Parametric cognitive load reveals hidden costs in the neural processing of perfectly intelligible degraded speech

Harrison Ritz*[1,2], Conor Wild[1], and Ingrid Johnsrude[1,3]

1. Brain and Mind Institute, University of Western Ontario, London ON N6A 3K7, Canada
2. Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence RI 02912, USA
3. Departments of Psychology and Communication Sciences and Disorders, University of Western Ontario, London ON N6A 3K7, Canada

Corresponding Author:
Harrison Ritz
Department of Cognitive, Linguistic and Psychological Sciences
Brown University
Providence, RI 02912, USA
email: harrison.ritz@gmail.com

Abstract

Speech is often degraded by environmental noise or hearing impairment. People can compensate for degradation, but this is subjectively effortful. Previous research has identified frontotemporal networks involved in effortful perception, but so far has been unable to distinguish domain-general involvement in cognitive control from domain-specific involvement in speech perception. We overcame these limitations by measuring brain responses (via fMRI) to clear and degraded spoken sentences matched for intelligibility. On each trial, participants either attended to a sentence, or to a concurrent multiple object tracking task that imposed parametric cognitive load. Whereas the anterior insulae tracked domain-general task demands, anterior temporal cortex was selective to speech, with cognitive load revealing categorically different responses to clear speech and intelligibility-matched degraded speech. The results demonstrate that cognitive control is required to understand even very mildly degraded speech, underscoring the importance of interactions between perceptual and control systems for real-world speech comprehension.

148 / 150 words

Keywords: speech perception, cognitive control, functional magnetic resonance imaging

Introduction

In perfect listening conditions, the comprehension of speech is seemingly effortless for healthy young people. However, everyday listening conditions are rarely as good as in the laboratory, and speech understanding is often compromised by noisy environments, low-fidelity digital communication, and by hearing impairment across a large portion of the population including older adults. When such degradation occurs, people must recruit effortful cognitive control to successfully perceive speech (Broadbent, 1958; Eckert et al., 2016; Fedorenko, 2014; Heald & Nusbaum, 2014; Johnsrude & Rodd, 2016; Pichora-Fuller et al., 2016; Rouault & Koechlin, 2018; Vaden et al., 2013). Experimenters have quantified listening effort by measuring participants' perception under distraction, finding domain-general performance costs to auditory perception (Herrmann et al., 2018; Herrmann & Johnsrude, 2018; Macdonald & Lavie, 2011; Raveh & Lavie, 2015) and speech perception (Downs, 1982; Feuerstein, 1992; Hicks & Tharpe, 2002; Fraser et al., 2010; Gagne et al., 2017; Hunter & Pisoni, 2018).

Neuroimaging experiments have supported these behavioral results, finding that cingulo-opercular regions associated with cognitive control (Shenhav et al., 2013) and temporal regions associated with high-level speech perception (Hickok & Poeppel, 2007) are sensitive to intelligibility (Davis & Johnsrude, 2003; Eckert et al., 2016), lose speech sensitivity during distracting tasks (Sabri et al., 2008; Wild et al., 2012), and predict perceptual accuracy (Wild et al., 2012; Vaden et al., 2013, 2015, 2016). These experiments generally support a role for domain-general control networks in speech perception, but previous work has been limited in its ability to reveal the specific division of labor across these systems.

Previous neuroimaging experiments have localized control-dependent speech processes by contrasting responses when speech is either easier or harder to understand, during tasks in which speech is either relevant or irrelevant (Sabri et al., 2008; Wild et al., 2012). This methodology cannot distinguish regions that subserve speech-specific functions from those involved in domain general control processes, because intelligibility-sensitivity does not necessarily imply speech-sensitivity, as has been previously assumed. For example, a domain general region that monitors task performance would appear sensitive to speech intelligibility during comprehension tasks, but only because intelligibility is strongly correlated with accuracy. In contrast, a region that has a domain-specific involvement in effortful speech processing should show intelligibility responses that depend

on concurrent cognitive load, whether or not speech is task-relevant. These two functions are likely to be organized hierarchically, with domain-general control processes in inferior frontal of the ventral speech pathway, and speech-selective processing in more caudal temporal regions of this pathway (Davis & Johnsrude, 2003; Evans & Davis, 2015; Hickok & Poeppel, 2007).

In the current study, we compared the perception of clearly spoken sentences against two levels of degradation: one matched for intelligibility with clear speech (near-perfect word report accuracy), and one with only slightly lower intelligibility (>90% word report accuracy). Since clear speech and intelligibility-matched degraded speech differ in their perceptual clarity, they also differ in putative control demands. These stimuli not only allow us to distinguish between performance monitoring and processing demands, but they also help simulate the experience of individuals with hearing aids, who commonly report listening to be cognitive effortful even at high levels of perceptual accuracy (Hicks & Tharpe, 2002; Lunner et al., 2016; McGarrigle et al., 2014; Pichora-Fuller et al., 2016).

In a previous neuroimaging experiment, we found a set of frontal and temporal regions exhibiting intelligibility-dependent responses when participants attended to speech, but not when they attended to either visual or auditory distractor tasks (Wild et al., 2012). In that study, the clear and degraded speech were not matched on intelligibility, and the difficulty of the distractor tasks was not manipulated. Here, we expand on this design by comparing BOLD responses to intelligibility-matched clear and degraded speech under parametric levels of cognitive load. As a distraction task, presented concurrently with spoken sentences, we used a multiple object tracking (MOT) task (Pylyshyn & Storm, 1988; Scholl, 2009). This is a visuospatial attention task with a well-characterized linear response profile for behavioral (Alvarez & Franconeri, 2007; Bettencourt & Somers, 2009; Pylyshyn & Storm, 1988) and BOLD (Culham et al., 1998, 2001; Howe et al., 2009; Bettencourt, 2010) performance indices. We parametrically manipulated the cognitive demands of the MOT task by changing the number of dots to be simultaneously tracked, predicting a specific linear load profile. This design allowed us to distinguish regional involvement when speech intelligibility responses depend on the relevance of speech for the current task (*task-dependent* control) from the amount of control that is available to aid perception (*load-dependent* control).

We demonstrate that the focus of attention -- whether individuals were listening to speech or tracking dots -- had a strikingly different effects on neural response to clear and degraded speech in high-level speech regions. Whereas the responses in anterior insulae were consistent with domain-general performance monitoring, anterior temporal cortex was selectively recruited for speech perception, with a strikingly different response profile for clear and intelligibility-match degraded speech under parametric cognitive load. These results reveal the division of labor within a classical fronto-temporal speech network, where cognitive control enhances speech perception in challenging listening conditions.

Results

Task overview

At the beginning of each trial of the fMRI session, participants were instructed to perform one of two tasks (see Figure 1). During LISTEN trials, they reported whether they understood the 'gist' of a sentence that was either not degraded (Clear), degraded but as intelligible as clear speech (NV12), or degraded below the intelligibility of clear speech (NV6; still over 90% intelligible). On a subset of LISTEN trials participants instead heard unintelligible speech (Rotated), a spectrotemporal-matched acoustic baseline for speech.

During TRACK trials, participants performed MOT, tracking either 1, 3, 4, or 6 pseudorandomly moving dots among 12 distractors, and then reporting which out of three highlighted dots had been a member of the tracked set (33% chance rate). Participants saw different numbers of moving dots during LISTEN trials and heard different kinds of speech during TRACK trials, in a fully crossed factorial design consisting of Task (2 levels), Speech Type (3 levels), and number of dots (4 levels). After the scanning session, participants performed a recognition memory test for sentences presented during the scan session (visually presented one at a time), as a secondary measure of their speech comprehension.

Task Performance

During LISTEN trials, participants reported whether they understood the gist of each sentence. Participants reported understanding almost all of the intelligible speech trials (Clear: 98.1%; NV12: 97.9%; NV6: 93.8%), and almost none of the Rotated trials (5.3%). These scores were similar to the word-report accuracy collected from a separate group of pilot participants (all $BF_{10} \leq 0.5$; see Figure

2A). Gist scores differed among intelligible speech types ($F_{(1.26, 26.6)} = 12.1$, $p < .001$, $\eta^2 = .365$). Whereas Clear and NV12 did not differ ($p_{Holm} = .648$, $BF_{10} = 0.246$), gist scores were higher for both Clear and NV12 compared to NV6 (Clear: $t_{(21)} = 3.44$, $p_{Holm} = .005$; NV12: $t_{(21)} = 3.98$, $p_{Holm} = .002$).

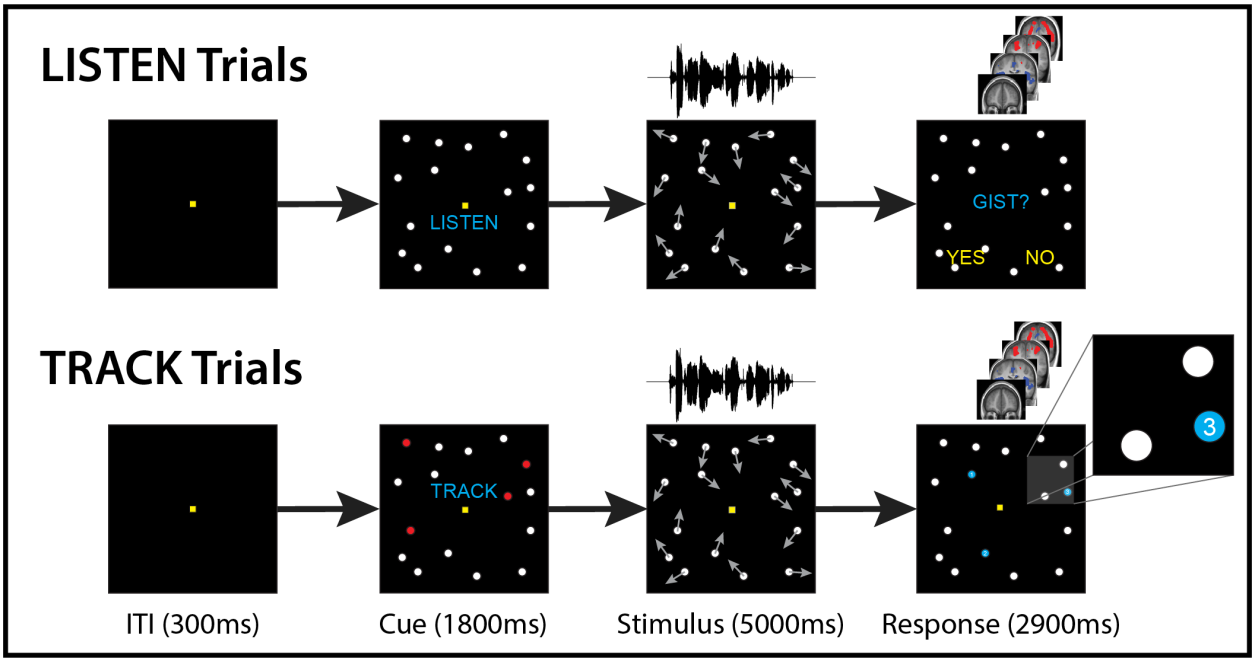

Figure 1. *Trial Timecourse.* At the beginning of each trial, participants were first cued to focus on speech (LISTEN) or focus on tracking (TRACK). They then both heard speech and saw moving dots, making a response during the whole-brain fMRI acquisition (occurring 4 seconds after stimulus midpoint). Speech stimuli were ordinary sentences (e.g., 'Her handwriting was very difficult to read') that were either clear (undistorted), 12-band noise-vocoded, or 6-band noise-vocoded, and during LISTEN trials participants reported whether they understood the 'gist' of each sentence. During tracking, participants tracked 1, 3, 4, or 6 moving dots among 12 distractors, and then reported which queried dot had been a member of the tracked set.

During TRACK trials, participants tracked 1, 3, 4, or 6 moving dots and then selected the member of the tracked set with a three-alternative forced choice. Participants' tracking accuracy linearly decreased as load increased (logistic mixed-effects regression: $\beta = -0.44$, $t_{(21)} = -11.0$, $p < .001$), from 94% accuracy for 1 dot to 60% accuracy for 6 dots. Participants consistently performed above chance (33%), even at the highest level of Load (6 dots: $t_{(21)} = 11.1$, $p < .001$; see Figure 2B).

Sensitivity scores (d') indexing how well participants could distinguish sentences heard during the experiment from foils during the post-scan recognition memory test are depicted in Figure 2C. Sentences from all conditions were remembered better than chance (one-sample t-tests against 0: all $p_{Holm} < .001$). We ran a 3 (Speech) × 2 (Task) repeated-measures ANOVA on d' scores,

finding that participants remembered sentences better during LISTEN than TRACK trials ($F_{(1, 23)}$ = 81.0, $p < .001$, $\eta^2 = .779$).

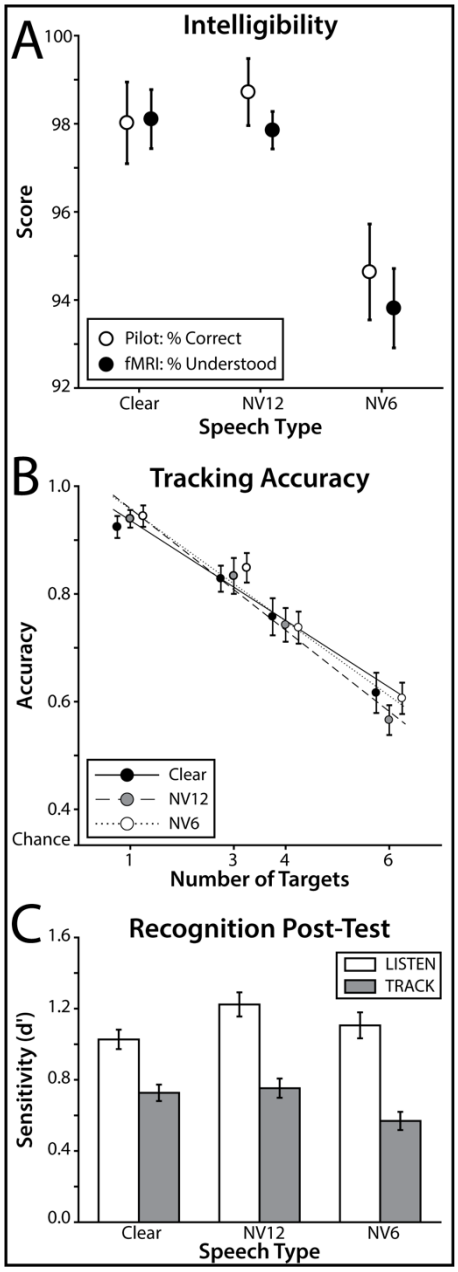

Figure 2. *Behavioral Results*. **A**: Intelligibility. Intelligibility scores across speech types were similar whether measured as objective word report accuracy (behavioral pilot; n = 12) or as subjective gist report (scanner experiment; n = 22). **B**: Tracking Accuracy. When participants tracked more targets, their tracking accuracy declined. Participants' accuracy remained above chance (33%) at all levels of tracking load. **C**: Recognition Post-Test. After the main experiment, participants performed a surprise memory test for the speech stimuli, deciding whether written probes had been heard previously or were novel. Memory sensitivity was quantified with d', comparing hit and false alarm rates. All error bars indicate within-participant SEM (Morey, 2008).

Recognition memory also depended on Speech type ($F_{(1.94, 44.7)}$ = 5.64, p = .007, $\eta^2 = .197$): memory for NV12 speech was significantly better than for NV6 ($t_{(23)}$ = 3.17, p_{Holm} = .008) and marginally better than for Clear speech ($t_{(23)}$ = 2.26, p = .057). The interaction between Task and Speech Type was only marginally significant ($F_{(1.99, 46.0)}$ = 2.84, p = .069, $\eta 2$ = .110), suggesting that memory for NV12 speech may have benefited from the LISTEN task more than Clear speech ($t_{(23)}$ = 2.31, p_{Holm} = .076). This interaction matches our previous observations of stronger memory performance for degraded than clear speech when these are both attended (Wild et al., 2012). These memory results suggest that performance of the MOT task disrupted speech processing, and that attention to mildly degraded sentences enhances processing.

Task-specific neural responses

Participants appeared to orient their attention depending on the task cue (Figure 3). Consistent with previous studies, LISTEN trials elicited greater activity across temporal and lateral prefrontal cortices (Davis & Johnsrude, 2003; Scott et al., 2000), whereas TRACK trials elicited greater activity in posterior parietal and superior frontal cortices (Culham et al., 2001; Howe et al., 2009).

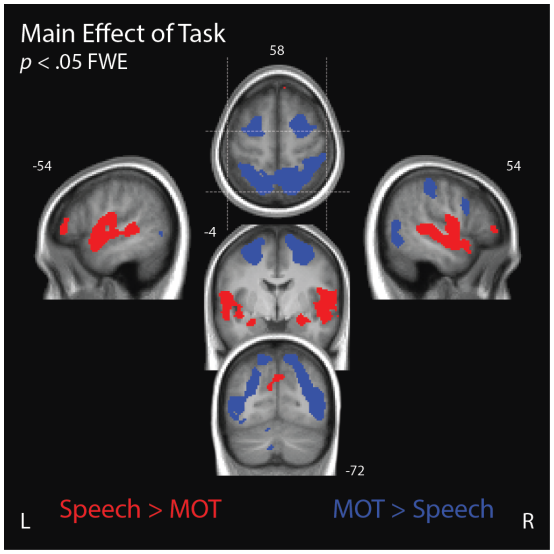

Figure 3. *Main effect of Task*. Voxels that exhibited a significant main effect of Task were colored according to whether they exhibited a greater response to LISTEN than TRACK, or vice versa ($p < .05$, whole-brain FWE). Activation is plotted on the mean participant T1-weighted structural MR image, with dashed lines on the axial slice indicating the location of the sagittal and coronal slices. See supplementary materials for coordinate table.

We tested the simple main effect of Speech Type during LISTEN trials only, as we hypothesized that speech processing would depend on attention (see Figure 4A). Comparing the activity elicited by Clear, NV12, NV6, and Rotated speech during LISTEN trials, we observed a simple main effect of Speech Type across temporal and cingulo-opercular cortices. Temporal lobe voxels appeared to be sensitive to the intelligibility of speech, exhibiting progressively greater activity as gist report accuracy increased across the four speech types (green voxels; Davis & Johnsrude, 2003; Wild et al., 2012). In contrast, cingulo-opercular voxels exhibited greater activity for NV6 speech than for clear and NV12 speech (blue voxels), consistent with these regions responding more when stimuli are degraded (Eckert et al., 2016; Wild et al., 2012). These hypothesis-driven contrasts were not exhaustive, and some regions showed a main effect of speech with a different pattern of activation (white voxels).

Despite the highly similar intelligibility of Clear and NV12, our neural measures distinguished these speech types. Contrasting Clear vs NV12 during LISTEN revealed a significant peak in the left STG ($F_{(1, 23)} = 80.46$, $p < .001$, whole-brain FWE) and a marginally significant peak in the right STG ($F_{(1, 23)} = 40.91$, $p = .069$). These clusters partially overlapped with intelligibility-sensitive regions. Both STG regions were more sensitive to Clear than to NV12 speech. No voxels exhibited a significantly stronger response to NV12 than to Clear.

Finally, we tested for the simple parametric effect of tracking load during TRACK. In many of the regions that were more active for TRACK than LISTEN (main effect of task), BOLD activity was positively correlated with tracking load (see Figure 4B; green voxels), consistent with previous reports (Bettencourt, 2010; Culham et al., 1998, 2001; Howe et al., 2009; Jovicich et al., 2001;

Tomasi et al., 2004). We also observed negative correlations with tracking load in the left supramarginal gyrus and angular gyri bilaterally (magenta voxels).

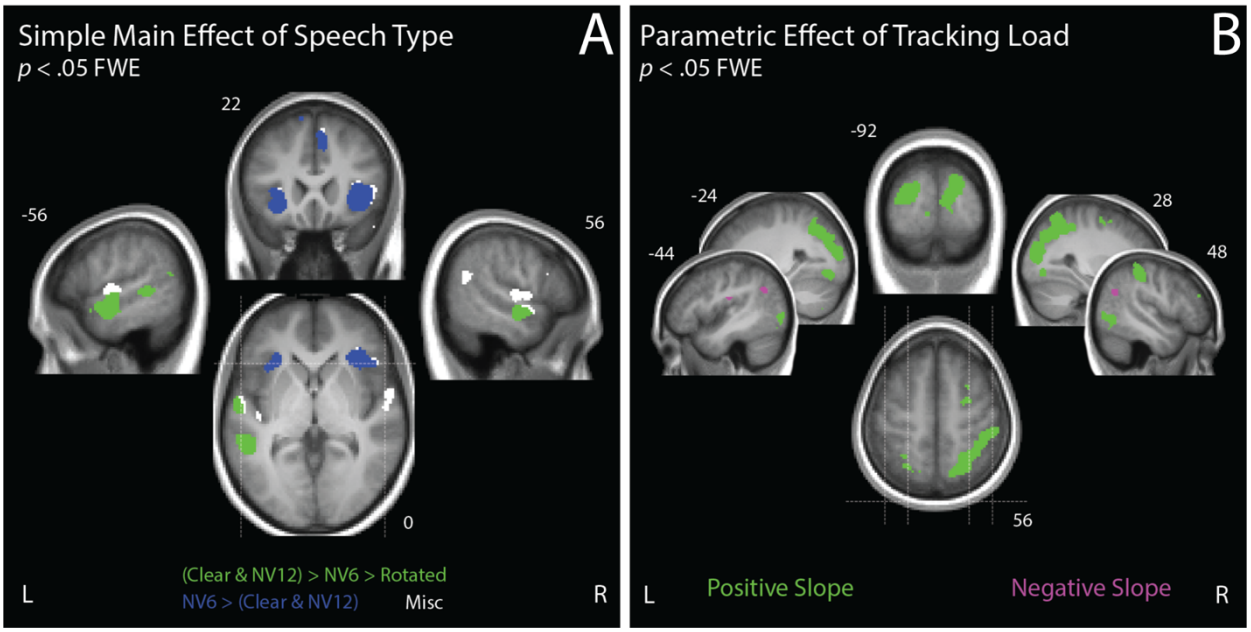

Figure 4. *Task-Specific Simple Main Effects.* **A:** Simple Main Effect of Speech Type. Voxels that exhibited a significant simple main effect of Speech Type (Clear, NV12, NV6, or Rotated) during LISTEN are colored according to hypothesized contrasts (Wild et al., 2012). Green voxels indicate a greater response for more intelligible speech and blue voxels indicate a greater response for NV6 compared to more intelligible Clear and NV12 speech. (White voxels exhibited any simple main effect pattern not captured by these contrasts.) **B:** Parametric Effect of Tracking Load. Voxels that exhibited a significant parametric effect of the number of dots tracked during TRACK are colored green if they show a positive relationship, and magenta if they show a negative relationship. In both images, activation is shown on the mean participant T1-weighted structural MR image, and dashed lines on the axial slice indicate the location of the sagittal and coronal slices. See supplementary materials for coordinate table.

Domain-general response in anterior insulae

Our primary hypotheses concern the degree to which speech processing requires attention under different levels of degradation. Accordingly, we tested our 2- and 3-way interactions within a large speech-sensitive mask based on our previous investigation (Wild et al., 2012), which was fully independent of the current experiment. We defined our mask as voxels exhibiting either a significant main effect of Speech Type or a Speech Type × Task interaction in this previous experiment (see Figures 4 and 5 from Wild et al., 2012).

We observed a significant interaction between Task (LISTEN and TRACK) and Speech Type (Clear, NV12, and NV6) in the anterior insulae bilaterally, consistent with our previous experiment (Wild et al. 2012b; see Figure 5). To compare the response profiles across hemispheres, we ran a Hemisphere × Speech Type × Task mixed ANOVA on the parameter estimates from these regions. The hemisphere factor did not influence our interaction effect (BF_{10} = .201), so we averaged parameter estimates across above-threshold voxels in this region across hemispheres.

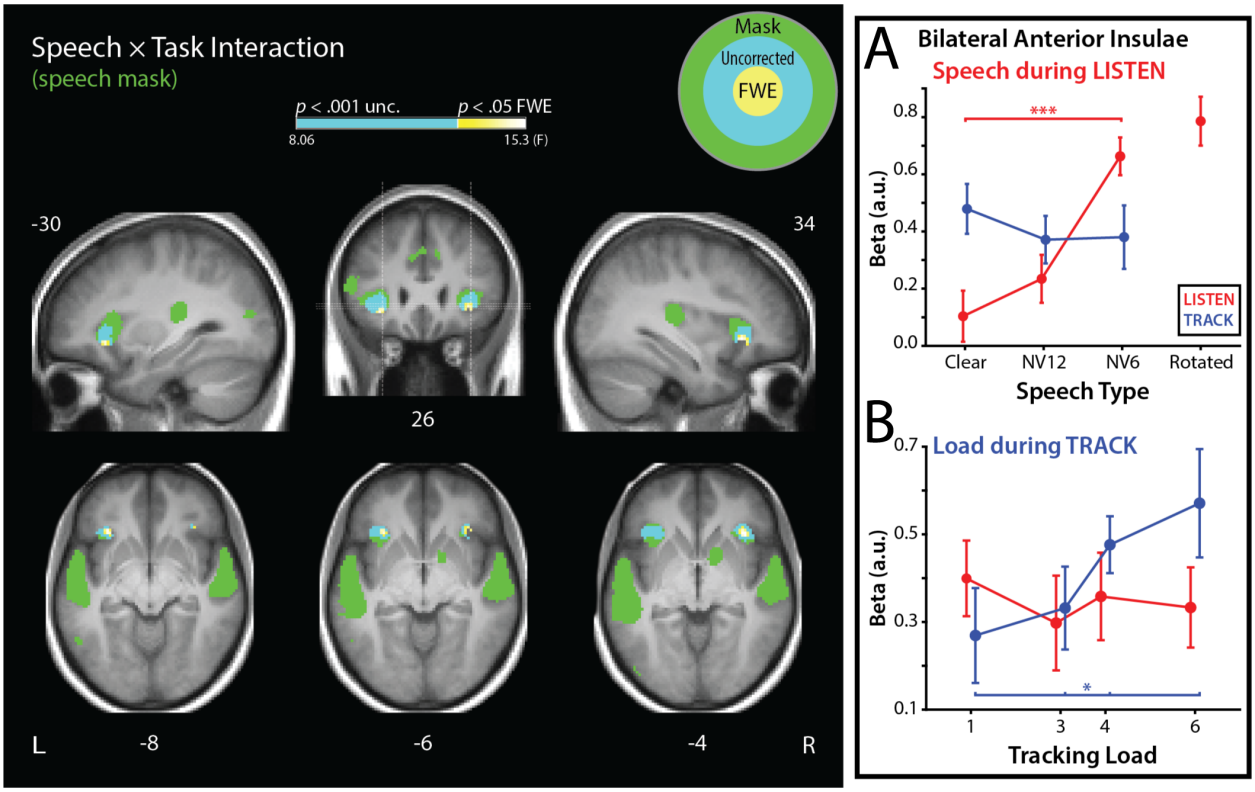

Figure 5. *Speech × Task Interaction.* Analyses were performed within an independent mask of speech-sensitive cortex (green; see text). Cyan voxels exhibited an interaction between Speech Type and Task at an uncorrected threshold (p < .001). Voxels that exhibited a significant interaction at a corrected threshold are indicated with a heat map corresponding to their F-statistic (p < .05, within-mask FWE). **A**: Parameter estimates extracted from above-threshold voxels show a significant simple main effect of Speech Type only during LISTEN (red). **B**: A post-hoc analysis found a significant positive parametric effect of Load only during TRACK (blue). Error bars indicate SEM adjusted for within-subject measurements (Morey, 2008). Activation is plotted on the mean participant T1-weighted structural MR image, and dashed lines on the coronal slice indicate the location of the sagittal and axial slices. See supplementary materials for coordinate table.

In this insular region was a simple main effect of Speech Type during LISTEN ($F_{(1.87, 43.1)}$ = 18.65, p < .001) that was not significant during TRACK ($F_{(1.53, 35.2)}$ = 0.458, p = .585; BF_{10} = .172; see Figure 5A). During LISTEN, the anterior insulae's response was greater for NV6 than Clear speech ($t_{(23)}$ = 5.81, p_{Holm} < .001), and NV12 speech ($t_{(23)}$ = 5.10, p_{Holm} < .001). Activation during LISTEN for Clear and NV12 speech did not differ (p_{Holm} = .229; BF_{10} = .423). This pattern of elevated

activity for difficult-to-understand degraded speech (NV6), only when this speech is task-relevant, is consistent with the response profile observed in (Wild et al., 2012).

To further characterize the task-dependent role of the anterior insulae, we also tested whether the effect of tracking load was evident in these insular voxels (see Figure 5B). We found that the insular response linearly increased with Load during TRACK ($t_{(23)} = 2.22$, $p = .036$), with a stronger Load effect during TRACK than LISTEN ($t_{(23)} = 2.55$, $p = .018$). Together, these signals suggest that the insulae's response reflected the performance of the currently attended task.

Domain-specific response in anterior temporal cortex

Our analysis of primary interest examined whether there are speech-sensitive regions in which the effect Speech Type depends on the load during TRACK trials, and in particular whether this cognitive load dissociates processing of Clear speech from intelligibility-matched degraded speech (NV12). Using the same speech-sensitive mask as our Speech Type × Task analysis, we examined the interaction of Speech Type × Task on the parametric Load modulators (effectively examining the Speech × Task × Load interaction). We found that this interaction was significant in anterior portions of the superior temporal gyri bilaterally (aSTG; see Figure 6). As with the insulae, we found that this interaction was similar across hemispheres ($BF_{10} = .301$), so we averaged the parameter estimates across above-threshold voxels in both hemispheres.

During LISTEN, the effect of Load was not significant, nor was there a Load × Speech Type interaction ($F_{(2, 46)} = 1.38$, $p = .267$, $BF_{10} = .278$). This was expected, since Load predictors during LISTEN only indexed the number of (task-irrelevant) dots on the screen. In contrast, during TRACK, the parametric Load effect depended on Speech Type ($F_{(2, 46)} = 12.13$, $p < .001$; see Figure 6A). The Load effect was apparent for Clear speech, with activity decreasing as load increased beyond 1-item MOT. In contrast, for NV12 and NV6 speech, activity during TRACK was at floor even for 1-item MOT, eliciting a response no stronger than for unintelligible rotated speech ($Load_{Clear} – Load_{NV12}$: $t_{(23)} = -4.041$, $p_{Holm} < .001$; $Load_{Clear} - Load_{NV6}$: $t_{(23)} = -2.92$, $p_{Holm} = .016$). Across all of the Speech conditions in both tasks, only Clear speech during TRACK exhibited a significant effect of Load (Clear during TRACK: $t_{(23)} = -3.20$, $p_{bonferroni} = .024$; all other $p_{uncorrected} \geq .16$ and $BF_{10} \leq .545$).

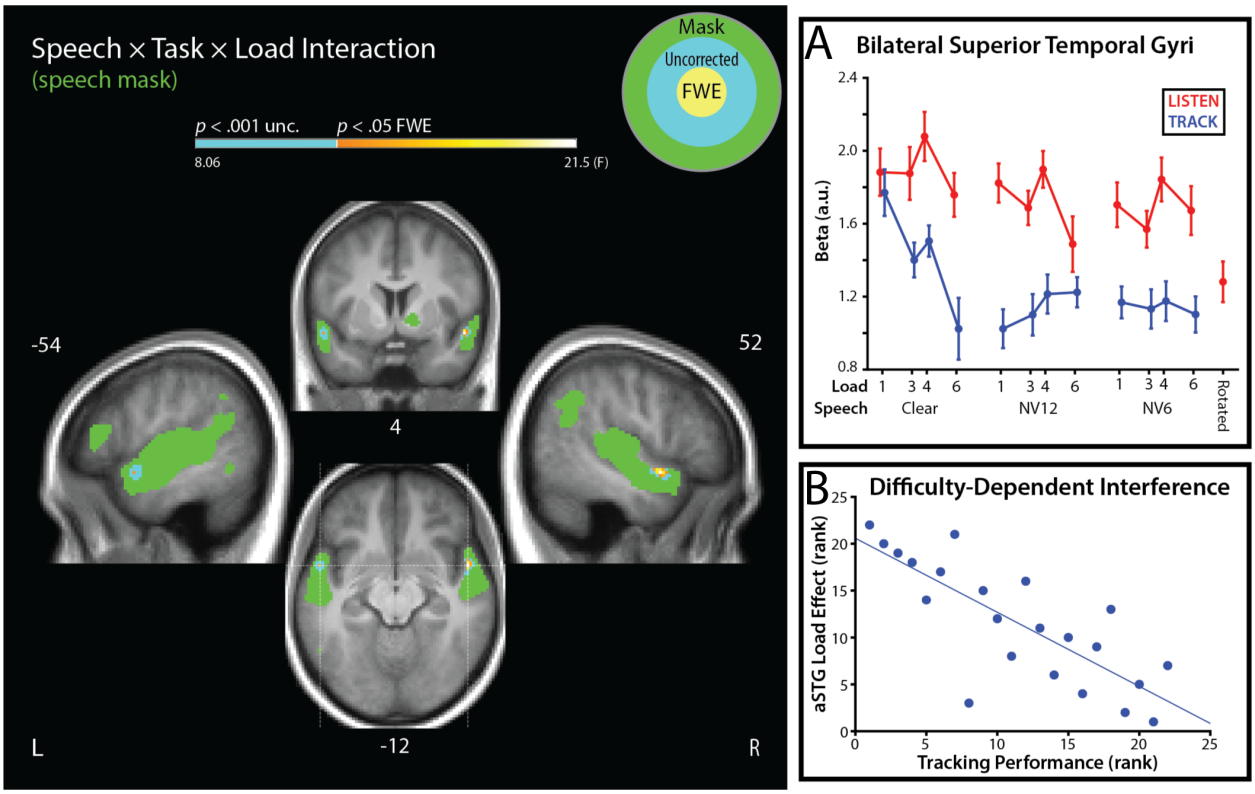

Figure 6. *Speech × Task × Load Interaction.* Analyses were performed within an independent mask of speech-sensitive cortex (green; see text). In cyan voxels, the slope relating BOLD activation to tracking load depended on both Task and Speech Type ($p < .001$, uncorrected). Voxels that exhibited a significant interaction at a corrected threshold are indicated with a heat map corresponding to their F-statistic ($p < .05$, within-mask FWE). **A**: Parameter estimates extracted from above-threshold voxels show a different load response for Clear and degraded speech during TRACK (blue), with degraded speech yielding activation in these regions at floor level (defined by the Rotated-speech point) at all tracking loads. In marked contrast, activity for Clear speech did not depend on task when Load was low (1-item MOT), but then linearly declined with increasing tracking load. **B**: Participants who had more difficulty with the tracking task (lower Accuracy / RT) had a stronger interaction between Speech Type and Load during TRACK. Error bars indicate within-subject SEM (Morey, 2008). Activation is plotted on the mean participant T1-weighted structural MR image, and dashed lines on the coronal slice indicate the location of the sagittal and axial slices. See supplementary materials for coordinate table.

Another way to compare our Speech conditions is to examine, within each Speech Type, the MOT load at which differences between tasks begin to arise. Within each Speech Type, therefore, we compared the response at each level of Load during TRACK to the overall response during LISTEN. When one target was being tracked (lowest load), the STG response for clear speech was similar between TRACK and LISTEN ($t_{(23)} = -1.02$, $p_{\text{uncorrected}} = .32$, $BF_{10} = 0.344$). In marked contrast, activity evoked by degraded speech depended strongly on Task: activity for both NV12 and NV6 was substantially lower during TRACK than LISTEN, even at the weakest level of Load ($NV12_{1\text{-target}}$: $t_{(23)} = -6.07$, $p_{\text{Holm}} < .001$; $NV6_{1\text{-target}}$: $t_{(23)} = -5.76$, $p_{\text{Holm}} < .001$). When tracking three or

more objects, STG activity was always lower for TRACK than LISTEN, and did not differ among speech types (Effect of Speech Type when Load > 1: $BF_{10} = 0.038$).

Complementing our neural measures, we also examined whether individual differences in the strength of this Load by Speech Type interaction was correlated with participants' task performance. We found that participants' with a stronger aSTG Load effect during TRACK ($Load_{(NV12, NV6)}$ - $Load_{Clear}$) had worse average overall tracking accuracy (Spearman's correlation: $\rho_{(20)} = -.46$, $p = .032$) and slower median reaction times ($\rho_{(20)} = .52$, $p = .014$). We validated the generalizability of these individual differences using a leave-one-out cross-validation procedure. A measure of processing efficiency (accuracy / RT) was strongly correlated with aSTG Load effects within-sample ($\rho_{(20)} = -.79$, $p < .001$; see Figure 6B), and regression predictions for held-out participants strongly correlated with their performance ($\rho_{(20)} = .74$, $p < .001$). Participants with stronger neural indicators of load-dependent interference on speech processing performed more poorly on the MOT task, suggesting that our aSTG neural measures reflect the subjective task demands.

In sum, the response to clear speech in anterior temporal cortex was similar regardless of the focus of attention when tracking was easy, but linearly declined to the same low level as for degraded speech with increasing tracking load. This neural index of interference was more severe for participants that were overall worse at the tracking task. The response profile for clear speech was fundamentally different from that for equally intelligible degraded speech, with activity for this degraded speech at the same level as unintelligible Rotated speech, even at weakest level of tracking load.

Discussion

Intelligibility responses in the anterior portion of the ventral speech pathways depend on attention (Eckert et al., 2016; Sabri et al., 2008; Wild et al., 2012). In the current experiment, we found that these regions can be fractionated based on whether speech-sensitivity depends on the current task or the available processing capacity. Activity in the anterior insulae appeared to reflect the demands of the instructed task. This region responded more strongly to more degraded speech only when speech was task-relevant, and activity depended linearly on tracking load only during MOT (Figure 5). In contrast, sensitivity to speech in anterior temporal regions depended both on the type of speech and, for clear speech, on concurrent cognitive demands (Figure 6). This load-dependent

response in bilateral temporal lobes strongly dissociated clear speech from intelligibility-matched degraded speech: clear speech was unaffected by the weakest level of distraction, at which the degraded speech response was already reduced to baseline. These observations functionally parcellate speech-sensitive cortex in the inferior frontal and superior temporal regions based on their relationship to cognitive control, demonstrating substantial costs of distraction under natural, perfectly intelligible, levels of speech degradation.

The anterior insulae play an important role in cognitive control (Bunge et al., 2002; Cieslik et al., 2015; Dosenbach et al., 2006; Duncan & Owen, 2000; Fedorenko et al., 2013; Shenhav et al., 2013), and may support performance monitoring (Lamichhane et al., 2016; Vaden et al., 2013; Wager et al., 2005) and/or orienting towards salient events (Craig & Craig, 2009; Klein et al., 2007; Seeley et al., 2007; Ullsperger et al., 2010). In this experiment, activity in the anterior insulae was sensitive only to the demands of the instructed task: stronger responses to degraded speech only during LISTEN (as in Wild et al., 2012b), and positive linear dependence on tracking load only during TRACK. During LISTEN, this region exhibited a similar response for clear and intelligibility-matched degraded speech, also consistent with a generic role for performance monitoring (Vaden et al., 2013, 2015, 2016).

In anterior temporal cortex, we found that speech sensitivity depends on the cognitive demands of a distracting task. When Clear speech was task-irrelevant, the aSTG response linearly declined as tracking load increased, with a stronger decline predicting poorer tracking performance. This decline may reflect a decreased availability of attention to enhance speech perception or active suppression of this region to reduce interference, with both accounts implying shared capacity for speech perception and MOT (Broadbent, 1958; Kahneman, 1973). MOT is a relatively simple task designed to isolate attentional processes that index object locations (Cavanagh & Alvarez, 2005; Pylyshyn & Storm, 1988; Scholl, 2009), with recent theoretical (Franconeri et al., 2010) and computational (Srivastava & Vul, 2016) models proposing that a critical function of MOT is protecting target indices from interference (i.e., from 'swapping' a target with a distractor; Pylyshyn, 2004). During speech perception, there may be analogous competition between phonological, lexical, and semantic candidates (e.g., multiple potential interpretations of a sound or word), which is exacerbated by degradation (Luce & Pisoni, 1998; Marslen-Wilson, 1987; Miller et al., 1951; Novick et al., 2005; Rodd et al., 2002; Spivey et al., 2005; Thompson-Schill et al., 1997; Zhuang et al., 2011).

During both tasks, attention could plausibly be allocated in response to heightened uncertainty and competition (e.g., towards regions of target/distractor proximity in MOT, or proximal phonological candidates during speech), a core process in domain-general cognitive control (Berlyne, 1957; Miller & Cohen, 2001; Posner & Snyder, 1975).

When attention was on the MOT task the anterior temporal response to (task-irrelevant) intelligible degraded speech was eliminated, which contrasted markedly with the response during task-irrelevant clear speech. This profile may reflect 'maxed-out' processing capacity, or additional functions that are unavailable under distraction (e.g., functions that are goal-dependent). That processing capacity was entirely occupied by the MOT task is not likely, given that the response in anterior temporal regions to mildly degraded speech was at the baseline even when individuals were tracking a single object, which is a very modest level of load. Furthermore, the load effect was clearly evident for task-irrelevant clear speech, but not for degraded speech.

Instead, the processing of perfectly intelligible degraded speech in anterior temporal lobe regions appears to be gated by task goals. Consistent with this idea, activity in anterior insulae was determined by the demands of the attended task, plausibly in the service of top-down control over anterior temporal cortex (Novick et al., 2005; Wild et al., 2012; Eckert et al., 2016). The insulae and anterior temporal lobe share extensive anatomical connections via the uncinate fasciculus and extreme capsule (Kier et al., 2004; Petrides & Pandya, 1988, 2007; Romanski et al., 1999), which have long been thought to facilitate speech perception (Wernicke, 1908). Neuropsychological and neuroimaging evidence supports a role for this network in semantic processing (Dick & Tremblay, 2012; Hickok & Poeppel, 2007; Saur et al., 2008). For example, electrostimulation to extreme capsule fibers in the anterior insulae reliably induce 'semantic paraphasias', with patients replacing target words with semantically-related competitors (e.g., brush → comb; Duffau et al., 2005), a potential complement to the target-distractor swaps that characterize MOT performance (Pylyshyn, 2004; Franconeri et al., 2010; Srivastava & Vul, 2016). While these similarities are promising, further research is needed to fully characterize the neural interactions that support selective attention during speech perception.

Consistent with enhanced top-down control during degraded speech perception, recognition memory tended to be better for NV12 speech than Clear speech when it was the focus

of attention (as in Wild et al., 2012; see also Hirshman & Mulligan, 1991; Nairne, 1988). However, these findings contrast with previous research that has documented poorer memory for degraded speech (Murphy et al., 2000; Pichora-Fuller et al., 1995; Rabbitt, 1966; Surprenant et al., 1999). In many of these previous experiments, stimuli lacked the contextual constraints of full sentences (Murphy et al., 2000; Rabbitt, 1966; Surprenant et al., 1999), suggesting that the use of syntactic or semantic context to enhance speech intelligibility also enhances memory (Novick et al., 2005).

We found that task interference effects were strikingly different between clear and intelligibility-matched degraded speech, supporting an essential role for cognitive control at even the mildest levels of perceptual difficulty. These findings echo reports from individuals with hearing impairments, that sustained perception of (amplified) speech is cognitive fatiguing. Nearly one in four people fitted with hearing aids report rarely using them, and one in five are neutral about, or dissatisfied with, their hearing aids (McCormack & Fortnum, 2013). The "listening effort" that is required to understand speech through hearing aids may be an important reason for this lack of enthusiasm. Our results demonstrate that even minor distractions during perception (i.e., tracking a single target) disrupts processing of mildly degraded speech: this illustrates the need to consider cognitive load when assessing and accommodating listeners with hearing impairment.

Online Methods

Participants

Twenty-six individuals (15 females; M_{age} = 21.5, SD_{age} = 3.86) participated in this experiment after providing informed consent in accordance with the research ethics board at the University of Western Ontario. Participants were right-handed, native English speakers, with self-reported normal (or corrected-to-normal) vision and self-reported normal hearing. Two participants were removed before analysis, due to dislodged earbuds or excessive movement during scanning, leaving 24 participants for the subsequent analyses.

Experimental Design

On every trial, participants both heard a sentence and saw moving dots (See Figure 1). At the beginning of each trial, we instructed participants to either attend to the speech ('LISTEN'), or to perform a visual tracking task ('TRACK'). Across trials, we manipulated which task participants performed (2 levels), the clarity of speech that participants heard (3 levels), and the number of dots

that participants saw on their screen (4 levels), generating 24 factorial conditions. Participants experienced 3 trials from each condition in each of the 3 scanning runs, for a total of 216 experimental trials. Participants also experienced two types of control trial: 24 silent, fixation-only trials and 24 LISTEN trials with rotated NV speech (see below), distributed equally across the three runs. We block-randomized conditions within each scanner run to minimize the effect of low-frequency drift.

Speech Task (LISTEN)

Due to a technical error, the comprehension and tracking data during scanning were lost for 2 participants, leaving 22 participants for behavioral analyses.

We used the same materials used in (Wild et al., 2012): 216 everyday sentences, all recorded by the same female speaker of Canadian English (e.g., 'His handwriting was very difficult to read'). Stimuli were presented diotically via foam-tipped insert earphones (Sensimetrics, Belmont, USA) at a comfortable listening level. The sentences were 6-13 words long; 1.2-4.7 seconds in duration; and were split into six lists that were closely matched on the number of words, the sentence duration, and the summed word frequency (Thorndike and Lorge written frequency). These lists were assigned to the six Speech × Task conditions, counterbalanced across participants.

The clarity of the speech stimuli was manipulated using noise vocoding (Shannon et al., 1995). Each speech signal was filtered into logarithmically spaced frequency bands, with boundaries chosen to be equally spaced along the basilar membrane (Greenwood, 1990). The amplitude envelope within each frequency band was extracted and convolved with white noise that was band-limited to the same frequency range. Previous work has found that intelligibility depends on the number of bands (Davis & Johnsrude, 2003; Shannon et al., 1995). In this experiment, we used highly intelligible noise-vocoded stimuli, filtered with 12 (NV12) and 6 (NV6) bands, as well as Clear (un-manipulated) speech. Piloting and previous experiments have determined that people can accurately report nearly 100% of the words from both Clear and NV12 sentences, whereas word-report of NV6 speech is poorer, but still greater than 90% (see Figure 2A). Unintelligible, spectro-temporally matched, control stimuli were generated by "spectral rotation": during the vocoding process, we permuted the assignment of speech envelopes to their noise envelopes (i.e., randomized over frequency bands;

(Blesser, 1972)).

Volumes were collected using a sparse acquisition protocol (Hall et al., 1999), in which our speech stimuli were presented during the silent period (9 seconds) between scans. The onset of each scan began 4 seconds after the midpoint of each sentence and tracking task, sampling the hemodynamic response near its peak amplitude. On LISTEN trials, participants had 2.8 seconds near the end of the 9-sec silent period to indicate with a yes/no keypress (dominant hand) whether they had understood the gist of the sentence (see Figure 1).

Multiple Object Tracking Task (TRACK)

Between 13 and 18 dots were on the screen throughout every trial, regardless of the task. All dots had a diameter of ~1 degree of visual angle and were shown against a black screen spanning ~ 20 × 20 degrees. Dots were stationary for 1.8 seconds, and then moved pseudorandomly around the screen at an approximate speed of 1.8 degrees per second, with dots repelling 180 degrees away from other dots or the edge of the screen at a 0.5-degree proximity.

On TRACK trials, participants tracked a subset of the moving dots (multiple object tracking, MOT; Pylyshyn & Storm, 1988). On these trials, 1, 3, 4, or 6 target dots were highlighted in red for 1.8 seconds before movement. Participants were instructed to keep their gaze on a fixation cue in the center of the screen and track these dots covertly. After 5 seconds of tracking, the dots froze in place, and three dots (one randomly selected target and two foils) were highlighted in blue and labelled '1', '2', and '3'. Participants had 2.8 seconds to indicate with a 3-alternative keypress which of the numbered dots was a target, without feedback (see Figure 1).

Pre-Training and Memory Post-Test

Prior to the scanning session, participants practiced both the speech and tracking tasks. First, participants were familiarized with NV speech, in order to bring their comprehension performance to asymptote (Davis et al., 2005). Over 24 trials, participants heard a noise-vocoded sentence, indicated whether they had understood the gist of the sentence, and then received feedback by hearing the vocoded sentence again while also reading it on the screen (following the recommendations from Davis et al., 2005, Experiment 3). MOT training proceeded over 24 trials. On the first 12, the number of targets began at 1 and increased (to 3, 4 and 6) after each correct

tracking response, and decreased after each incorrect response. On the latter 12 trials, the number of targets on each trial was randomly selected (from 1, 3, 4 or 6).

After the scanning session, we tested participants on their recognition memory for the sentences they had heard. On each trial, participants saw a written sentence on a computer screen, and indicated with a keypress whether they remembered this sentence from the experiment ('OLD'), or whether it was new ('NEW'). Participants were tested on all 216 sentences from the experiment, along with 108 foil sentences. Foil sentences differed from target sentences in both their topic and their content words. During the scanning session, participants were unaware that memory would be tested, ensuring incidental memory encoding.

fMRI acquisition

Images were acquired on the 3.0T Siemens Prisma MRI system at the University of Western Ontario. T1-weighted structural images were collected at the beginning of each session using a single-shot EPI (FoV: 256mm^2; resolution: 1mm isotropic; slice thickness: 1mm with 50% gap; TE: 2.98ms; TR: 2300ms; flip angle: 9°). T2*-weighted functional volumes were acquired across the whole brain using a 4-factor interleaved multi-band gradient EPI (FoV: 192mm^2; resolution: 2.5mm isotropic; slice thickness: 2.5mm with 10% gap; 52 slices; TE: 30ms; TA: 1000ms; TR: 10sec; flip angle: 70°). Acquisition was transverse-oblique, angled away from the eyes.

fMRI preprocessing and analysis

fMRI data were preprocessed and analyzed using SPM12 (Wellcome Centre for Neuroimaging, London, UK), following standard preprocessing steps including realignment, coregistration, and simultaneous segmentation and normalization to MNI (ICBM452) space. Normalization parameters were calculated from the structural image and applied to functional images coregistered to the mean of each run, resampling the images at 2mm^3. The normalized images were spatially smoothed using a 3D Gaussian kernel with an 8mm FWHM.

Statistical parametric maps for each subject were estimated using a general linear model containing onset indicators for rotated speech and the six combinations of Speech (Clear, NV6, and NV12) by Task (LISTEN and TRACK) conditions. The model also included Load parametric modulators for the six speech × task conditions, based on the dots on the screen. For LISTEN

trials, the parametric modulators only captured the number of dots on the screen (c.f., visual load), whereas for TRACK trials, these modulators also captured the effect of tracking load. These models also included run-specific modulators including the six spatial realignment parameters, as well as a run intercept and linear trend. Modulators were mean centered and not orthogonalized (allowing control modulators to compete for variance with task modulators). Due to the long TR (10 seconds; 9 second silent gap between successive scans) in our sparse acquisition design, we modelled trial activation using a finite-impulse response model without serial autocorrelations. Contrast maps for main effects and interactions were calculated at the subject level and tested against zero at the group level using a factorial partitioned-error repeated-measures ANOVA (Henson & Penny, 2003).

We analyzed participants' behavior using custom MATLAB (R2018a) scripts and JASP (0.8.3) for ANOVA and Bayesian analyses (using the default Cauchy prior). Note that Bayes factors (BF_{10}) less than 1/3 provide moderate evidence supporting the null hypothesis (e.g., that two groups are the same; see Jarosz & Wiley, 2014). Follow-up fMRI analyses were performed using MATLAB and JASP. For our follow-up interaction analyses, we utilized a second general linear model that included all twenty-four Speech × Task × Load conditions, along with our run-specific nuisance terms (see above). We followed-up omnibus ANOVAs with post-hoc t-tests, correcting for multiple comparisons with the Holm procedure for sequential tests. Brain-behavior relationships were cross-validated by fitting a linear regression model to predict BOLD contrasts from behavior while holding-out one participant at a time, using this model to predict each held-out participant's BOLD contrast from their behavior, and then correlating the predicted and observed BOLD contrasts.

Data and Code Availability

All data and code are available upon request.

Author contributions

H.R.: data collection; H.R & I.J.: Conception, design, interpretation, writing; H.R. & C.W.: data analysis

Competing Interests

None.

References

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, *7*(13), 14.1–10.

Berlyne, D. E. (1957). Uncertainty and conflict: a point of contact between information-theory and behavior-theory concepts. *Psychological Review*, *64*(6p1), 329.

Bettencourt, K. (2010). *Functional MRI and behavioral investigations of capacity limits in human visual attention* (D. Somers (ed.)) [3411712, Boston University]. https://search.proquest.com/docview/577653442?accountid=9758

Bettencourt, K. C., & Somers, D. C. (2009). Effects of target enhancement and distractor suppression on multiple object tracking capacity. *Journal of Vision*, *9*(7), 9.

Blesser, B. (1972). Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *Journal of Speech, Language, and Hearing Research: JSLHR*, *15*(1), 5–41.

Broadbent, D. E. (1958). *Perception and communication*. Elsevier.

Bunge, S. A., Hazeltine, E., Scanlon, M. D., Rosen, A. C., & Gabrieli, J. D. E. (2002). Dissociable contributions of prefrontal and parietal cortices to response selection. *NeuroImage*, *17*(3), 1562–1571.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, *9*(7), 349–354.

Cieslik, E. C., Mueller, V. I., Eickhoff, C. R., Langner, R., & Eickhoff, S. B. (2015). Three key regions for supervisory attentional control: evidence from neuroimaging meta-analyses. *Neuroscience and Biobehavioral Reviews*, *48*, 22–34.

Craig, A. D., & Craig, A. D. (2009). How do you feel--now? The anterior insula and human awareness. *Nature Reviews. Neuroscience*, *10*(1).

Culham, J. C., Brandt, S. A., Cavanagh, P., Kanwisher, N. G., Dale, A. M., & Tootell, R. B. H.

(1998). Cortical fMRI activation produced by attentive tracking of moving targets. *Journal of Neurophysiology*, *80*(5), 2657–2670.

Culham, J. C., Cavanagh, P., & Kanwisher, N. G. (2001). Attention response functions: characterizing brain areas using fMRI activation during parametric variations of attentional load. *Neuron*, *32*(4), 737–745.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, *23*(8), 3423–3431.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology. General*, *134*(2), 222.

Dick, A. S., & Tremblay, P. (2012). Beyond the arcuate fasciculus: consensus and controversy in the connectional anatomy of language. *Brain: A Journal of Neurology*, *135*(12), 3529–3550.

Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., Burgund, E. D., Grimes, A. L., Schlaggar, B. L., & Petersen, S. E. (2006). A core system for the implementation of task sets. *Neuron*, *50*(5), 799–812.

Downs, D. W. (1982). Effects of hearing aid use on speech discrimination and listening effort. *The Journal of Speech and Hearing Disorders*, *47*(2), 189–193.

Duffau, H., Gatignol, P., Mandonnet, E., Peruzzi, P., Tzourio-Mazoyer, N., & Capelle, L. (2005). New insights into the anatomo-functional connectivity of the semantic system: a study using cortico-subcortical electrostimulations. *Brain: A Journal of Neurology*, *128*(4), 797–810.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*(10), 475–483.

Eckert, M. A., Teubner-Rhodes, S., & Vaden, K. I., Jr. (2016). Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. *Ear and Hearing*, *37*

Suppl 1(Suppl 1), 101S – 10S.

Evans, S., & Davis, M. H. (2015). Hierarchical Organization of Auditory and Motor Representations in Speech Perception: Evidence from Searchlight Similarity Analysis. *Cerebral Cortex* , *25*(12), 4772–4788.

Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology*, *5*, 335–335.

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 201315235.

Feuerstein, J. F. (1992). Monaural versus binaural hearing: ease of listening, word recognition, and attentional effort. *Ear and Hearing*, *13*(2), 80–86.

Franconeri, S. L., Jonathan, S. V., & Scimeca, J. M. (2010). Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity. *Psychological Science*, *21*(7), 920–925.

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research: JSLHR*, *53*(1), 18–33.

Gagne, J.-P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing*, *21*, 2331216516687287.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America*, *87*(6), 2592–2605.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., Gurney, E. M., & Bowtell, R. W. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping*, *7*(3), 213–223.

Heald, S., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, *8*, 35.

Henson, R. N. A., & Penny, W. D. (2003). ANOVAs and SPM. *Wellcome Department of Imaging Neuroscience, London, UK.*

Herrmann, B., & Johnsrude, I. S. (2018). Neural signatures of the processing of temporal patterns in sound. *Journal of Neuroscience*, *38*(24), 5466–5477.

Herrmann, B., Maess, B., & Johnsrude, I. S. (2018). Aging Affects Adaptation to Sound-Level Statistics in Human Auditory Cortex. *Journal of Neuroscience*, *38*(8), 1989–1999.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393.

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research: JSLHR*, *45*(3), 573–584.

Hirshman, E., & Mulligan, N. (1991). Perceptual interference improves explicit memory but does not enhance data-driven processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *17*(3), 507.

Howe, P. D., Horowitz, T. S., Morocz, I. A., Wolfe, J., & Livingstone, M. S. (2009). Using fMRI to distinguish components of the multiple object tracking task. *Journal of Vision*, *9*(4), 10–10.

Hunter, C. R., & Pisoni, D. B. (2018). Extrinsic Cognitive Load Impairs Spoken Word Recognition in High- and Low-Predictability Sentences. *Ear and Hearing*, *39*(2), 378–389.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

Johnsrude, I. S., & Rodd, J. M. (2016). Chapter 40 - Factors That Increase Processing Demands When Listening to Speech. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 491–502). Academic Press.

Jovicich, J., Peters, R. J., Koch, C., Braun, J., Chang, L., & Ernst, T. (2001). Brain areas specific for attentional load in a motion-tracking task. *Journal of Cognitive Neuroscience*, *13*(8), 1048–1058.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Citeseer.

Kier, E. L., Staib, L. H., Davis, L. M., & Bronen, R. A. (2004). MR imaging of the temporal stem: anatomic dissection tractography of the uncinate fasciculus, inferior occipitofrontal fasciculus, and Meyer's loop of the optic radiation. *AJNR. American Journal of Neuroradiology*, *25*(5), 677–691.

Klein, T. A., Endrass, T., Kathmann, N., Neumann, J., von Cramon, D. Y., & Ullsperger, M. (2007). Neural correlates of error awareness. *NeuroImage*, *34*(4), 1774–1781.

Lamichhane, B., Adhikari, B. M., & Dhamala, M. (2016). The activity in the anterior insulae is modulated by perceptual decision-making difficulty. *Neuroscience*, *327*, 79–94.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1.

Lunner, T., Rudner, M., Rosenbom, T., Ågren, J., & Ng, E. H. N. (2016). Using speech recall in hearing aid fitting and outcome evaluation under ecological test conditions. *Ear and Hearing*, *37*, 145S – 154S.

Macdonald, J. S. P., & Lavie, N. (2011). Visual perceptual load induces inattentional deafness. *Attention, Perception & Psychophysics*, *73*(6), 1780–1789.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71–102.

McCormack, A., & Fortnum, H. (2013). Why do people fitted with hearing aids not wear them? *International Journal of Audiology*, *52*(5), 360–368.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: what exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper." *International Journal of Audiology*, *53*(7), 433–440.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*(5), 329.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau. *Tutorials in Quantitative Methods for Psychology , 4*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.665.4286

Murphy, D. R., Craik, F. I. M., Li, K. Z. H., & Schneider, B. A. (2000). Comparing the effects of aging and background noise on short-term memory performance. *Psychology and Aging*, *15*(2), 323.

Nairne, J. S. (1988). The mnemonic value of perceptual identification. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *14*(2), 248.

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective & Behavioral Neuroscience*, *5*(3), 263–281.

Petrides, M., & Pandya, D. N. (1988). Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey. *The Journal of Comparative Neurology*, *273*(1), 52–66.

Petrides, M., & Pandya, D. N. (2007). Efferent association pathways from the rostral prefrontal cortex in the macaque monkey. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(43), 11573–11586.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., & Mackersie, C. L. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*,

37, 5S – 27S.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608.

Posner, M., & Snyder, C. (1975). *Attention and cognitive control.*

Pylyshyn, Z. (2004). Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities. *Visual Cognition*, *11*(7), 801–822.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197.

Rabbitt, P. (1966). Recognition: Memory for words correctly heard in noise. In *Psychonomic Science* (Vol. 6, Issue 8, pp. 383–384). https://doi.org/10.3758/bf03330948

Raveh, D., & Lavie, N. (2015). Load-induced inattentional deafness. *Attention, Perception & Psychophysics*, *77*(2), 483–492.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245–266.

Romanski, L. M., Bates, J. F., & Goldman-Rakic, P. S. (1999). Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. *The Journal of Comparative Neurology*, *403*(2), 141–157.

Rouault, M., & Koechlin, E. (2018). Prefrontal function and cognitive control: from action to language. *Current Opinion in Behavioral Sciences*, *21*, 106–111.

Sabri, M., Binder, J. R., Desai, R., Medler, D. A., Leitl, M. D., & Liebenthal, E. (2008). Attentional and linguistic interactions in speech perception. *NeuroImage*, *39*(3), 1444–1456.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., & Abel, S. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences*, nas. 0805234105.

Scholl, B. J. (2009). What have we learned about attention from multiple object tracking (and vice versa). *Computation, Cognition, and Pylyshyn*, 49–78.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain: A Journal of Neurology*, *123*(12), 2400–2406.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., & Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, *27*(9), 2349–2356.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398.

Srivastava, N., & Vul, E. (2016). Attention Modulates Spatial Precision in Multiple-Object Tracking. *Topics in Cognitive Science*, *8*(1), 335–348.

Surprenant, A. M., Neath, I., & LeCompte, D. C. (1999). Irrelevant speech, phonological similarity, and presentation modality. *Memory* , *7*(4), 405–420.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(26), 14792–14797.

Tomasi, D., Ernst, T., Caparelli, E. C., & Chang, L. (2004). Practice-induced changes of brain function during visual attention: a parametric fMRI study at 4 Tesla. *NeuroImage*, *23*(4), 1414–1421.

Ullsperger, M., Harsay, H. A., Wessel, J. R., & Ridderinkhof, K. R. (2010). Conscious perception of

errors and its relation to the anterior insula. *Brain Structure & Function*, *214*(5-6), 629–643.

Vaden, K. I., Jr, Kuchinsky, S. E., Ahlstrom, J. B., Teubner-Rhodes, S. E., Dubno, J. R., & Eckert, M. A. (2016). Cingulo-opercular function during word recognition in noise for older adults with hearing loss. *Experimental Aging Research*, *42*(1), 67–82.

Vaden, K. I., Kuchinsky, S. E., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2015). Cortical activity predicts which older adults recognize speech in noise and when. *Journal of Neuroscience*, *35*(9), 3929–3937.

Vaden, K. I., Kuchinsky, S. E., Cute, S. L., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2013). The cingulo-opercular network provides word-recognition benefit. *Journal of Neuroscience*, *33*(48), 18979–18986.

Wager, T. D., Sylvester, C.-Y. C., Lacey, S. C., Nee, D. E., Franklin, M., & Jonides, J. (2005). Common and unique components of response inhibition revealed by fMRI. *NeuroImage*, *27*(2), 323–340.

Wernicke, C. (1908). Diseases of the nervous system. *New York: Appleton*, 265–324.

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: the processing of degraded speech depends critically on attention. *Journal of Neuroscience*, *32*(40), 14010–14021.

Zhuang, J., Randall, B., Stamatakis, E. A., Marslen-Wilson, W. D., & Tyler, L. K. (2011). The interaction of lexical semantics and cohort competition in spoken word recognition: an fMRI study. *Journal of Cognitive Neuroscience*, *23*(12), 3778–3790.