiment 1) or through filtering (Experiment 2) – the role of cognition became observable. Indeed, the WM/split-listening relationship was *strongest* for the collocated condition in Experiment 2. This may point to a role for WM in bottom-up (segregation) processes rather than top-down attentional control processes, although this is somewhat speculative given the exploratory nature of the analyses.

Finally, it is worth briefly noting the language background of our participants. Not only did all participants report having English as their first language, the majority (~70%) also described their nationality as "United Kingdom", suggesting that they were native speakers of British English, the variety and accent used in the experimental stimuli. However, the remaining minority of participants were from a mixture of locations where other varieties of English are spoken, such as Canada and Ireland. It is possible that listening to a non-native variety of English may have impacted these participants' performance and/or the nature or degree of their recruitment of WM. Future studies should seek to address the role of non-native languages and dialects during divided-attention listening.

Conclusion

In these studies, we explored the effect of spatial separation between two target talkers during divided-attention listening using a novel split-listening task, in which participants had to attend to the two talkers simultaneously over varying degrees of spatial separation. We also measured working memory (using the letter-number sequencing (LNS) task) in order to assess the relationship between split-listening performance and cognitive abilities. When natural speech was used (Experiment 1), performance on the split-listening task was dominated by the degree of energetic masking present in the stimuli, with split-listening scores increasing monotonically as spatial separation between target talkers increased. When energetic masking was removed (Experiment 2), this spatial benefit disappeared and a small spatial cost was found for the greatest degree of spatial separation. Additionally, a significant relationship between split-listening performance and working memory was observed only for Experiment 2. Taken together, these results point to both spatial benefits and spatial costs during divided-attention listening, with spatial separation of targets creating a trade-off between spatial release from masking on the one hand and an increased need for top-down spatial attentional control on the other. They also suggest that the presence of energetic masking may in some cases make it difficult to measure effects related to the recruitment of cognitive resources. Further work is needed to determine the extent to which this trade-off between spatial benefits and spatial costs, and the associated relationships with cognition, are observed when using different paradigms assessing divided-attention listening. More generally, these results highlight the usefulness of the data- vs. resource-limit framework in considering the relationship between acoustic and cognitive factors during speech perception. Future studies should endeavour to assess the ability of this framework to account for speech perception performance across a range of different tasks and listening environments.

Funding

This work was supported by the Leverhulme Trust [grant number RPG-2018–152]. The funder had no involvement in the study design, the collection, analysis and interpretation of data, the writing of the report, or the decision to submit the article for publication.

CRediT authorship contribution statement

Sarah Knight: Conceptualization, Methodology, Formal analysis, Writing – original draft, Visualization. **Lyndon Rakusen:** Methodology, Software, Investigation, Data curation. **Sven Mattys:** Conceptualization, Methodology, Supervision, Writing – review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All stimuli, data and analysis code are available at https://osf.io/vznhu/.

Acknowledgements

We are grateful to Ronan McGarrigle, Alex Mepham and Sophie Meekings for their advice and feedback.

Appendix

Supplementary Table 1. Significance tests from the initial GLMM for Experiment 1. No effects or interactions involving the design factor (mixed, blocked) are significant.

condition	$X^2(3) = 43.48, p < 0.001$
design	$X^2(1) = 1.01, p = 0.32$
LNS	$X^2(1) = 2.90, p = 0.09$
condition x design	$X^2(3) = 3.22, p = 0.36$
condition x LNS	$X^2(3) = 2.06, p = 0.56$
design x LNS	$X^2(1) = 1.65, p = 0.20$
condition x design x LNS	$X^2(3) = 1.43, p = 0.70$

Supplementary Table 2. Significance tests from the initial GLMM for Experiment 2. No effects or interactions involving the design factor (mixed, blocked) are significant.

condition	$X^2(3) = 8.00, p < 0.05$
design	$X^2(1) = 0.00, p = 0.99$
LNS	$X^2(1) = 9.93, p < 0.01$
condition x design	$X^2(3) = 0.55, p = 0.91$
condition x LNS	$X^2(3) = 6.23, p = 0.10$
design x LNS	$X^2(1) = 0.72, p = 0.40$
condition x design x LNS	$X^2(3) = 7.08, p = 0.07$

References

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology, 47* (sup2), S53–S71.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407.

Arbogast, T. L., Mason, C. R., & Kidd, G., Jr (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America, 112*(5), 2086–2098.

Arehart, K. H., Souza, P., Baca, R., & Kates, J. M. (2013). Working memory, age and hearing loss: Susceptibility to hearing aid distortion. *Ear and Hearing, 34*(3), 251.

Arlinger, S., Lunner, T., Lyxell, B., & Kathleen Pichora-Fuller, M. (2009). The emergence of cognitive hearing science. *Scandinavian Journal of Psychology, 50*(5), 371–384.

Axelrod, S., Guzy, L. T., & Diamond, I. T. (1968). Perceived rate of monotic and dichotically alternating clicks. *The Journal of the Acoustical Society of America, 43*(1), 51–55.
Axelrod, S., & Powazek, M. (1972). Dependence of apparent rate of alternating clicks on azimuthal separation between sources. *Psychonomic Science, 26*(4), 217–218.
Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417–423.
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.
Besser, J., Koelewijn, T., Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2013). How linguistic closure and verbal working memory relate to speech recognition in noise—a review. *Trends in Amplification, 17*(2), 75–93.
Best, V., Gallun, F. J., Ihlefeld, A., & Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening. *The Journal of the Acoustical Society of America, 120*(3), 1506–1516.
Best, V., Swaminathan, J., Kopčo, N., Roverud, E., & Shinn-Cunningham, B. (2018). A "buildup" of speech intelligibility in listeners with normal hearing and hearing loss. *Trends in Hearing, 22*.
Bianco, R., & Chait, M. (2022). No link between Speech-in-noise perception and Auditory short-term memory–evidence from a large cohort of older and younger listeners. PsyArXiv. https://doi.org/10.31234/osf.io/mjg4q.
Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America, 107*(2), 1065–1066.
Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America, 109*(3), 1101–1109.
Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition, 1*(1), 9. https://doi.org/10.5334/joc.10
Colflesh, G. J., & Conway, A. R. (2007). Individual differences in working memory capacity and divided attention in dichotic listening. *Psychonomic Bulletin & Review, 14*(4), 699–703.
Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review, 8*(2), 331–335.
Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America, 123*(1), 414–427.
Culling, J.F., & Stone, M.A. (2017). Energetic masking and masking release. In Middlebrooks, J., Simon, J., Popper, A., Fay, R. (Eds.) *The Auditory System at the Cocktail Party*. Springer Handbook of Auditory Research, vol 60. Springer, Cham.
Ebbinghaus, H. (1913). Retention and obliviscence as a function of the time (H. A. Ruger & C. E. Bussenius, Trans.). In H. Ebbinghaus & H. A. Ruger, C. E. Bussenius (Trans.), Memory: A contribution to experimental psychology (pp. 62–80). Teachers College Press.
Edmonds, B. A., & Culling, J. F. (2006). The spatial unmasking of speech: Evidence for better-ear listening. *The Journal of the Acoustical Society of America, 120*(3), 1539–1545.
Engle, R. W., & Kane, M. J. (2004). Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 44, pp. 145–199). Elsevier Science.
Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America, 106*(6), 3578–3588.
Füllgrabe, C., & Rosen, S. (2016). On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds. *Frontiers in Psychology, 7*, 1268. https://doi.org/10.3389/fpsyg.2016.01268
Gatehouse, S., Naylor, G., & Elberling, C. (2003). Benefits from hearing aids in relation to the interaction between the user and the environment. *International Journal of Audiology, 42*(sup1), 77–85.
Grange, J. A., & Culling, J. F. (2016a). The benefit of head orientation to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 139*(2), 703–712.
Grange, J. A., & Culling, J. F. (2016b). Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions. *The Journal of the Acoustical Society of America, 140*(6), 4061–4072.
Heinrich, A., Henshaw, H., & Ferguson, M. A. (2016). Only behavioral but not self-report measures of speech perception correlate with cognitive abilities. *Frontiers in Psychology, 7*, 576.
Heinrich, A., & Knight, S. (2016). The contribution of auditory and cognitive factors to intelligibility of words and sentences in noiseP. van Dijk, et al. (Eds.). *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing, Advances in Experimental Medicine and Biology, 894*.
Heinrich, A., & Knight, S. (2020). Reproducibility in cognitive hearing research: theoretical considerations and their practical application in multi-lab studies. *Frontiers in Psychology*, 1590.
Humes, L. E. (2007). The contributions of audibility and cognitive factors to the benefit provided by amplified speech to older adults. *Journal of the American Academy of Audiology, 18*(07), 590–603.
Ihlefeld, A., & Shinn-Cunningham, B. (2008). Spatial release from energetic and informational masking in a selective speech identification task. *The Journal of the Acoustical Society of America, 123*(6), 4369–4379.
Janse, E., & Andringa, S. J. (2021). The roles of cognitive abilities and hearing acuity in older adults' recognition of words taken from fast and spectrally reduced speech. *Applied Psycholinguistics, 42*(3), 763–790.
Kidd, G., & Colburn, H. S. (2017). Informational masking in speech recognition. In J. Middlebrooks, J. Simon, A. Popper, & R. Fay (Eds.), *The Auditory System at the Cocktail Party. Springer Handbook of Auditory Research* (vol 60). Cham: Springer.
Litovsky, R. Y. (2012). Spatial release from masking. *Acoustics Today, 8*(2), 18–25.
MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109*(1), 35–54.
Mandler, G., & Mandler, J. M. (1964). Serial position effects in sentences. *Journal of Memory and Language, 3*(3), 195.
Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7–8), 953–978.
Michalek, A. M., Ash, I., & Schwartz, K. (2018). The independence of working memory capacity and audiovisual cues when listening in noise. *Scandinavian Journal of Psychology, 59*(6), 578–585.
Mondor, T. A., & Zatorre, R. J. (1995). Shifting and focusing auditory spatial attention. *Journal of Experimental Psychology: Human Perception and Performance, 21*(2), 387.
Neher, T., Behrens, T., Carlile, S., Jin, C., Kragelund, L., Petersen, A. S., & Schaik, A. V. (2009). Benefit from spatial separation of multiple talkers in bilateral hearing-aid users: Effects of hearing loss, age, and cognition. *International Journal of Audiology, 48*(11), 758–774.
Ng, E. H. N., & Rönnberg, J. (2020). Hearing aid experience and background noise affect the robust relationship between working memory and speech recognition in noise. *International Journal of Audiology, 59*(3), 208–218.
Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*(1), 44–64.
Pinto, D., Agmon, G., & Golumbic, E. Z. (2020). The Role of Spatial Separation on Selective and Distributed Attention to Speech. bioRxiv. https://doi.org/10.1101/2020.01.27.920785.
Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06*, University of Cambridge, Cambridge.
Rönnberg, J., Rudner, M., Lunner, T., & Zekveld, A. A. (2010). When cognition kicks in: Working memory and speech understanding in noise. *Noise and Health, 12*(49), 263-269.
Rönnberg, J., Holmer, E., & Rudner, M. (2019). Cognitive hearing science and ease of language understanding. *International Journal of Audiology, 58*(5), 247–261.
Rönnberg, J., Holmer, E., & Rudner, M. (2021). Cognitive hearing science: Three memory systems, two approaches, and the ease of language understanding model. *Journal of Speech, Language, and Hearing Research, 64*(2), 359–370.
Rönnberg, J., Signoret, C., Andin, J., & Holmer, E. (2022). The Cognitive Hearing Science perspective on perceiving, understanding, and remembering language: The ELU model. *Frontiers in Psychology, 13*.
Rothauser, E. H. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*, 225–246.
Rudner, M., Rönnberg, J., & Lunner, T. (2011). Working memory supports listening in noise for persons with hearing impairment. *Journal of the American Academy of Audiology, 22*(03), 156–167.
Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., … Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11*(9), 1141–1152.
Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182–186.
Treisman, A. M. (1971). Shifting attention between the ears. *Quarterly Journal of Experimental Psychology, 23*(2), 157–167.
Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language, 62*(4), 392–406.
Waters, G. S., & Caplan, D. (1996). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review, 103*(4), 761–772.
Wearing, A. J. (1971). Word class and serial position in the immediate recall of sentences. *Psychonomic Science, 25*(6), 338–340.
Wechsler, D. (1997). *Wechsler Adult Intelligence Scale III*. New York: The Psychological Corporation.
Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072.