Cueing listeners to attend to a target talker progressively improves word report as the duration of the cue-target interval lengthens to 2000 ms

Emma Holmes[a], Padraig T. Kitterick[b,c], and A. Quentin Summerfield[a,d]

[a]Department of Psychology, University of York, UK;

[b]NIHR Nottingham Hearing Biomedical Research Unit, UK;

[c]Division of Clinical Neuroscience, School of Medicine, University of Nottingham, UK;

[d]Hull York-Medical School, University of York, UK

Correspondence should be addressed to Emma Holmes at eholme5@uwo.ca or the following postal address: The Brain and Mind Institute, Natural Sciences Centre, Room 120, Western University, London, Ontario, N6A 5B7, Canada

Abstract

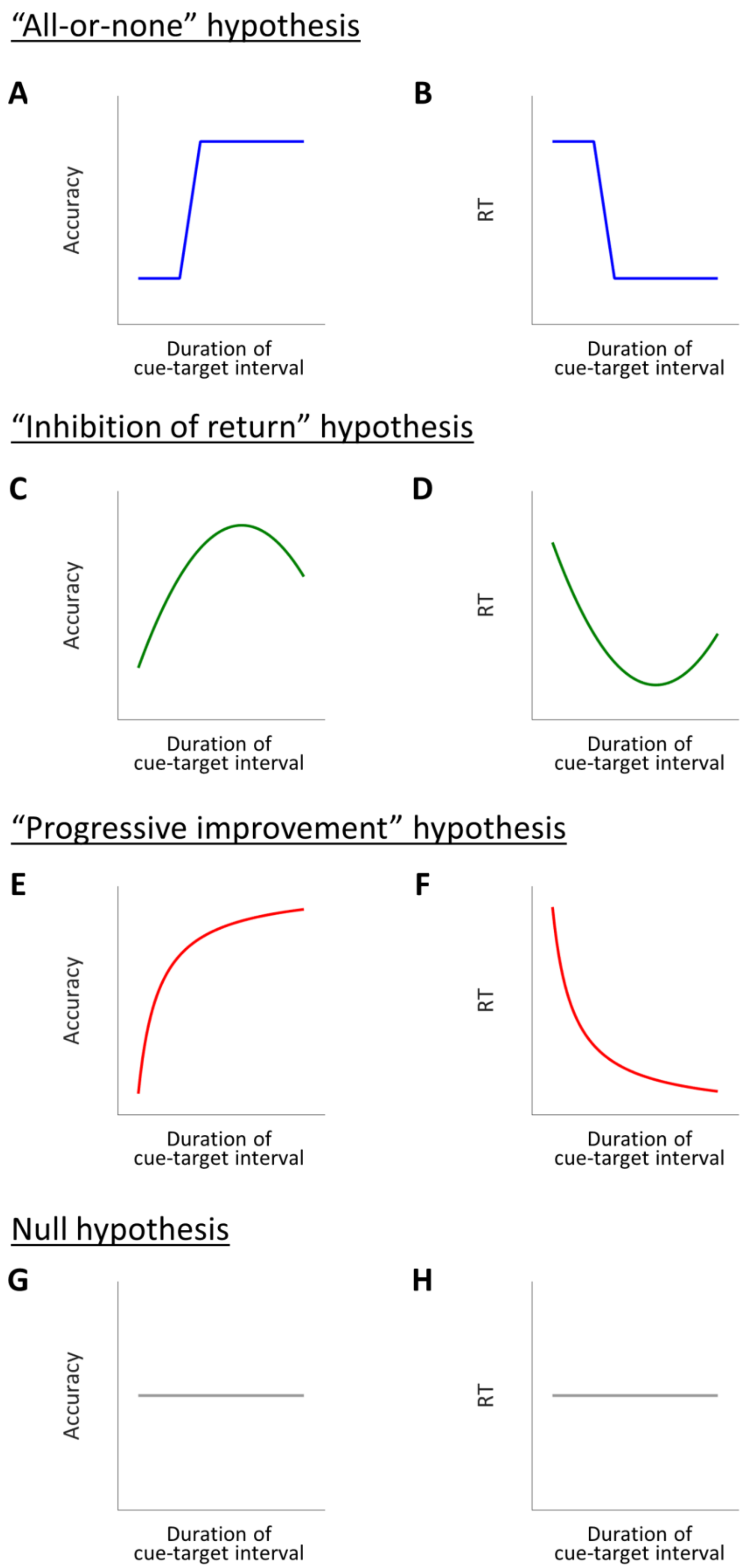
Endogenous attention is typically studied by presenting instructive cues in advance of a target stimulus array. For endogenous visual attention, task performance improves as the duration of the cue-target interval increases up to 800 ms. Less is known about how endogenous *auditory* attention unfolds over time or the mechanisms by which an instructive cue presented in advance of an auditory array improves performance. The current experiment used five cue-target intervals (0, 250, 500, 1000, and 2000 ms) to compare four hypotheses for how preparatory attention develops over time in a multi-talker listening task. Young adults were cued to attend to a target talker who spoke in a mixture of three talkers. Visual cues indicated the target talker's spatial location or their gender. Participants directed attention to location and gender simultaneously ('objects') at all cue-target intervals. Participants were consistently faster and more accurate at reporting words spoken by the target talker when the cue-target interval was 2000 ms than 0 ms. In addition, the latency of correct responses progressively shortened as the duration of the cue-target interval increased from 0 to 2000 ms. These findings suggest that the mechanisms involved in preparatory auditory attention develop gradually over time, taking at least 2000 ms to reach optimal configuration, yet providing cumulative improvements in speech intelligibility as the duration of the cue-target interval increases from 0 to 2000 ms. These results demonstrate an improvement in performance for cue-target intervals longer than those that have been reported previously in the visual or auditory modalities.

Cueing listeners to attend to a target talker progressively improves word report as the duration of the cue-target interval lengthens to 2000 ms

The ability to direct attention selectively to a stimulus of interest in complex visual and acoustic environments is essential for performing a variety of tasks; for example, searching visually for an item in a cluttered room, reading the words contained in a paragraph of text, or understanding a friend speak at a noisy party (for a review, see Duncan, 2006). Experiments demonstrating the ability to direct endogenous selective attention have typically presented a cue that instructs participants to attend to a stimulus defined by a particular characteristic (or combination of characteristics) in a target array. Participants perform better on selective attention tasks when they are cued to attributes of visual or acoustic target stimuli before the target is presented than when no cue is presented or when the cue and target are revealed simultaneously (e.g. Koch, Lawo, Fels, & Vorländer, 2011; Lu et al., 2009; Richards & Neff, 2004). Although previous experiments have examined the time-course of preparatory attention for visual target stimuli, we do not fully understand the mechanisms involved in preparation for auditory stimuli and their sensitivity to the duration of the cue-target interval. The current experiment aimed to improve understanding of this mechanism by systematically investigating how accuracy, reaction times (RTs), and errors in a multi-talker listening task are affected by the duration of the cue-target interval.

With respect to the expected effects of increasing the duration of the cue-target interval on the accuracy and latency of speech intelligibility, there are at least four possibilities: (1) the duration of the cue-target interval does not improve intelligibility until it reaches a criterion duration, beyond which longer intervals do not improve intelligibility further; evidence in favour of this hypothesis would suggest that the mechanism underlying preparatory attention develops over the criterion length of time, but does not improve intelligibility until it reaches a fully prepared state, which is maintained at longer intervals (“All-or-none” hypothesis); (2) there is an optimal cue-target interval, after which further increases in the duration of the cue-target interval lead to worse intelligibility; evidence for this hypothesis would suggest that the mechanism underlying preparatory attention requires a specific amount of time to reach an optimal configuration, then returns to baseline (“Inhibition of return” hypothesis); (3) longer cue-target intervals continue to improve intelligibility progressively as the duration of the cue-target interval increases to the longest interval tested; evidence for this hypothesis would suggest that the mechanism underlying preparatory attention develops gradually, with some improvement in intelligibility gained when the mechanism is in a partially prepared state and, if a threshold or optimum time exists, it is longer than the intervals tested (“Progressive improvement” hypothesis); or (4) the duration of the cue-target interval has no effect on intelligibility, suggesting that either preparatory attention is not engaged or that the intervals tested are shorter than the threshold duration of time required for preparatory attention to affect task performance (null hypothesis). Fig. 1 illustrates predictions for the relationships between the length of the cue-target interval and accuracy and RTs under the four hypotheses.

Fig. 1 [next page]. Cartoon illustrations of four possible hypotheses. (**A,B**) Accuracy and RTs are unaffected by the duration of the cue-target interval. (**C,D**) The duration of the cue-target interval does not improve accuracy or decrease RTs until the interval reaches a threshold of time, beyond which longer intervals do not improve accuracy or RTs further. (**E,F**) There is an optimal cue-target interval, after which further increases in the duration of the cue-target interval lead to worse accuracy and slower RTs. (**G,H**) Longer cue-target intervals continue to improve accuracy and shorten RTs progressively as the duration of the cue-target interval increases.



"All-or-none" hypothesis
A
Accuracy
Duration of cue-target interval
B
RT
Duration of cue-target interval
"Inhibition of return" hypothesis
C
Accuracy
Duration of cue-target interval
D
RT
Duration of cue-target interval
"Progressive improvement" hypothesis
E
Accuracy
Duration of cue-target interval
F
RT
Duration of cue-target interval
Null hypothesis
G
Accuracy
Duration of cue-target interval
H
RT
Duration of cue-target interval

"Progressive improvement" hypothesis

In the visual modality, there is evidence for a progressive improvement in detection and discrimination performance, at least for intervals up to 800 ms. For example, Lu *et al.* (2009) asked participants to discriminate the orientation of a Gabor patch presented at a target location, when three other Gabor patches were presented at different locations. Before the target array was revealed, a centrally-presented arrow cue indicated the location of the target stimulus. Lu *et al.* measured participants' contrast thresholds at cue-target intervals between 0 and 240 ms. Contrast thresholds were better when the cue was presented 240 ms before the target array than when the cue was presented at the same time as the target array (i.e. for the 0-ms cue-target interval). A similar pattern of results was reported by Yamaguchi, Tsuchiya, and Kobayashi (1994) using a set of longer cue-target intervals (200, 500, and 800 ms) and a task in which participants had to detect a target asterisk stimulus presented either ipsilateral or contralateral to the direction conveyed by a central arrow cue. They found a significant difference in RTs between all three cue-target intervals, with RTs becoming progressively shorter as the duration of the cue-target interval lengthened. Together, these studies suggest that preparatory attention develops progressively over time, such that increasing the duration of the cue-target interval allows participants to better prepare for, and respond to, visual target stimuli. This finding is consistent with an underlying mechanism by which participants can improve the detection and discrimination of visual target stimuli by partially preparing for a target stimulus, with greater improvement achieved as longer time is available for preparatory attention to develop.

In the auditory modality, it is less clear whether longer cue-target intervals progressively improve task performance. One reason is that the duration of the cue-target interval has differed between experiments: ranging between 100 ms before the target (Koch et al., 2011) to cueing at the beginning of each block (Brungart & Simpson, 2007; Ericson, Brungart, & Brian, 2004; Kitterick, Bailey, & Summerfield, 2010). Nevertheless, studies measuring brain activity during the cue-target interval are consistent with the idea that preparatory attention develops incrementally over time. When participants prepared to detect a low-level target tone, Voisin *et al.* (2006) found an increase in blood-oxygen level-dependent (BOLD) activity in the superior temporal cortex during the cue-target interval, which increased in amplitude progressively over time. Similarly, Holmes et al. (2017) cued participants to attend to a target talker in a mixture of talkers and found that activity measured using electro-encephalography (EEG) increased in amplitude during the cue-target interval. Given that similar increases in brain activity have been observed during the cue-target interval in preparation for visual tasks—and this activity has been linked to improved performance on a visual discrimination task (Giesbrecht, Weissman, Woldorff, & Mangun, 2006)—we might expect to observe a progressive improvement in performance for auditory tasks as the duration of the cue-target interval lengthens, similar to that observed in visual cueing experiments.

“All-or-none” hypothesis

One study by Richards and Neff (2004; Experiment 4) examined how different length cue-target intervals affected performance on an auditory detection task. They varied the amount of time between the offset of an instructive cue (which was a preview of the target) and the onset of a target tone. The target tone was presented at the same time as a multi-tone masker, which contained frequencies in the same range (200–5000 Hz) as the target. Participants had to indicate whether or not the target tone occurred within the mixture. Detection thresholds were better when the preview was presented 50 ms before the target than when it was presented 5 ms before the target. However, participants achieved no further improvement when the interval increased from 50 to 500 ms. This pattern of results is very different to the pattern observed by Yamaguchi *et al.* (1994) for detection of visual target stimuli, which showed improvements in contrast thresholds up to 800 ms. Instead, the results of Richard and Neff imply that a criterion duration is necessary for successful preparation, beyond which increasing the time available for preparation does not affect detection performance. This finding implies that auditory preparatory attention may be all-or-none. In more detail, preparatory attention might require a threshold duration of time (of approximately 50 ms) to develop after an instructive cue is presented to improve auditory detection performance, after which no further improvements are gained as the duration of the cue-target interval is lengthened.

It is possible that the duration of the cue-target interval affects performance differently between the visual and auditory modalities, although there are several other possible explanations for different patterns of performance between the experiment of Richards and Neff (2004) and the aforementioned experiments in the visual modality. First, given that the cue presented by Richards and Neff was a preview of the target tone, it could be argued that at least some of its influence was exogenous, drawing attention to the cued frequency region. In the same experiment, Richards and Neff (2004) found an improvement in thresholds when the masker was cued than when no cue was presented, although thresholds did not differ as a function of the duration of the interval between the masker cue and the target. Given that exogenous and endogenous cues seem to rely on different functional processes (e.g. Jonides, 1981), have different neural substrates (e.g. Corbetta, Patel, & Shulman, 2008), and are affected differently by the duration of the cue-target interval in the visual modality (e.g. Lu et al., 2009), the different patterns of results may reflect different processes underlying exogenous and endogenous attention. Another possible explanation for the difference is that the studies in the visual modality used abstract arrow stimuli as cues, which would require more interpretation and may thus take more time to process than would the preview of the target stimulus used by Richards and Neff (2004). Therefore, it is currently unclear whether the all-or-none mechanism implied by the results of Richards and Neff (2004) generalises to other auditory tasks that rely on endogenous (but not exogenous) attention.

“Inhibition of return” hypothesis

Another possible mechanism is that preparatory attention could improve the detection or discrimination of target auditory stimuli until a criterion duration, after which the mechanism returns to its baseline state and the improvement in performance diminishes. This idea is similar to inhibition of return (e.g. Posner & Cohen, 1984; Shulman, Remington, & McLean, 1979; Tsal, 1983), which describes an improvement in the accuracy and latency by which target stimuli are detected at short cue-target intervals (between about 100 and 300 ms), but an impairment in accuracy and latency at longer cue-target intervals (between about 500 and 3000 ms). Inhibition of return has been observed under exogenous attention in the auditory and visual modalities (e.g. Spence & Driver, 1998) and is not typically found in behavioural studies of endogenous attention. Nevertheless, experiments examining brain activity in response to endogenous auditory cues are consistent with a mechanism by which preparatory attention is evoked shortly after an instructive cue, but then returns to baseline, with a similar time course as inhibition of return.

For example, a recent experiment (Holmes, Kitterick, & Summerfield, 2016) examined preparatory brain activity using electro-encephalography (EEG) in a multi-talker listening task. A visual cue instructed participants to attend a target talker (based on the talker’s spatial location or gender) 1000 ms before two talkers started speaking. Holmes *et al.* (2016) isolated preparatory brain activity by contrasting event-related potentials (ERPs) in this multi-talker listening condition with a control condition in which the same visual stimuli were presented, but which had no implications for auditory attention. Preparatory EEG activity began approximately 50 ms after the visual cue was revealed. At first, this result appears consistent with those of Richards and Neff (2004), who found better detection thresholds for cue-target intervals of 50 ms than 5 ms. However, Richards and Neff (2004) found similar detection thresholds for cue-target intervals between 50 and 500 ms, suggesting an all-or-none mechanism. If preparatory attention was all-or-none, preparatory brain activity would be expected to last for the entire duration of the cue-target interval. In contrast to this prediction, Holmes *et al.* found that preparatory brain activity lasted 600 ms, ending 650 ms after the instructive cue rather than lasting for the remainder of the 1000-ms cue-target interval. This result implies that preparatory brain activity returns to baseline after 650 ms and, thus, leads to the prediction that accuracy and RTs for intervals longer than 650 ms would be similar to those obtained at 0-ms cue-target intervals. Richards and Neff (2004) did not test intervals greater than 500 ms and, therefore, a decrement in performance at intervals longer than 650 ms would not be observable in their results.

No experiments to our knowledge have compared behavioural performance across different-length cue-target intervals in a multi-talker listening task similar to that used by Holmes *et al.* (2016). Although, it is well-established that listeners achieve better speech intelligibility during multi-talker listening when they know in advance the target talker's spatial location (Best, Marrone, Mason, Kidd, & Shinn-Cunningham, 2009; Best, Ozmeral, & Shinn-Cunningham, 2007; Ericson et al., 2004; Kidd, Arbogast, Mason, & Gallun, 2005) or their identity (Kitterick et al., 2010) compared to when they have little or no time to prepare for talker attributes before a target talker begins to speak.

A related line of research has examined the effect of varying the length of time available for switching attention from one talker to another during multi-talker listening. The pattern of speech intelligibility across different-length switching intervals is consistent with the "Inhibition of return" hypothesis. While it is possible that these task-switching experiments engage different cognitive processes to those involved in preparing attention during a cue-target interval, switching attention has previously been proposed to engage preparatory processes (Meiran, Chorev, & Sapir, 2000); thus, findings from the task-switching literature could provide useful predictions for how different-length cue-target intervals might influence intelligibility during multi-talker listening.

It is well-established across a variety of tasks that there is a switch cost—that is, RTs are longer and accuracy is worse when participants have to attend to a different stimulus attribute to the previous trial than when participants maintain attention on the same stimulus attribute (e.g. Koch et al., 2011; Meiran et al., 2000; Rogers & Monsell, 1995). During multi-talker listening, participants are worse at classifying spoken digits and letters when they are instructed to attend to a different talker to the previous trial than the same talker (e.g. Koch et al., 2011) or when they are required to monitor a different talker in the second part of a trial to the first part of a trial compared to monitoring the same talker throughout (e.g. Larson & Lee, 2013).

The multi-talker listening switch cost is reduced when participants are given longer intervals over which to switch their attention, until a criterion duration (Koch et al., 2011; Larson & Lee, 2013; Meiran et al., 2000). For example, Larson and Lee (2013) presented participants with two simultaneous sequences of spoken letters, which differed in fundamental frequency. Participants were instructed to respond the second time the letter "E" occurred in the attended sequence of letters. At the beginning of each trial, participants received an auditory cue (preview of the target) that indicated which sequence of letters they should attend to in the first part of the trial and a visual cue that instructed them to either maintain attention on the same talker throughout the trial or switch attention to the other talker half-way through the trial. On every trial, there was a silent gap of variable duration after the first three digits were spoken during which participants either had to switch attention to the other talker or maintain attention on the same talker. When participants were required to switch attention to a different talker during the silent interval, accuracy and RTs were significantly better for moderate-duration gaps (400 and 600 ms) than shorter-duration gaps (100 and 200 ms), but became worse when the gap duration increased to 800 ms. This result is consistent with the "Inhibition of return" hypothesis, with an optimal time for switching attention that lasts approximately 600 ms. This finding implies that the processes that underlie attentional switching take a finite duration of time, but those processes return to a neutral state once the optimal duration is exceeded.

Current Experiment

To investigate the mechanism underlying endogenous preparatory auditory attention, we varied the length of the cue-target interval in a multi-talker listening task. Young adults reported words spoken by a target talker who spoke in a mixture of three talkers. A visual cue presented on each trial indicated the spatial location (left or right) or gender (male or female) of the target talker. We used a longer set of cue-target intervals than those used in previous experiments (which typically have not exceeded 800 ms) because we expected to find improvements in performance up to 600–800 ms and possibly a decrease in performance at longer intervals. Thus, we used five different-length cue-target intervals between 0 and 2000 ms. Despite the "All-or-none" mechanism implied by the results of Richards and Neff (2004)—who examined different-length cue-target intervals in the auditory modality and found a criterion duration at 50 ms—we expected the best cue-target duration in the current experiment to be longer. First, the experiment of Richards and Neff (2004) may have engaged some exogenous attentional processes whereas the current experiment investigated endogenous attention. The current experiment also used visual cues that were more abstract than the auditory cues used by Richards and Neff (2004), and may thus require longer time to process. Instead, we expected to find either a "Progressive Improvement" pattern, similar to previous experiments examining endogenous visual attention and consistent with studies showing that preparatory brain activity increases over time, or an "Inhibition of return" pattern, consistent with experiments examining task-switching and the finding that preparatory EEG activity returns to baseline in a previous multi-talker listening experiment. Although most of the previous experiments have analysed either RTs or accuracy (but not both), we analysed both accuracy and RTs in the current experiment. It has often been assumed that accuracy and RT reflect the same underlying mechanism, but it has previously been suggested that the two mechanisms might in fact differ (see Prinzmetal, McCool, & Park, 2005; van Ede, De Lange, & Maris, 2012).

Previous studies have reported that errors during multi-talker listening typically consist of words spoken by competing talker(s), rather than words that were not spoken on that trial (Brungart & Simpson, 2002; Darwin, 2006). We also aimed to examine the types of errors made on trials in which participants did not correctly report the target words because errors can provide additional information about the mechanism by which preparatory attention improves speech intelligibility. For the current three-talker listening task, participants had to report a colour and a number key word spoken by the target talker. The two competing talkers also spoke different colour and number key words. If participants did not segregate the talkers correctly, then errors would be expected to consist of words that were not spoken on that trial. If participants segregated the talkers but the segregated words were not assigned to the correct sources, then errors would consist of a colour word that was spoken by one talker and a number word that was spoken by a different talker. Whereas, if participants segregated the talkers and assigned consecutive colour and number words to their correct sources, but attended to the incorrect source, then errors would consist of colour-number combinations that were spoken by one of the two distracting talkers. The aim was to investigate whether the proportions of these error types differed between the cue-target interval conditions, to help elucidate the mechanism by which different preparation time conditions improve speech intelligibility. We predicted that longer cue-target intervals might allow participants to better segregate the talkers and, thus, errors for longer cue-target intervals would be more likely to consist of words that were spoken by one or more of the competing talkers on that trial, rather than words that were not spoken on that trial. In addition, we predicted that longer cue-target intervals might promote streaming, such that errors would be more likely to consist of colour-number combinations that were spoken by one of the two distracting talkers for longer cue-target intervals, rather than a mixture of colour and number words that were spoken by two different talkers.

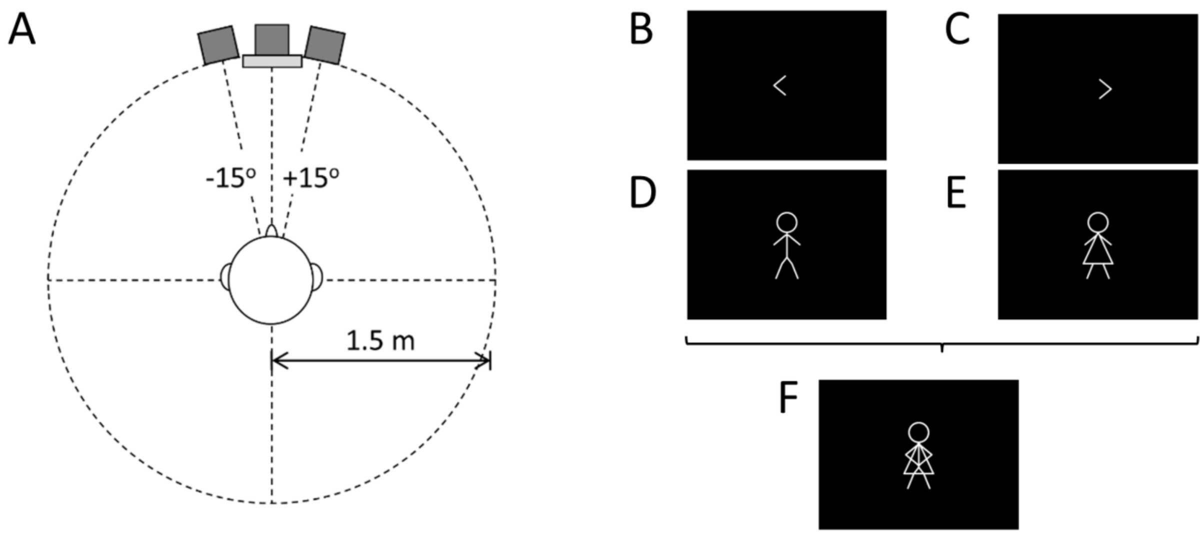


Fig. 2. (**A**) Layout of loudspeakers (dark grey squares) and visual display unit (light grey rectangle) relative to a participant's head. Visual cues for location (**B,C**) and gender (**D,E**). A visual composite stimulus (**F**) was created by overlaying the four visual cues.

We conducted an additional analysis to explore whether participants were attending to the location and gender of a talker in combination (i.e. using ‘object-based’ attention), or to only the cued attribute on each trial (i.e. using location- or feature-based attention). The logic of this analysis arose from the well-established switch cost effect—the finding that RTs are longer when participants have to switch attention to a different attribute than when participants maintain attention on the same attribute (Monsell & Driver, 2000; Rogers & Monsell, 1995)—and resembles analyses that have been used to identify object-based attention in previous studies (e.g., Ihlefeld & Shinn-Cunningham, 2008). This analysis focussed on trials in which participants received the same visual cue on two consecutive trials. We compared trials in which the non-cued attribute remained the same as the previous trial with trials in which the non-cued attribute changed. The rationale behind this approach was that accuracy and RTs would be influenced by the non-cued dimension on the previous trial if participants used 'object-based' attention, but not if they used location-based attention when they were cued to attend to the left or right talker and feature-based attention when they were cued to attend to the male or female talker. We also hypothesised that the extent to which participants used 'object-based' attention would depend on the cue-target interval.

Method

Participants

Participants were 20 young adults (10 male), aged 18–24 years (mean [M] = 19.6, standard deviation [SD] = 1.8). They were self-declared native English speakers with no history of hearing problems. The study was approved by the Research Ethics Committee of the Department of Psychology of the University of York.

Apparatus

The experiment was conducted in a 5.3 m x 3.7 m single-walled test room (Industrial Acoustics Co., NY) located within a larger sound-treated room. Participants sat facing three loudspeakers (Plus XS.2, Canton, Germany) arranged in a circular arc at a height of 1 m at 0° azimuth (fixation) and at 15° to the left and right (Fig. 2). The loudspeakers were visible to participants. A 15-inch visual display unit (VDU; NEC AccuSync 52VM) was positioned directly below the central loudspeaker. Participants were instructed to fixate on the centre of the visual display unit until they were ready to respond, although their heads were not restrained.

Stimuli

Visual cues. Four visual cues, "left", "right", "male", and "female", were defined by white lines on a black background. Left and right cues were leftward- and rightward-pointing chevrons, respectively; male and female cues were stick figures (Fig. 2B–E). A composite visual stimulus was created by overlaying the four cues (Fig. 2F).

Acoustic stimuli. Acoustic stimuli were modified phrases from the Co-ordinate Response Measure corpus (CRM; Moore, 1981) spoken by native British-English talkers (Kitterick et al., 2010). The original stimuli (which were spoken versions of entire CRM sentences; e.g. “Ready Baron, go to green two now”) were cut so that each sentence had the form "<colour> <number> now”. In the cut phrases, the onset of the <colour> word was the same across talkers, and the onsets of the other two words occurred at approximately the same time across talkers (owing to minor differences in speaking rate between talkers). There were four colours (‘blue’, ‘red’, ‘green’, ‘white’) and four numbers (‘one’, ‘two’, ‘three’, ‘four’). An example is ‘Green two now’. Phrases spoken by one male talker and one female talker were selected from the corpus. An additional female talker was selected from the corpus, whose voice was manipulated to sound like a child’s voice by raising the fundamental frequency and the frequencies of the formants using Praat (Version 5.3.08; http://www.praat.org/). The average duration of the phrases was 1.4 s. The levels of the digital recordings of the sentences were normalised to the same root mean square (RMS) power.

The average presentation level of concurrent triplets of CRM phrases was set to 63 dB(A) SPL (range 61.6–66.2 dB) measured with a B&K (Brüel & Kjær, Nærum, Denmark) Sound Level Meter (Type 2260 Investigator) and 0.5-inch Free-field Microphone (Type 4189) placed in the centre of the arc at the height of the loudspeakers with the participant absent.

Procedure

At the start of each trial, a fixation cross was presented for 1000 ms (Fig. 3). Next, a visual composite stimulus was presented, which faded to reveal the visual cue for each trial. The fade lasted 200 ms. The total amount of time between the onset of the visual composite stimulus and the onset of the acoustical stimuli was fixed at 3000 ms. Although, the relative durations of the visual composite stimulus and visual cue within the 3000-ms interval varied quasi-randomly from trial to trial. There were five possible intervals between the full reveal of the visual cue and the onset of the acoustical stimuli: 0,

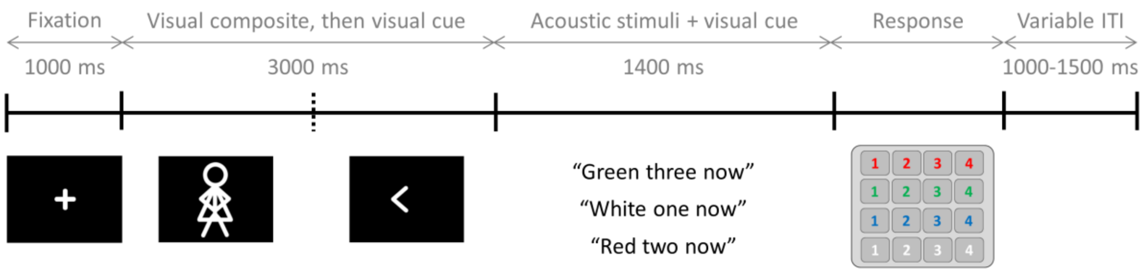


Fig. 3. Schematic of trial structure. The interval between the onset of the visual composite and the onset of the acoustic stimuli was fixed at 3000 ms on every trial. The duration of the visual cue was 0, 250, 500, 1000, or 2000 ms; there was a 200-ms fade between the visual composite and visual cue; the visual composite was presented for the remainder of the 3000-ms interval. Stimuli for example trials are displayed below, with an example of the visual stimuli (left trial), acoustical stimuli (centre) and response buttons (right).

250, 500, 1000, and 2000 ms. For a trial with a 2000-ms cue-target interval, the visual composite was presented for 800 ms, the fade between the visual composite and visual cue lasted 200 ms, and the visual cue was fully revealed for 2000 ms before the acoustic stimuli began.

Both the visual cue duration and the cue type (left, right, male, and female) were randomly interleaved within blocks. One phrase was played from each loudspeaker (left, centre, and right) with the same onset time but a different colour-number combination. The "child" voice was always played from the central loudspeaker and was never the target. Of the remaining two voices, one was always the male and the other was always the female and they were presented equally often at the left and right loudspeakers. The three talker identities remained the same over the course of the experiment.

The visual cue directed attention to the target talker and varied quasi-randomly from trial to trial. The cue remained on the screen throughout the duration of the acoustic stimuli. Participants were instructed to report the colour-number combination in the target sentence by pressing a coloured digit on a touch screen directly in front of their chair. They were instructed to respond as quickly and as accurately as possible. The coloured digits appeared on the screen before each trial and participants were able to respond at any time during the trial. Participants were instructed to look at the central video screen at the beginning of each trial. The inter-trial interval varied randomly from 1000 to 1500 ms. Each participant completed 360 trials (72 for each cue duration and, within this, 18 trials for each of the different visual cues), with a break every 40 trials.

The logic behind the design was that, on every trial, there was a fixed time interval (3000 ms) between the onset of the visual composite stimulus and the onset of the acoustic stimuli. This aspect ensured that any differences between different cue-target intervals must be explained by differences in the duration of time for which participants received information about the location or gender of the upcoming talker. Any advantage for longer cue-target intervals, therefore, could not be explained by a general increase in arousal for longer cue-target intervals or by changes in the predictability of the onset time of the acoustic stimuli.

Prior to the main task, participants completed two sets of familiarisation trials. In the first set, 12 trials were presented in which *either* the male or female talker was presented on each trial from the left or right loudspeaker. The aim was to familiarise participants with the left and right locations and with the male and female talkers that would be used in the main task. The trial structure was the same as the main task, with the exception that only the male or female talker was presented on each trial. The second set of familiarisation trials were identical to the main task. Participants completed 4 trials (1 for each visual cue). Each trial contained all three voices. During both sets of familiarisation trials, the cue-target interval varied quasi-randomly from trial-to-trial.

Analyses

Accuracy and RTs. Trials were separated into attend-location (average left/right cues) and attend-gender (average male/female cues) groups, separately for each of the five cue-target interval conditions. For each condition, we calculated the percentage of trials in which participants correctly identified both the colour and number (i.e. the "colour-number combination") spoken by the target talker. We also calculated average RTs, measured from the onset of the acoustic stimuli, on trials in which participants correctly identified the colour-number combination. The patterns of significance for accuracy and RTs across the five cue-target intervals were the primary method we used to distinguish which of the four hypotheses described in Fig. 1 were supported by the data.

To gain another perspective on which hypothesis was most likely, we also fitted four models to the accuracy and RT data, separately for attend-location and attend-gender trials. The "All-or-none" hypothesis was modelled as a step function (consisting of one step) with three free parameters (a, b, and c), as shown by the following equations:

$$\{x: x < a\}\ y = b \qquad [1]$$
$$\{x: x \geq a\}\ y = c$$

Where x corresponds to the duration of the cue-target interval and y corresponds to accuracy (%) or RT (seconds).

The "Inhibition of return" hypothesis was modelled as a quadratic function with three free parameters:

$$y = ax^2 + bx + c \qquad [2]$$

The "Progressive improvement" hypothesis was modelled as an exponential function with three free parameters:

$$y = ae^{bx} + c \qquad [3]$$

Finally, the null hypothesis was modelled as a flat linear function with one free parameter:

$$y = a \qquad [4]$$

We used the 'fminsearch' function in MATLAB 2014b (The MathWorks, Inc., Natick, MA, USA) to fit each function to the average data and we compared the percent of variance in the data that were explained by each of the four models.

Errors. When participants did not correctly identify the colour-number combination, responses were categorised into one of four different types of error. The reported colour-number combination could be: (1) spoken by the opposite-gender talker that was presented from the contralateral location ("opposite-gender" error), (2) spoken by the "child" talker that was presented from the central location ("child" error), (3) a mixture of words spoken by the target and a non-target talker *or* a mixture of words spoken by the two non-target talkers ("mix" error), or (4) not spoken by any mixture of the talkers on that trial ("absent" error).

The percentages of the four types of error were assessed in relation to the percentages expected if participants guessed randomly with a uniform distribution. The expected percentages were: 6.7% "opposite-gender" error, 6.7% "child" error, 40.0% "mix" error, and 46.7% "absent" error.

Trial-by-trial analysis. Colour-number accuracy and RTs were compared between trials in which the array of talkers had the same compared to a different configuration to the previous (n-1th) trial. For the attend-gender condition, trials in which the target talker was the same gender and had the same location as the previous trial were compared with trials in which the target talker was the same gender but had a different location. For the attend-location condition, trials in which the target talker was the same gender and had the same location as the previous trial were compared with trials in which the target talker was the opposite gender but had the same location.

Results

Accuracy

Fig. 4A illustrates the accuracy of reporting the colour-number combination (plots for individual participants are displayed in Supplemental Fig. 1). A 5 x 2 repeated-measures ANOVA showed a significant main effect of cue-target interval [$F(2.9, 54.1) = 3.50$, $p = 0.023$, $\omega^2 = 0.11$]. Bonferroni-corrected post-hoc tests showed significantly better colour-number accuracy for the 2000-ms than 0-ms interval ($p = 0.001$). After Bonferroni correction, none of the other cue-target intervals had colour-

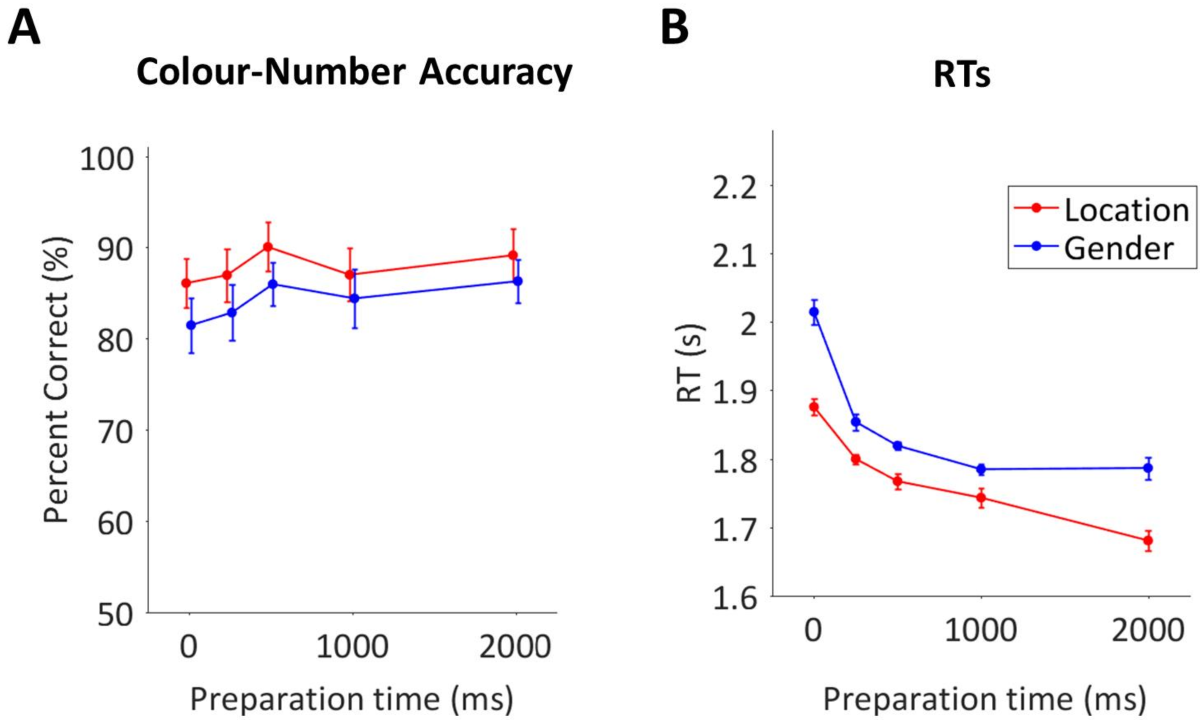


Fig. 4. Accuracy and reaction time (RT) results. (**A**) Mean percentage of trials in which participants correctly identified the colour-number combination spoken by the target talker. (**B**) Mean RTs of correct trials, relative to the onset of acoustic stimuli.

number accuracies that were significantly better than the 0-ms interval ($p > 0.08$). Comparing adjacent intervals, there were no significant differences in accuracy ($p > 0.13$). This pattern of results is most consistent with the "All-or-none" hypothesis and is partially consistent with the "Progressive improvement" hypothesis. The pattern of results is also consistent with the "Inhibition of return" hypothesis with an optimal interval that occurs at or longer than 2000 ms.

Participants achieved better colour-number accuracy in the attend-location condition (M = 87.8%, SD = 4.7) than the attend-gender condition (M = 84.2%, SD = 5.2) [$F(1, 19) = 13.75$, $p = 0.001$, ω^2

= 0.38]. There was no significant interaction between cue-target interval and cue type [$F(4, 76) = 0.24$, $p = 0.92$, $\omega^2 = -0.04$].

To gain another perspective on which of the hypotheses best explained the accuracy data, we fitted four models to the average results, separately for attend-location and attend-gender trials. The fitted equations are listed in Table 1. Fig. 5A and 5C illustrate the fitted functions alongside the data and Fig. 5B and 5D illustrate the R^2 values for each model fit. None of the models explained more than 53% of the variance in attend-location trials or more than 82% of the variance in attend-gender trials. Consistent with the pattern of significance, the step model ("All-or-none" hypothesis) explained most variance in the data for both attend-location and attend-gender trials; the 'step' for the best-fitting functions occurred between 400 and 450 ms. Thus, although all of the four models leave a substantial portion of the variance unexplained, out of the four models tested the accuracy results most closely resemble the "All-or-none" hypothesis.

Table 1. Equations fitted to accuracy and RT results for models corresponding to each of the four hypotheses (H1, H2, H3, H4).

Dependent variable	Model	Fitted equation	R^2
Accuracy (attend-location)	H1: Step function	$\{x: x < 409\}\ y = 86.5$ $\{x: x \geq 409\}\ y = 88.7$	0.53
	H2: Quadratic function	$y = -8.14*10^{-7} x^2 + 2.60*10^{-3} x + 86.7$	0.28
	H3: Exponential function	$y = 5500\ e^{1.86*10e\text{-}7\ x} + -5.42$	0.25
	H4: Flat linear function	$y = 87.8$	0
Accuracy (attend-gender)	H1: Step function	$\{x: x < 446\}\ y = 82.2$ $\{x: x \geq 446\}\ y = 85.5$	0.82

	H2: Quadratic function	$y = -1.23*10^{-6}\, x^2 + 4.20*10^{-3}\, x + 82.6$	0.70
	H3: Exponential function	$y = 2.85*10^{-7}\, e^{2.85*10e\text{-}7\, x} - 5.35$	0.58
	H4: Flat linear function	$y = 84.7$	0
RTs (attend-location)	H1: Step function	$\{x: x < 458\}\ y = 1.82$ $\{x: x \geq 458\}\ y = 1.73$	0.65
	H2: Quadratic function	$y = 7.47*10^{-8}\, x^2 - 2.31*10^{-4}\, x + 1.85$	0.91
	H3: Exponential function	$y = 0.167\, e^{-2.40*10e\text{-}3\, x} + 1.70$	0.95
	H4: Flat linear function	$y = 1.79$	0
RTs (attend-gender)	H1: Step function	$\{x: x < 420\}\ y = 1.93$ $\{x: x \geq 420\}\ y = 1.80$	0.62
	H2: Quadratic function	$y = 1.49*10^{-7}\, x^2 - 3.90*10^{-4}\, x + 1.98$	0.88
	H3: Exponential function	$y = 0.210\, e^{-4.70*10e\text{-}3\, x} + 1.79$	0.99
	H4: Flat linear function	$y = 1.88$	0

RTs

RTs became shorter as the duration of the cue-target interval lengthened (Fig. 4B; plots for individual participants are displayed in Supplemental Fig. 2). There was a significant main effect of cue-target interval [$F(1.4, 27.4) = 213.40$, $p < 0.001$, $\omega^2 = 0.91$]. Contrasts showed significantly shorter RTs for the 250-ms [$F(1, 19) = 590.86$, $p < 0.001$, $\omega^2 = 0.97$], 500-ms [$F(1, 19) = 442.39$, $p = 0.001$, $\omega^2 = 0.95$], 1000-ms [$F(1, 19) = 297.37$, $p < 0.001$, $\omega^2 = 0.93$], and 2000-ms [$F(1, 19) = 283.25$, $p < 0.001$, $\omega^2 = 0.93$] intervals than for the 0-ms interval. Bonferroni-corrected post-hoc tests also showed significant differences in RTs between all adjacent cue-target intervals ($p \leq 0.001$). The significant improvement in RTs between all adjacent cue-target intervals supports the “Progressive Improvement” hypothesis

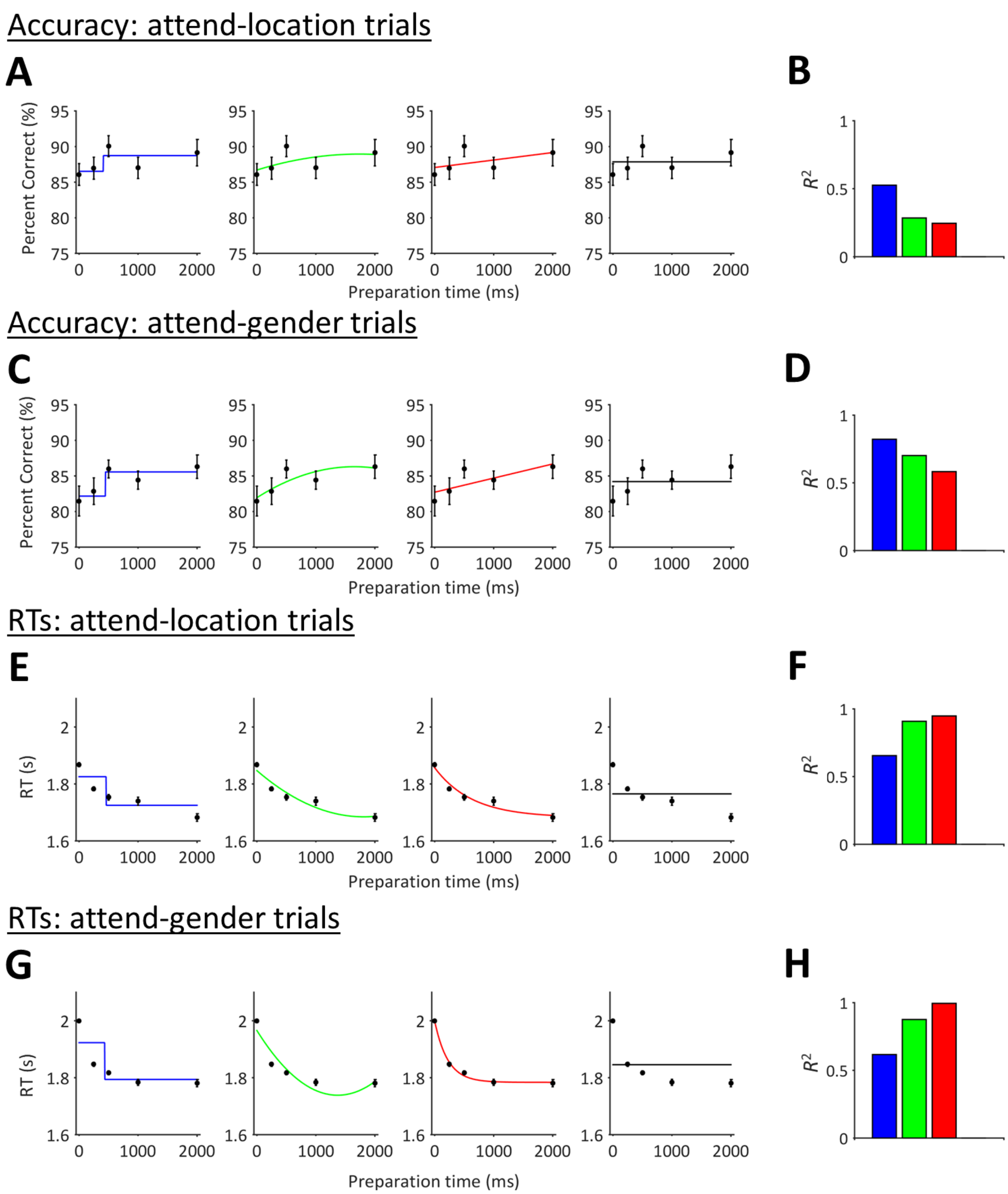


Fig. 5. Model fits to the accuracy and reaction time (RT) data [blue = step function, green = quadratic function, red = exponential function, black = flat linear function]. (**A**,**C**) The accuracy results for attend-location and attend-gender trials, respectively, are plotted alongside the fitted equations corresponding

to the four models. (**B**,**D**) Comparison of R^2 values for the four models. (**E**,**G**) and (**F**,**H**) show the same for the RT results.

because it is the only hypothesis that predicts that RTs should continue to improve significantly as the duration of the cue-target interval lengthens (compare Fig. 4B and 1H).

RTs were significantly shorter in attend-location (M = 1.8 s, SD < 0.1) than attend-gender (M = 1.9 s, SD < 0.1) trials [$F(1, 19) = 461.39$, $p < 0.001$, $\omega^2 = 0.96$]. There was also a significant two-way interaction between cue-target interval and cue type [$F(2.0, 38.4) = 103.13$, $p < 0.001$, $\omega^2 = 0.83$]. The interaction was underpinned by a greater improvement in RTs for attend-gender than attend-location trials between 0 and 250 ms [$F(1, 19) = 172.54$, $p < 0.001$, $\omega^2 = 0.89$] and between 500 and 1000 ms [$F(1, 19) = 5.57$, $p = 0.029$, $\omega^2 = 0.18$]; whereas, there was a similar improvement in RTs for the two attention condition between 250 and 1000 ms [$F(1, 19) = 0.46$, $p = 0.51$, $\omega^2 = -0.02$] and greater improvement for attend-location than attend-gender trials between 1000 and 2000 ms [$F(1, 19) = 90.65$, $p < 0.001$, $\omega^2 = 0.81$].

The fitted equations for the RT data are listed in Table 1. Fig. 5E and 5G illustrate the fitted functions alongside the data and Fig. 5F and 5H illustrate the R^2 values. The exponential function, which corresponded to the "Progressive improvement" hypothesis, explained more variance than any of the other models for both conditions. It explained 95% of the variance for attend-location trials and 99% of the variance for attend-gender trials. Thus, the best-fitting hypothesis from the modelling analysis matches that implied by the pattern of significance for RTs.

Errors

The largest percentage of errors were "mix" errors (M = 78.4%, SD = 9.1), where the reported colour-number combination was spoken by a mixture of the presented talkers. The second largest percentage of errors were "absent" errors (M = 17.2%, SD = 8.7), where the colour and/or number was not spoken by any of the talkers on that trial. Participants made "opposite-gender" errors (M = 3.5%, SD

= 4.3) and "child" errors (M = 1.0%, SD = 1.3) on a low proportion of trials. The percentages of "mix" [$t(19) = 19.33$, $p < 0.001$] and "absent" [$t(19) = 15.70$, $p < 0.001$] errors were significantly greater than their expected values, whereas the percentages of "opposite-gender" [$t(19) = 3.99$, $p = 0.001$] and "child" [$t(19) = 24.25$, $p < 0.001$] errors were significantly smaller than their expected values.

A 4 x 5 x 4 repeated-measures ANOVA investigated whether the types of errors (4 levels: "opposite-gender", "child", "mix", and "absent" errors) differed significantly between the different-length cue-target intervals (5 levels: 0, 250, 500, 1000, and 2000 ms) or between cue types (4 levels: left, right, male, and female). There was a significant main effect of error type [$F(1.5, 27.6) = 367.20$, $p < 0.001$, $\omega^2 = 0.95$]. Bonferroni-corrected post-hoc tests showed that the percentage of "opposite-gender" errors did not differ significantly from the percentage of "child" errors ($p = 0.18$), but there were significant differences between the percentages of all other error type combinations ($p < 0.001$).

There was no significant difference in the percentages of errors for different-length cue-target intervals [error type * cue-target interval interaction: $F(4.5, 84.5) = 0.57$, $p = 0.71$, $\omega^2 = -0.02$] and no significant difference in the percentages of errors across the four different cue types [error type * cue type interaction: $F(3.1, 59.8) = 1.09$, $p = 0.36$, $\omega^2 < 0.01$]. There was no significant three-way interaction [$F(8, 152) = 1.45$, $p = 0.18$, $\omega^2 = 0.02$].

Trial-by-trial analysis

A 5 x 2 x 2 ANOVA, conducted separately for the accuracy and RT data, showed no significant 2- or 3-way interactions between cue-target interval (0–2000 ms) and talker configuration (same/different) or cue type (location/gender). Thus, for plotting and subsequent analyses, accuracy and RTs were collapsed across cue-target intervals.

Fig. 6 shows that participants achieved better colour-number accuracy (Fig. 6A) and faster RTs (Fig. 6B) when the configuration of talkers was the same as the previous trial than when it was different.

We performed a 2 x 2 ANOVA, separately for colour-number accuracy and RTs, with the factors of configuration (same/different) and cue type (location/gender).

Colour-number accuracy was significantly better when participants were cued to location than gender [$F(1, 19) = 4.94$, $p = 0.039$, $\omega^2 = 0.16$], which is consistent with the results reported above (Fig. 4A). Trials with the same configuration as the previous trial (M = 91.2%, SD = 3.9) displayed significantly better accuracy than trials with a different configuration (M = 85.4%, SD = 6.1) [$F(1, 19) = 23.4$, $p < 0.001$,

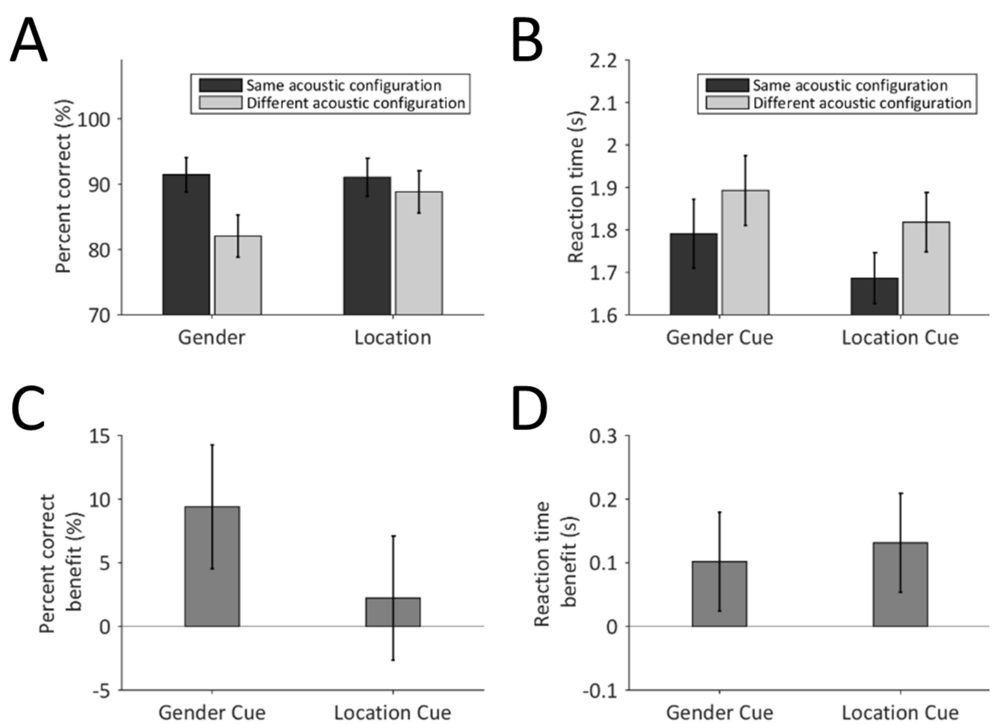


Fig. 6. Results from the trial-by-trial analysis. (**A**) Accuracy for reporting the colour-number combination spoken by the target talker, separated by cue type (location/gender), when the acoustic configuration was either the same (i.e. the location and gender of the target talker was the same) or different (i.e. the target talker varied on the uncued dimension) to the previous trial. (**B**) Accuracy benefit, calculated as the difference in percent correct when the acoustic configuration was the same as the previous trial

compared to when it was different. (**C,D**) Equivalent plots for RTs. Error bars display within-subjects 95% confidence intervals.

ω^2 = 0.52]. There was also a significant two-way interaction, with gender trials leading to a larger difference in accuracy between the same and different configuration conditions than location trials [$F(1, 19) = 4.75$, $p = 0.042$, $\omega^2 = 0.15$] (Fig. 6A).

RTs were significantly shorter when participants were cued to location than gender [$F(1, 19) = 22.88$, $p < 0.001$, $\omega^2 = 0.51$], which is consistent with the results reported above (Fig. 4B). There was also a main effect of configuration, with same-configuration trials (M = 1.7 s, SD = 0.2) displaying significantly shorter RTs than different-configuration trials (M = 1.9 s, SD = 0.2) [$F(1, 19) = 20.58$, $p < 0.001$, $\omega^2 = 0.48$]. The interaction between configuration and cue-type was not significant [Fig. 4D; $F(1, 19) = 0.32$, $p = 0.58$, $\omega^2 = -0.03$].

Discussion

During three-talker listening, RTs for reporting key words spoken by a target talker systematically shortened as the duration of the cue-target interval increased from 0 to 2000 ms. The current results are consistent with previous multi-talker listening experiments in which trials with advance cues were compared to trials with no advance cues. Those experiments demonstrated a behavioural advantage from knowing the spatial location (Best et al., 2007; Ericson et al., 2004; Kidd et al., 2005) or the identity (Kitterick et al., 2010) of a target talker before he or she begins to speak. However, those experiments did not compare speech intelligibility across more than two cue-target intervals. The current results build upon those of previous multi-talker listening experiments by showing that the *duration* of the cue-target interval affects the accuracy and latency of speech intelligibility.

The current results provide strong evidence for an improvement in multi-talker listening as the duration of the cue-target interval increases to 2000 ms, demonstrated by both better accuracy and shorter RTs for the 2000-ms than the 0-ms cue-target interval. In addition, RTs became significantly shorter as the duration of the cue-target interval lengthened between 0 and 2000 ms. The “Progressive improvement” hypothesis is the only one of the four that we compared that predicts this pattern of results. Thus, we judge that the results are most consistent with the “Progressive improvement” hypothesis—longer cue-target intervals continue to improve intelligibility progressively as the duration of the cue-target interval increases.

The accuracy results showed a significant improvement only for the 2000-ms cue-target interval compared to the 0-ms interval. This result is most consistent with the “All-or-none” hypothesis. Based on the modelling results, the threshold at which accuracy improves is most likely to occur between 400 and 450 ms. There are several different reasons why we might have found differences between the accuracy and RT results. First, there might have been a speed-accuracy trade-off, where improvements in accuracy with increasing durations of preparation time were sacrificed for shorter RTs. This effect would obscure improvements in accuracy with longer cue-target intervals. Second, effects on accuracy may have been difficult to detect in this experiment, because accuracy was > 80% in the 0-ms condition. Although accuracy was not at ceiling level, the effects might be smaller at this level of accuracy than if the task was more difficult, meaning that significant differences between adjacent cue-target intervals were not observed. A third explanation is that accuracy and RTs depend on different underlying processes (Prinzmetal, McCool, & Park, 2005; van Ede, De Lange, & Maris, 2012), such that an “All-or-none” model is sufficient to explain the accuracy results, whereas a combination of the “All-or-none” and “Progressive improvement” hypotheses explain the RT data. These possibilities are indistinguishable based on the current results. Nevertheless, taken together, the results imply that the mechanism for preparatory auditory attention develops over time—speech spoken by a target talker is able to be identified more quickly when listeners are able to partially prepare for the location or gender of the target talker, with further improvements when longer time is available for preparatory attention to develop.

The effect of cue-target interval on accuracy and RTs cannot be explained by an increase in general arousal that was unspecific for the cued attribute. First, the different cue-target interval conditions were randomised within blocks. Second, the time between the onset of the fixation cross and that at which the talkers began to speak was identical for all trial types. Instead, the difference in RTs must arise from preparing for the cued attribute. This advantage could be explained by an enhancement in processing of the target stimulus or a reduction in interference from the masker talker.

Participants might have used the 200-ms fade time between the visual composite and visual cue to prepare for the cued attribute because the cue would be detectable before it was fully revealed. Although this aspect of the design would not affect comparisons between trials with different length cue-target intervals, it means that participants would have had up to 200 ms to prepare for the cued attribute in the 0-ms condition. Given that the steepest shortening of response latency occurred at the shortest cue-target intervals (Fig. 3), the difference between the 0 and 2000-ms cue-target intervals might therefore underestimate the difference that would have been obtained if the 200-ms fade was not present.

“Progressive Improvement” hypothesis

The pattern of results that emerges from the current experiment most closely aligns with the ”Progressive improvement” hypothesis because there was a progressive shortening of RTs as the duration of the cue-target interval increased from 0 to 2000 ms (compare Fig. 4B and Fig. 1H). The “Progressive improvement” hypothesis is the only one of the four that we tested that predicts significantly shorter RTs as the duration of the cue-target interval increases. The “All-or-none” hypothesis predicts an improvement in RTs between two adjacent intervals, but no difference at shorter or longer intervals. The “Inhibition of return” hypothesis predicts an improvement as the cue-target interval lengthens for short intervals, but it also predicts that RTs should become significantly worse as the interval lengthens beyond an optimal duration. Consistent with the idea that the data most closely align with the “Progressive improvement” hypothesis, the model corresponding to this hypothesis provided the best fit to the RT data, explaining more than 95% of the variance (Fig. 5F,H).

Although the pattern of significance for the accuracy data is most consistent with the “All-or-none” hypotheses, it is also consistent with the “Progressive improvement” hypothesis, if we assume that differences between adjacent cue-target intervals were too small to detect with the intervals used here. The “Progressive improvement” hypothesis is the only of the four we tested that is consistent with both the pattern of RT and pattern of accuracy results. Of course, it is possible that the accuracy and RT results reflect different underlying processes, so would be best fit by different combinations of models (discussed below).

We found an improvement in RTs as the cue-target interval lengthened that was similar to the improvement in the detection of target stimuli (Yamaguchi et al., 1994) and contrast thresholds (Lu et al., 2009) observed in previous studies of endogenous visual preparatory attention. The current experiment used longer intervals than the intervals that have previously been used in visual endogenous cueing tasks; thus, based on the current results, performance in visual endogenous cueing tasks might also improve at cue-target intervals of 2000 ms compared to shorter intervals, although that improvement would have been missed by the choice to compare only shorter (≤ 800 ms) intervals.

The neural mechanisms that underlie auditory preparatory attention have been less well studied than their visual counterpart. Given that the current pattern of results aligns closely with those observed in vision, it is possible that similar neural mechanisms underlie preparatory attention for stimuli in both sensory modalities. Preparatory visual endogenous attention is likely underpinned by increased activity in the neural circuits involved in selectively attending to the cued stimulus attribute (Giesbrecht, Woldorff, Song, & Mangun, 2003; Slagter, Giesbrecht, & Kok, 2007; Woldorff et al., 2004). For example, the amplitude of pre-target BOLD activity in visual cortex correlates positively with performance on a visual discrimination task (Giesbrecht, Weissman, Woldorff, & Mangun, 2006). Thus, in preparation for visual target arrays, the amplitude of pre-target activity may increase with longer cue-target intervals, progressively improving task performance. Consistent with this idea, the contingent negative variation (CNV; Walter, Cooper, Aldridge, McCallum, & Winter, 1964), an ERP thought to reflect anticipation of an upcoming stimulus (e.g. Chennu et al., 2013), develops after a cue is presented and builds up over time until a target occurs. The latency of the CNV is correlated with the length of subjective judgements of interval duration (Ruchkin, McCalley, & Glaser, 1977), perhaps reflecting anticipation of the time at which a target stimulus will occur. Greater CNV amplitudes have also been shown to relate to better detection of acoustic target stimuli (Rockstroh, Müller, Wagner, Cohen, & Elbert, 1993), suggesting that an increase in neural population activity may help participants to detect target stimuli.

Similar to findings in the visual modality (e.g. Giesbrecht et al., 2006), several studies have shown that brain areas involved in selectively attending to a feature of a target talker (such as their spatial location or the fundamental frequency of their voice) are also activated during the cue-target interval (e.g. Hill & Miller, 2010; Lee et al., 2013). Hill and Miller (2010) found that preparing to attend to the location or fundamental frequency of a target talker evoked activity in a left-dominant fronto-parietal network, including inferior frontal gyrus, dorsolateral pre-central sulcus, inferior parietal sulcus, and the superior parietal lobule. Importantly, activity within that network differed significantly when participants prepared to attend to location compared to fundamental frequency. Consistent with the idea that the amplitude of pre-target activity may increase with longer cue-target intervals, Voisin *et al.* (2006) showed that, when participants were asked to detect a tone in silence on a side cued by a central arrow stimulus, BOLD activity in superior temporal cortex contralateral to the expected direction increased in amplitude over the silent cue-target interval.

In addition, preparatory activity resembling the CNV was found in two previous EEG experiments using similar multi-talker listening tasks as that used in the current experiment (Holmes et al., 2016, 2017). Holmes *et al.* (2016) observed CNV-like activity during a 1000-ms cue-target interval when adults were cued to attend to the location or gender of a target talker who spoke in a mixture of two talkers. Another study by the same authors (Holmes *et al.*, 2017) used a task more closely resembling the current experiment because there were three talkers in the mixture. They used a cue-target interval of 2000 ms and found preparatory activity during location trials in two distinct phases: the first started soon after the cue was revealed, whereas the second occurred throughout the 1000 ms immediately before the target talkers started speaking. The finding of preparatory brain activity lasting longer than 1000 ms is consistent with the finding that RTs for attend-location trials in the current experiment were shorter for the 2000-ms compared to the 1000-ms cue-target interval. Furthermore, both of these EEG experiments (Holmes et al., 2016, 2017) found that the duration of significant preparatory activity was shorter on gender than location trials—consistent with the current result that RTs on attend-gender trials did not improve between 1000-ms and 2000-ms cue-target intervals as much as for attend-location trials.

There are several different mechanisms that could underlie an increase in amplitude in preparatory neural circuits related to successful preparatory attention. There is substantial evidence that such activity could arise from increased gain in neuronal populations tuned to the target stimulus (e.g. Chawla, Rees, & Friston, 1999; O'Connell, Barczak, Schroeder, & Lakatos, 2014; Treue & Martinez-Trujillo, 1999) and perhaps also suppression of responses in neuronal populations tuned to non-target stimuli (Gazzaley, Cooney, McEvoy, Knight, & D'Esposito, 2005; Jensen & Mazaheri, 2010; O'Connell et al., 2014; Seidl, Peelena, & Kastner, 2012). An alternative (but not mutually exclusive) possibility is that the distribution of activity across neurons may shift over the cue-target interval—changing from a more broadly activated area of neurons to an increasingly smaller population that is more specific for the target stimulus, possibly underpinned by a change in receptive field size (e.g. Anton-Erxleben, Stephan, & Treue, 2009; Fritz, Shamma, Elhilali, & Klein, 2003; Womelsdorf, Anton-Erxleben, Pieper, & Treue, 2006). When participants are cued to the location of an upcoming talker, the distribution of active neurons may shift from a group of neurons whose preferred locations include a large range of spatial locations on either side of the target location to a set of neurons that are more narrowly tuned to the target location. Another possible mechanism is that neuronal populations tuned to a target location or identity (e.g. for the talker's fundamental frequency [f_0] or vocal tract length [VTL]) may display more synchronised oscillatory activity (Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008; O'Connell et al., 2015, 2014) during longer cue-target intervals, increasing the sensitivity of entrained neuronal populations to the attended attribute (Fries, 2001, 2005). By any of these mechanisms (or a combination of multiple mechanisms), preparatory attention could potentially prime the neural circuitry required for attending to a target stimulus, allowing listeners to more quickly process and respond to target speech.

"All-or-none" hypothesis

Rather than showing a progressive improvement in RTs with longer cue-target intervals, a different prediction was that the results could show an all-or-none pattern. If the mechanism underlying preparatory attention in the current experiment was all-or-none, we should have found a significant difference in RTs between two adjacent cue-target intervals around a criterion duration, but no difference between intervals that were shorter or longer than those adjacent to the criterion. The "All-or-none" hypothesis seems plausible based on the accuracy results alone, although the pattern of RT results is incompatible with the "All-or-none" hypothesis. One possibility is that the accuracy and RT results reflect different underlying mechanisms, such that the "All-or-none" hypothesis underlies the accuracy data, but an additional mechanism (that progressively improves over time) is necessary to explain the RT data.

A pattern of results consistent with the "All-or-none" hypothesis was observed in a previous auditory cueing study by Richards and Neff (2004). Similar to the pattern of results observed for detection thresholds by Richards and Neff (2004), we found a significant difference in RTs between the two shortest intervals tested (in the current experiment, 0 and 250 ms; in the experiment of Richards and Neff, 5 and 50 ms). However, unlike the pattern of results found by Richards and Neff, we also found improvements in RTs between intervals longer than 250 ms. Differences between the current results and those of Richards and Neff (2004) could be explained by differences in the type of attention elicited by the different cues (exogenous in the previous study, endogenous in the current study). In addition, differences in the task and stimuli might contribute because the current experiment involved discrimination of speech stimuli, whereas the task used by Richards and Neff (2004) involved detection of a 1000 Hz tone and they measured contrast thresholds. This idea is consistent with the finding that the time course of visual spatial attention is influenced by properties of a visual target array (Cheal & Lyon, 1992) and difficulty of the task (Lyon, 1987). Nevertheless, Richards and Neff used an accuracy-based measure, which is consistent with the idea that accuracy can be explained by an "All-or-none" mechanism, whereas an additional mechanism is required to explain the RT results.

"Inhibition of Return" hypothesis

Although the model corresponding to the "Inhibition of return" hypothesis provided a reasonable fit to the RT data, explaining approximately 90% of the variance, it did not provide as good a fit as the model corresponding to the "Progressive improvement" hypothesis. Furthermore, the patterns of significance for the accuracy and RT data were inconsistent with the "Inhibition of return" hypothesis. This hypothesis predicts that RTs should become significantly longer and/or accuracy should become significantly worse as the cue-target interval lengthens beyond an optimal duration, but we observed neither of those results. Although it is possible that the optimal interval is longer than the durations we tested, this idea seems unlikely because previous experiments observing this type of pattern have found an optimal interval at approximately 400–600 ms, which is within the range of the intervals that we tested.

The current experiment reveals a different pattern of results across cue-target intervals than previous experiments requiring participants to switch attention during multi-talker listening. The key difference between previous experiments requiring participants to switch attention (e.g., Larson & Lee, 2013) and the current experiments is that those experiments varied the interval provided to switch attention from one sequence to another (but the interval between the onsets of the cue and the switching interval remained constant), whereas the current experiments varied the cue-target interval (i.e. the length of time between the onset of an instructive cue and the onset of target speech). Switching attention has sometimes been thought to engage preparatory processes (Meiran et al., 2000), which may or may not be the same as the processes elicited during the cue-target interval in the current experiment. The different time courses of performance benefit between the current results and the results of Larson and Lee (2013) imply that different processes are involved in preparing attention during a cue-target interval than those required to switch attention to a different talker. To better understand how the mechanisms that occur during these two intervals differ, future experiments could directly compare the time courses of brain activity elicited by attentional preparation and attentional switching.

The “Inhibition-of-return” hypothesis was also motivated by the results of a previous experiment showing that EEG activity in preparation for multi-talker listening lasted approximately 600 ms, but returned to baseline for the remainder of the 1000-ms cue-target interval (Holmes et al., 2016). However, this previous experiment used a two-talker listening task, whereas a three-talker task was employed in the current experiment. The two-talker task had a lower perceptual load than the current three-talker task and participants achieved better accuracy for reporting key words (95%, on average, compared to 86% in the current experiment). Thus, participants may not have gained such a large benefit from deploying preparatory attention throughout the preparatory interval in the two-talker task used by Holmes et al. (2016) as in the more demanding task used in the current experiment.

Consistent with the idea that differences in the duration of preparatory brain activity may depend on the task, a two-talker version of the current experiment (unpublished) showed no difference in RTs across the same cue-target intervals as those used in the current experiment. The two-talker experiment presented full-length CRM sentences from ± 30° azimuth. In the two-talker experiment, accuracy was already at ceiling for the 0-ms cue-target interval; thus, increasing the duration of the cue-target interval did not lead to better speech intelligibility performance. Previous experiments have also reported differences in the benefit of preparation between tasks of different difficulty. For example, Ericson *et al.* (2004) found that, during three-talker listening, knowing the spatial location of an upcoming target talker led to better speech intelligibility than when participants had no information about the target location; whereas, for two-talker listening, accuracy was near-ceiling accuracy even when participants received no cue. Similar results have also been reported by Brungart *et al.* (2001). The results are also consistent with those reported by Lu *et al.* (2009) in the visual modality—they only found improvements in contrast thresholds for longer cue-target intervals when they used stimuli with a high level of external noise (i.e. using noise images that were superimposed on the test stimuli) but not when they used stimuli with lower levels of external noise. Taken together, these findings suggest that advance cueing is not necessary for accurate speech intelligibility, but can improve the accuracy and/or latency of speech intelligibility in challenging multi-talker listening situations.

Comparison between location and gender cues

We found significantly faster RTs on attend-location than attend-gender trials. This finding could be explained by differences in the extent to which the talkers were segregable based on their location (which differed by 30 degrees azimuth) compared to their fundamental frequency or vocal tract length, which participants may have used to determine the gender of the talkers. Another possible explanation for the difference is that the stick figures we used as cues for gender may take longer to interpret than chevrons because chevrons have sometimes been thought to direct attention relatively automatically (see Ristic & Kingstone, 2006). Differences in the length of time taken to interpret the cues could underlie the difference in RTs, particularly at the shorter cue-target intervals, which might (at least partially) explain why we found a significant interaction in RTs between the duration of the cue-target interval and the attended attribute (i.e., location or gender). The interaction could also be driven by greater benefits to RTs when the task is more difficult because baseline RTs at 0-ms were longer for attend-gender than attend-location trials.

Errors on incorrect trials

The possible origin of incorrect responses was inferred from the analysis of error types. Contrary to the prediction, there were no significant differences in the proportions of different errors made across the cue-target intervals, suggesting that advance cueing did not specifically facilitate the ability to segregate the target talker from the other talkers or to stream words spoken by each talker. However, the lack of significant difference might be explained by high accuracy overall, thus reducing the power for detecting differences in error proportions between cue-target intervals. For all of the cue-target intervals, the largest proportion of errors by far were "mix" errors, which consisted of a colour that was spoken by one talker and a number that was spoken by a different talker. Thus, the results suggest that, when participants do not correctly identify the colour-number combination, they successfully segregate words spoken by different talkers (because they report words that were present in the mixture), but fail to stream the talkers (i.e. they do not correctly identify colour and number key words that were spoken by the same talker).

Object-based attention

The trial-by-trial analysis provides evidence that participants attended simultaneously to both the location and the gender of the target talker across all cue-target intervals. On trials in which the visual cue was identical to the previous trial, RTs were shorter when the configuration of talkers remained the same as the previous trial compared to when the configuration changed from the previous trial (Fig. 6). By demonstrating that a task-irrelevant attribute influences speech intelligibility, this result is consistent with the idea of object-based attention, which has been reported in previous studies of visual (Nishida, Shibata, & Ikeda, 2014; Scholl, 2001) and auditory (Shinn-Cunningham, 2008; Ihlefeld & Shinn-Cunningham, 2008) attention. Thus, when participants attend to a talker based on cues for spatial location or gender, activity or phase-locking may increase in neurons that preferentially encode the spatial location of the target talker (e.g. those tuned to the left or right hemifield; for a review, see Salminen, Tiitinen, & May, 2012) in combination with neurons that preferentially encode the f_0 and/or VTL of the target talker (Mäkelä et al., 2002; Steinschneider, Nourski, & Fishman, 2013; Weston, Hunter, Sokhi, Wilkinson, & Woodruff, 2015); those features could potentially be 'bound' together by synchronous oscillatory activity (Clayton, Yeung, & Cohen Kadosh, 2015; Engel & Singer, 2001; Singer, 1993). Importantly, the same pattern of RTs for the trial-by-trial analyses were observed on attend-location and attend-gender trials, which is inconsistent with the alternative explanation that participants were *either* using space-based *or* feature-based attention on both types of trial. If participants were directing space-based attention during *both* attend-location and attend-gender trials, then RTs should be affected by the location of the talker on attend-gender trials, but RTs should *not* be affected by the gender of the talker on attend-location trials.

One possible reason why participants may have adopted attention to both location and gender in this task is that the acoustic stimuli were natural speech, which fluctuates in amplitude and frequency over time. When identifying words spoken during multi-talker listening in everyday life, it would be advantageous to monitor multiple attributes of a talker at once rather than focusing only on the talker's location or gender alone; the dynamic nature of speech means that differences in f_0 may best distinguish two talkers at some points in time (e.g. during voiced speech), whereas at other times, differences in spatial location, amplitude, or onset time may be more distinctive cues. In addition, when many talkers speak simultaneously, it may only be through a conjunction of multiple cues (e.g. location and f_0) that a target talker can be distinguished from other competing talkers. Relying on a combination of factors would also be useful when tracking a talker whose f_0 fluctuates over time (e.g. when a talker speaks emotively) or whose location varies over time (e.g. when the talker is moving whilst speaking).

We predicted that longer cue-target intervals might promote object-based attention. However, we found no difference in the trial-by-trial analysis between trials with different length cue-target intervals, thus suggesting that participants directed object-based attention to a target talker even in the 0-ms condition.

Implications

The finding that RTs progressively shortened up to cue-target intervals of 2000 ms has implications for other studies investigating endogenous cueing because previous exogenous visual experiments have typically tested cue-target intervals that are shorter than 1000 ms (e.g., Lu et al., 2009; Yamaguchi et al., 1994). Given there was an improvement in RTs for speech intelligibility in the current experiment between cue-target intervals of 1000 and 2000 ms, there may also be an improvement in visual tasks for cue-target intervals up to (and beyond) 2000 ms. Thus, future experiments examining the time-course of endogenous attention should consider intervals longer than 800 ms. Such studies could also investigate whether endogenous cueing continues to benefit performance at intervals longer than 2000 ms and aim to determine the interval at which performance eventually plateaus or declines.

The current results also have potential implications for research that examines populations for whom multi-talker listening is particularly challenging, including older people (Dubno, Dirks, & Morgan, 1984; Helfer & Freyman, 2008) and people with hearing loss (Gatehouse & Noble, 2004). Best *et al.* (2009) found that hearing-impaired listeners achieved better speech intelligibility when they were cued to the spatial location or time period at which the target talker would speak than when they received no cue. However, the speech intelligibility benefit gained by hearing-impaired listeners was smaller than that found for normally-hearing listeners. Future experiments could investigate whether hearing-impaired listeners gain improvements in speech intelligibility equivalent to that gained by normally-hearing listeners when longer-duration cue-target intervals are used. It is possible that, similar to normally-hearing listeners, longer cue-target intervals could improve speech intelligibility for hearing-impaired listeners. Given that ageing has been associated with a reduction in the speed of processing (Füllgrabe, Moore, & Stone, 2014), using longer cue-target intervals could potentially improve the performance of older adults so that it is equivalent to that of young normally-hearing listeners, when both groups are given sufficient time.

Conclusion

The results demonstrate that longer intervals between an endogenous visual cue for location or gender and the time at which a target talker starts to speak produces better accuracy and shorter-latency responses for reporting key words spoken by a target talker who speaks in a mixture of three talkers. Our main finding was that increasing the duration of preparation time up to 2000 ms progressively improved the latency with which participants correctly reported target words. This result demonstrates that the mechanism underlying preparatory attention unfolds over time—listeners are able to partially prepare for the location or gender of the target talker, with further improvements when longer time is available for preparatory attention to develop. However, listeners may need at least 2000 ms to optimally prepare for talker characteristics during multi-talker listening. This finding demonstrates improvements related to endogenous cueing for cue-target intervals that are longer than those that have previously been tested in the auditory or visual modalities. Thus, to maximise the benefit gained from endogenous cues, future experiments should use intervals up to and greater than 2000 ms.

Acknowledgements

This work was supported by a studentship from the Goodricke Appeal Fund to EH.

References

Anton-Erxleben, K., Stephan, V. M., & Treue, S. (2009). Attention reshapes center-surround receptive field structure in macaque cortical area MT. *Cerebral Cortex*, *19*(10), 2466–78. https://doi.org/10.1093/cercor/bhp002

Best, V., Marrone, N., Mason, C. R., Kidd, G., & Shinn-Cunningham, B. G. (2009). Effects of sensorineural hearing loss on visually guided attention in a multitalker environment. *Journal of the Association for Research in Otolaryngology*, *10*(1), 142–9. https://doi.org/10.1007/s10162-008-0146-7

Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *Journal of the Association for Research in Otolaryngology*, *8*(2), 294–304. https://doi.org/10.1007/s10162-007-0073-z

Brungart, D. S., & Simpson, B. D. (2002). Within-ear and across-ear interference in a cocktail-party listening task. *The Journal of the Acoustical Society of America*, *112*(6), 2985. https://doi.org/10.1121/1.1512703

Brungart, D. S., & Simpson, B. D. (2007). Cocktail party listening in a dynamic multitalker environment. *Perception & Psychophysics*, *69*(1), 79–91.

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–38. https://doi.org/10.1121/1.1408946

Chawla, D., Rees, G., & Friston, K. J. (1999). The physiological basis of attentional modulation in extrastriate visual areas. *Nature Neuroscience*, *2*(7), 671–6. https://doi.org/10.1038/10230

Cheal, M., & Lyon, D. R. (1992). Benefits from attention depend on the target type in location-precued discrimination. *Acta Psychologica*, *81*(3), 243–67. https://doi.org/10.1016/0001-6918(92)90020-E

Chennu, S., Noreika, V., Gueorguiev, D., Blenkmann, A., Kochen, S., Ibáñez, A., … Bekinschtein, T. A. (2013). Expectation and attention in hierarchical auditory prediction. *The Journal of Neuroscience*, *33*(27), 11194–205. https://doi.org/10.1523/JNEUROSCI.0114-13.2013

Clayton, M. S., Yeung, N., & Cohen Kadosh, R. (2015). The roles of cortical oscillations in sustained attention. *Trends in Cognitive Sciences*, *19*(4), 188–95. https://doi.org/10.1016/j.tics.2015.02.004

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, *58*(3), 306–24. https://doi.org/10.1016/j.neuron.2008.04.017

Darwin, C. J. (2006). Contributions of binaural information to the separation of different sound sources. *International Journal of Audiology*, *45*, S20-4. https://doi.org/10.1080/14992020600782592

Dubno, J. R., Dirks, D. D., & Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *The Journal of the Acoustical Society of America*, *76*(1), 87–96.

Duncan, J. (2006). EPS Mid-Career Award 2004: Brain mechanisms of attention. *The Quarterly Journal of Experimental Psychology*, *59*(1), 2–27. https://doi.org/10.1080/17470210500260674

Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences*, *5*(1), 16–25. https://doi.org/10.1016/S1364-6613(00)01568-0

Ericson, M. A., Brungart, D. S., & Brian, D. (2004). Factors that influence intelligibility in multitalker speech displays. *The International Journal of Aviation Psychology*, *14*(3), 313–334.

Fries, P. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, *291*(5508), 1560–63. https://doi.org/10.1126/science.1055465

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, *9*(10), 474–80. https://doi.org/10.1016/j.tics.2005.08.011

Fritz, J. B., Shamma, S. A., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, *6*(11), 1216–23. https://doi.org/10.1038/nn1141

Füllgrabe, C., Moore, B. C. J., & Stone, M. A. (2014). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience, 6*, 347. https://doi.org/10.3389/fnagi.2014.00347

Gatehouse, S., & Noble, W. (2004). The Speech, Spatial and Qualities of Hearing scale (SSQ). *International Journal of Audiology*, *43*, 85–99.

Gazzaley, A., Cooney, J. W., McEvoy, K., Knight, R. T., & D'Esposito, M. (2005). Top-down enhancement and suppression of the magnitude and speed of neural activity. *Journal of Cognitive Neuroscience*, *17*(3), 507–17. https://doi.org/10.1162/0898929053279522

Giesbrecht, B., Weissman, D. H., Woldorff, M. G., & Mangun, G. R. (2006). Pre-target activity in visual cortex predicts behavioral performance on spatial and feature attention tasks. *Brain Research*, *1080*(1), 63–72. https://doi.org/10.1016/j.brainres.2005.09.068

Giesbrecht, B., Woldorff, M. G., Song, A. W., & Mangun, G. R. (2003). Neural mechanisms of top-down control during spatial and feature attention. *NeuroImage*, *19*(3), 496–512. https://doi.org/10.1016/S1053-8119(03)00162-9

Helfer, K. S., & Freyman, R. L. (2008). Aging and speech-on-speech masking. *Ear and Hearing*, *29*(1), 87–98. https://doi.org/10.1097/AUD.0b013e31815d638b.Aging

Hill, K. T., & Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex*, *20*(3), 583–90. https://doi.org/10.1093/cercor/bhp124

Holmes, E., Kitterick, P. T., & Summerfield, A. Q. (2016). EEG activity evoked in preparation for multi-talker listening by adults and children. *Hearing Research*, *336*, 83–100. https://doi.org/10.1016/j.heares.2016.04.007

Holmes, E., Kitterick, P. T., & Summerfield, A. Q. (2017). Peripheral hearing loss reduces the ability of children to direct selective attention during multi-talker listening. Hearing Research, 350, 160–172. https://doi.org/10.1016/j.heares.2017.05.005

Ihlefeld, A., & Shinn-Cunningham, B. G. (2008). Disentangling the effects of spatial cues on selection and formation of auditory objects. *The Journal of the Acoustical Society of America, 124*(4), 2224–35. https://doi.org/10.1121/1.2973185

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in Human Neuroscience*, *4*, 186. https://doi.org/10.3389/fnhum.2010.00186

Jonides, J. (1981). Voluntary versus automatic control over the mind's eye's movement. *Attention and Performance IX*, 187–203.

Kidd, G., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, *118*(6), 3804–15. https://doi.org/10.1121/1.2109187

Kitterick, P. T., Bailey, P. J., & Summerfield, A. Q. (2010). Benefits of knowing who, where, and when in multi-talker listening. *The Journal of the Acoustical Society of America*, *127*(4), 2498–508. https://doi.org/10.1121/1.3327507

Koch, I., Lawo, V., Fels, J., & Vorländer, M. (2011). Switching in the cocktail party: exploring intentional control of auditory selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1140–7. https://doi.org/10.1037/a0022189

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, *320*(5872), 110–3. https://doi.org/10.1126/science.1154735

Larson, E., & Lee, A. K. C. (2013). Influence of preparation time and pitch separation in switching of auditory attention between streams. *The Journal of the Acoustical Society of America*, *134*(2), EL165-71. https://doi.org/10.1121/1.4812439

Lee, A. K. C., Rajaram, S., Xia, J., Bharadwaj, H. M., Larson, E., Hämäläinen, M. S., & Shinn-Cunningham, B. G. (2013). Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. *Frontiers in Neuroscience*, *6*, 1–9. https://doi.org/10.3389/fnins.2012.00190

Lu, Z.-L., Tse, H. C.-H., Dosher, B. A., Lesmes, L. A., Posner, C., & Chu, W. (2009). Intra- and cross-modal cuing of spatial attention: Time courses and mechanisms. *Vision Research*, *49*(10), 1081–96. https://doi.org/10.1016/j.visres.2008.05.021

Lyon, D. R. (1987). *How quickly can attention affect form perception?* University of Dayton Research Institute.

Mäkelä, A. M., Alku, P., Mäkinen, V., Valtonen, J., May, P., & Tiitinen, H. (2002). Human cortical dynamics determined by speech fundamental frequency. *NeuroImage*, *17*(3), 1300–5. https://doi.org/10.1006/nimg.2002.1279

Meiran, N., Chorev, Z., & Sapir, A. (2000). Component processes in task switching. *Cognitive Psychology*, *41*(3), 211–53. https://doi.org/10.1006/cogp.2000.0736

Monsell, S., & Driver, J. (2000). *Control of cognitive processes: Attention and performance XVIII*. (S. Monsell & J. Driver, Eds.). Cambridge, Massechusetts: MIT Press.

Nishida, S., Shibata, T., & Ikeda, K. (2014). Object-based selection modulates top-down attentional shifts. *Frontiers in Human Neuroscience*, *8*, 90. https://doi.org/10.3389/fnhum.2014.00090

O'Connell, M. N., Barczak, A., Ross, D., McGinnis, T., Schroeder, C. E., & Lakatos, P. (2015). Multi-scale entrainment of coupled neuronal oscillations in primary auditory cortex. *Frontiers in Human Neuroscience*, *9*, 655. https://doi.org/10.3389/fnhum.2015.00655

O'Connell, M. N., Barczak, A., Schroeder, C. E., & Lakatos, P. (2014). Layer specific sharpening of frequency tuning by selective attention in primary auditory cortex. *The Journal of Neuroscience*, *34*(49), 16496–508. https://doi.org/10.1523/JNEUROSCI.2055-14.2014

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and Performance X: Control of Language Processes*, *32*, 531–556.

Prinzmetal, W., McCool, C., & Park, S. (2005). Attention: Reaction time and accuracy reveal different mechanisms. *Journal of Experimental Psychology: General, 134*(1), 73–92. https://doi.org/10.1037/0096-3445.134.1.73

Richards, V. M., & Neff, D. L. (2004). Cuing effects for informational masking. *The Journal of the Acoustical Society of America*, *115*(1), 289. https://doi.org/10.1121/1.1631942

Ristic, J., & Kingstone, A. (2006). Attention to arrows: pointing to a new direction. *Quarterly Journal of Experimental Psychology*, *59*(11), 1921–30. https://doi.org/10.1080/17470210500416367

Rockstroh, B., Müller, M., Wagner, M., Cohen, R., & Elbert, T. (1993). "Probing" the nature of the CNV. *Electroencephalography and Clinical Neurophysiology*, *87*(4), 235–41. https://doi.org/10.1016/0013-4694(93)90023-O

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–31.

Ruchkin, D. S., McCalley, M. G., & Glaser, E. M. (1977). Event related potentials and time estimation. *Psychophysiology, 14*(5), 451–455. https://doi.org/10.1111/j.1469-8986.1977.tb01311.x

Salminen, N. H., Tiitinen, H., & May, P. J. (2012). Auditory spatial processing in the human cortex. *The Neuroscientist, 18*(6), 602-12. https://doi.org/10.1177/1073858411434209

Scholl, B. J. (2001). Objects and attention: the state of the art. *Cognition*, *80*(1–2), 1–46.

Seidl, K. N., Peelena, M. V, & Kastner, S. (2012). Neural evidence for distracter suppression during visual search in real-world scenes. *J Neurosci*, *32*(34), 11812–11819. https://doi.org/10.1016/j.pestbp.2011.02.012.Investigations

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–6. https://doi.org/10.1016/j.tics.2008.02.003

Shulman, G. L., Remington, R. W., & McLean, J. P. (1979). Moving attention through visual space. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(3), 522–6.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, *55*, 349–74.

Slagter, H. A., Giesbrecht, B., & Kok, A. (2007). fMRI evidence for both generalized and specialized components of attentional control. *Brain Research*, 90–102. https://doi.org/10.1016/j.brainres.2007.07.097.fMRI

Spence, C. J., & Driver, J. (1998). Auditory and audiovisual inhibition of return. *Perception & Psychophysics*, *60*(1), 125–39.

Steinschneider, M., Nourski, K. V., & Fishman, Y. I. (2013). Representation of speech in human auditory cortex: Is it special? *Hearing Research*, *305*(1), 57–73. https://doi.org/10.1016/j.heares.2013.05.013

Treue, S., & Martinez-Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–9. https://doi.org/10.1038/21176

Tsal, Y. (1983). Movement of attention across the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(4), 523–30.

van Ede, F., De Lange, F. P., & Maris, E. (2012). Attentional cues affect accuracy and reaction time via different cognitive and neural processes. *The Journal of Neuroscience, 32*(30), 10408–12. https://doi.org/10.1523/JNEUROSCI.1337-12.2012

Voisin, J., Bidet-Caulet, A., Bertrand, O., & Fonlupt, P. (2006). Listening in silence activates auditory areas: a functional magnetic resonance imaging study. *The Journal of Neuroscience*, *26*(1), 273–8. https://doi.org/10.1523/JNEUROSCI.2967-05.2006

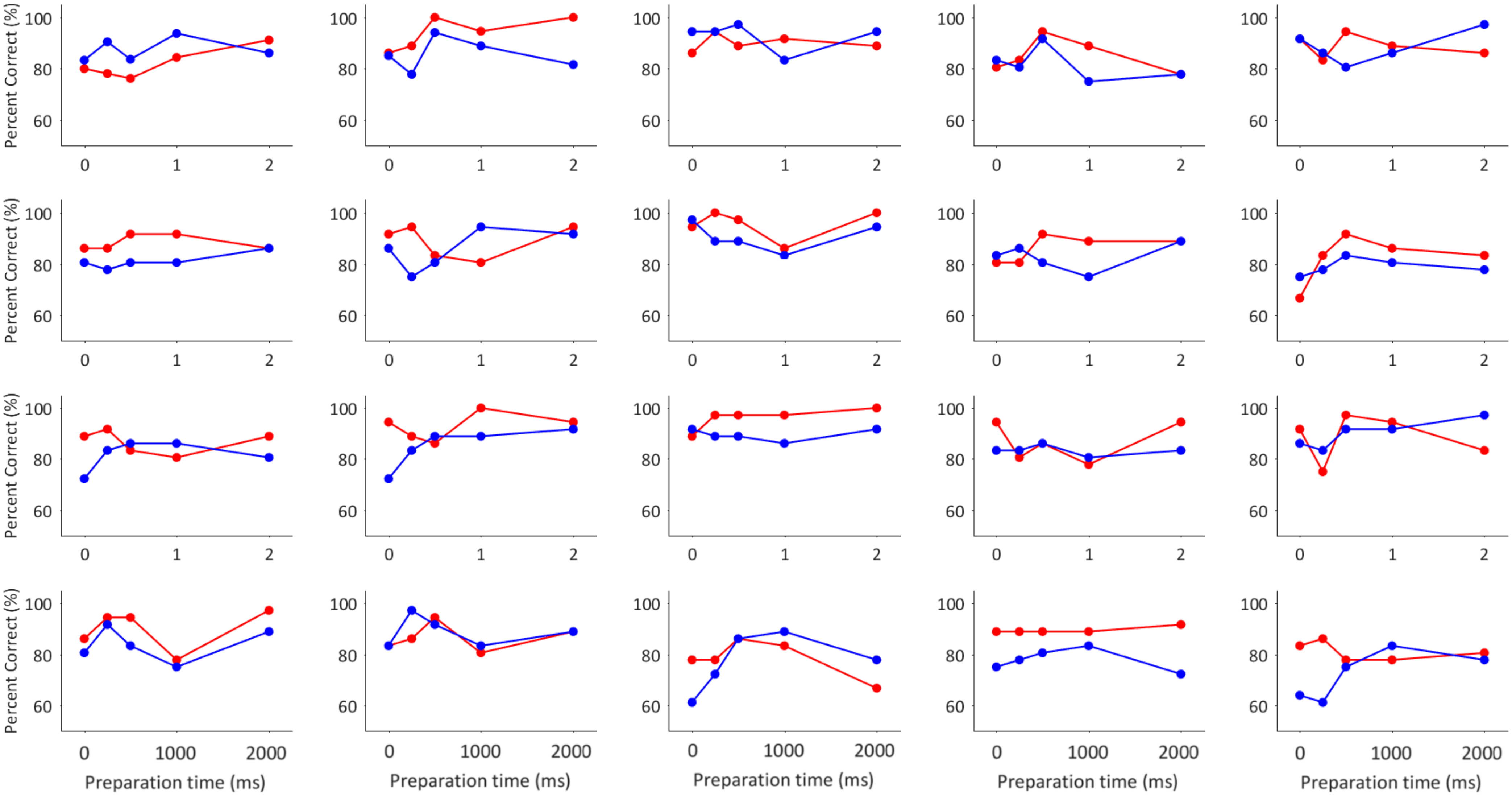
Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., & Winter, A. L. (1964). Contingent negative variation: An electric sign of sensori-motor association and expectancy in the human brain. *Nature*, *203*(4943), 380–4. https://doi.org/10.1038/203380a0

Weston, P. S. J., Hunter, M. D., Sokhi, D. S., Wilkinson, I. D., & Woodruff, P. W. R. (2015). Discrimination of voice gender in the human auditory cortex. *NeuroImage*, *105*, 208–14. https://doi.org/10.1016/j.neuroimage.2014.10.056

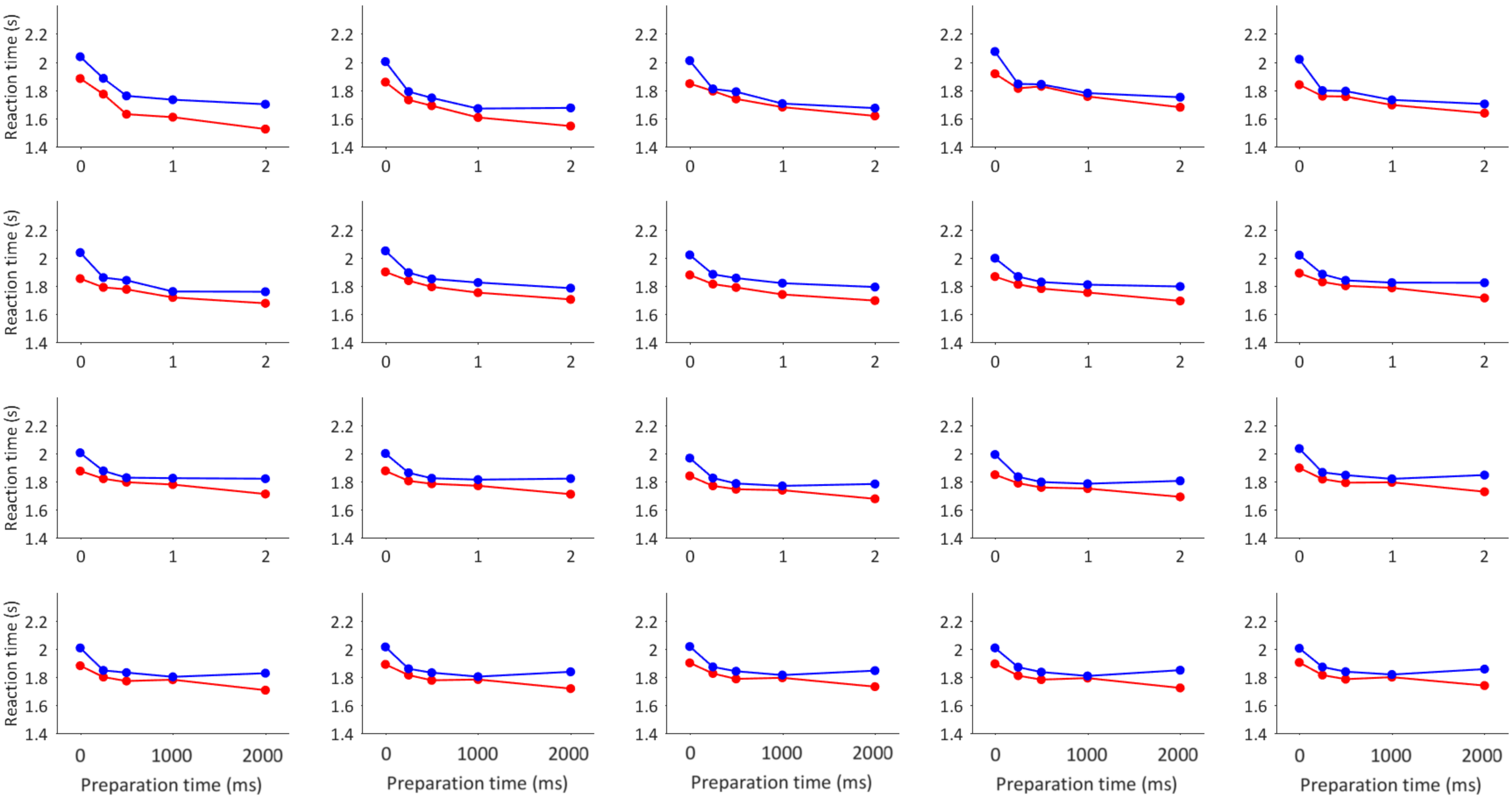
Woldorff, M. G., Hazlett, C. J., Fichtenholtz, H. M., Weissman, D. H., Dale, A. M., & Song, A. W. (2004). Functional parcellation of attentional control regions of the brain. *Journal of Cognitive Neuroscience*, *16*(1), 149–65. https://doi.org/10.1162/089892904322755638

Womelsdorf, T., Anton-Erxleben, K., Pieper, F., & Treue, S. (2006). Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nature Neuroscience*, *9*(9), 1156–60. https://doi.org/10.1038/nn1748

Yamaguchi, S., Tsuchiya, H., & Kobayashi, S. (1994). Electroencephalographic activity associated with shifts of visuospatial attention. *Brain*, *117*, 553–62. https://doi.org/10.1093/brain/117.3.553.



Supplemental Fig. 1. Mean percentage of trials in which participants correctly identified the colour-number combination spoken by the target talker. Each plot displays the results from one participant.



Supplemental Fig. 2. Mean RTs of correct trials, relative to the onset of acoustic stimuli. Each plot displays the results from one participant.