Can we prepare to attend to one of two simultaneous voices?

Stephen Monsell, Aureliu Lavric, Amy Strivens, and Emilia Paul

University of Exeter

Journal of Experimental Psychology: Human Perception and Performance (in press)

[Accepted 20 February 2019]

Running head: Preparation for a voice

Author Note

Stephen Monsell, Aureliu Lavric , Amy Strivens, and Emilia Paul, Psychology, College of Life and Environmental Sciences, University of Exeter.

Experiment 1 was conducted as an undergraduate final year research project by AS and EP under the supervision of SM and AL.

We thank Iring Koch, Eric Ruthruff and two anonymous referees for their comments. The data will be available from Open Research Exeter Repository [URL to follow]. Correspondence concerning this article should be addressed to Stephen Monsell, Psychology, College of Life and Environmental Sciences, Washington Singer Labs, University of Exeter, Exeter EX4 4QG, United Kingdom. E-mail: s.monsell@exeter.ac.uk

Abstract

We can selectively attend to one of two simultaneous voices sharing a source location. Can we endogenously select the voice before speech is heard? Participants heard two digit names, spoken simultaneously by a male and a female voice, following a visual cue indicating which voice's digit to classify as odd or even. There was a substantial cost in RT and errors when the target voice switched from one trial to the next. In Experiment 1, with a highly familiar pair of voices, the switch cost reduced by nearly half as the cue-stimulus interval increased from 50 to 800 ms, indicating (contrary to previous reports) effective endogenous preparation for a change of voice. No further reduction in switch cost occurred with a longer preparation interval — this "residual" switch cost may be attributable to attentional "inertia". In Experiment 2, with previously unfamiliar voices, the pattern of switch costs was very similar, though repeated attention to the same target voice over a run of trials improved performance more. Delaying the onset of one voice by 366 ms improved performance but the pattern of preparatory tuning effects was similar. Thus endogenous preparation for a voice is possible; but it is limited in efficacy, as for some other attentional domains.

Keywords: cocktail party problem, multitalker speech perception, selective attention, endogenous attention, control of attention.

Statement of public significance: We explore the mechanisms by which a person can attend to one of several simultaneous speakers, e.g. in a crowded restaurant. In experiments in which participants are cued to respond to one of two simultaneous speakers, we demonstrate that people can to some extent "tune" their attention to one voice in advance, but optimal tuning is achieved only after several words following a shift of attention to a different voice. This is of interest to those developing hearing aid technologies, or concerned with performance limitations in critical multitalker environments such as air traffic control.

A major starting point for research on attention in the 1950s and 1960s was the "cocktail party problem" identified by Cherry (1953): How can we attend to one speaker in an environment in which two or more people are speaking? Although spatial separation of the voices (e.g. one in each ear) makes selecting one voice much easier, selection remains possible when two voices come from the same location, based on monaural cues such as fundamental frequency, vocal tract size, the prosodic contours of the target and masking speech, accent and speech style (Brungart, Simpson, Ericson & Scott, 2001; Darwin & Hukin, 2000a); such cues are particularly important in realistic resonant environments (Darwin & Hukin, 2000b). It is thus undisputed that, in a multi-talker environment, we can selectively tune attention to a voice *per se*. The question we ask in this paper is: to what extent can we endogenously tune attention to one of two voices (not distinguished by location) *before* either of them speaks?

Around the end of the 1970s, research on attention shifted largely to visual stimuli (driven in part by the much greater control of visual stimuli afforded by the laboratory computers then becoming available than was possible for auditory stimuli presented by a twin-track tape recorder). The exploration of endogenous versus exogenous orientation of covert visual attention became a major research theme thanks to Posner's (1980, 2016) cuing paradigms. Meanwhile, research on understanding speech in a multi-talker environment has continued in the psycho-acoustic research community, focused largely on how properties of the two or more voices heard influence the degree to which they mask or interfere with each other or the ability to separate them as perceptual objects (see Bronkhorst, 2015; Darwin, 2008; Shinn-Cunningham & Best, 2015, for reviews).

Researchers have used a number of measures to assess the ability to attend selectively to one of several simultaneous talkers. One popular paradigm is the "call-sign" (or "coordinate response measure") paradigm, in which participants hear two or more speakers asynchronously saying sentences of the form "Ready Baron, go to green six now". On each trial the participant must listen for a target voice identified by a particular call sign (e.g. "Baron"), and attend to that message, trying to ignore the masking voice or voices saying similar messages (e.g. "Ready Eagle, go to white two now"). Performance is measured by the participant's ability to indicate the appropriate

color-number cell on a matrix. Among the numerous findings from this and other paradigms, the following are relevant to the present study:

- With perceptually dissimilar voices, the majority of errors in the call sign paradigm involve indicating the color, number, or both, spoken by the wrong voice (Brungart, 2001). Hence the difficulty is not principally in identifying speech sounds and words in the two messages, but in tuning attention to the appropriate voice. In the psychoacoustic literature this is referred to as *informational* masking, as distinct from more peripheral *energetic* masking due to the two messages containing energy in the same critical bands, rendering sounds indiscriminable (Bronkhorst, 2015; Brungart, 2001; Darwin, 2008).
- Listeners with normal hearing are capable of segregating two simultaneous talkers even when their spectral profile is quite similar, or the distractor voice is the louder of the two. But it is generally harder to listen selectively to one of two voices the more similar they are. The effective dimensions of similarity include fundamental frequency, vocal tract length, timbre, accent, speech rate and intonation pattern. The first two, on which male and female voices typically differ, are especially important, so that performance is generally better when the target and masker messages are spoken by a man and a woman than by two speakers of the same gender (e.g. Brungart, 2001; Brungart, Simpson, Ericson, & Scott, 2001; Shafiro & Gygi, 2007). But this advantage does not necessarily imply that attention can be tuned to voice gender per se; the effects of artificial manipulation of frequency and vocal tract length suggest that these properties are sufficient to account for the different-gender advantage (Darwin, Brungart, & Simpson, 2003).
- Prior familiarity with a voice helps: listeners are better at avoiding cross-talk between two simultaneous messages when the speaker of the target or the masking message is highly familiar – for example the listener's spouse or a friend (Johnsrude et al., 2013; Souza, Gehani, Wright & McCloy, 2013; Holmes, Domingo & Johnsrude, 2018)
- Advance knowledge of the voice to listen for helps. Freyman, Balakrishnan and Heffer (2004)

had participants report the content words of a nonsense sentence spoken against a background of two or more other speakers asynchronously saying other sentences; preceding this babble with a cue consisting of a partial presentation of the target recording minus the last word helped the participant lock on to the relevant voice and report the last word. In the call-sign paradigm keeping the speaker of the target message consistent over a block of trials improves performance (Brungart et al., 2001, Kitterick, Bailey, & Summerfield, 2010). In these studies the participant was informed of the consistency. Was the benefit of foreknowledge due to the exercise of voluntary attention or to exogenous input having already focused attention on the relevant voice? Bressler, Masud, Bharadwaj, & Shinn-Cunningham (2014) had participants report a sequence of five spoken digit names, each concurrent with three other voices saying a time-reversed digit name. Performance was much better in blocks of trials when the five digits were all said by the same voice, but the same advantage was also seen for an adjacent pair of digits in the same voice in a sequence shifting between voices, when the participant had no expectation that the voice would repeat. In this case the benefit was clearly due to exogenous priming of attention to a voice. Samson and Johnsrude (2016), using the call-sign paradigm, kept the target voice or the masker voice consistent for random sequences of between three and seven trials and found an improvement through a consistent run, though participants were largely unaware of the voice repetitions; again this (especially the effect of masker consistency) seems unlikely to reflect endogenous orienting to a voice.

Endogenous preparation. Our interest is in endogenous, or top-down, orientation of attention to a voice. One approach is to examine voluntary switches between simultaneous speech streams. Larson and Lee (2013) probed the time-course of switching between two simultaneous auditory sequences in the same location. Target and masking sequences were spoken at two different voice pitches (generated from the same female speaker), and consisted of three letter names, a gap of 200-800 ms, then three more letter names; participants responded on the basis of the number of "E"s in the attended sequence, ignoring "E"s in the unattended sequence. They were cued by hearing a sample of the voice to attend to, and told in advance of the whole sequence whether to switch

between the voices or not at the gap. Performance on the first letter name after the gap was worse when a switch was required at all gap durations, and a gap of 400-600 ms was required to achieve optimal performance: there is a substantial cost to switching between voices per se, and it takes a non-trivial amount of time[1]. There was little sign of a reduction in switch cost with more time to switch. Using a MEG version of the same paradigm, with a 600 ms gap, Larson and Lee (2014) detected preparatory brain activity specific to a pitch or location switch as early as the second of the first three letter names, suggesting that top-down control of attention to voice is initiated even while the last words before the gap are still being heard.

Holmes and colleagues (Holmes, 2014; Holmes, Kitterick & Summerfield, 2018) modified the call-sign paradigm to present a visual cue directing attention to the male or female speaker of two briefer spatially separated call-sign messages with a simultaneous onset, varying the interval between the cue and the onset between 0 and 2 sec. With a third masking voice in the centre, they found a systematic benefit of a longer interval to prepare. However, there are at least two possible sources of this benefit other than endogenous attention to a voice per se. One is that with a zero or short interval, processing the cue overlaps with processing the speech. The other is that there may be a generic alerting advantage of a longer preparation interval (as seen in effects on reaction time of a foreperiod following an uninformative warning signal, allowing merely “temporal” orienting of attention, Nobre & Heideman, 2015)[2]. Both sources may be ruled out by the combination of voice-cuing and switching we introduce below.

Hill and Miller (2010) and Lee et al. (2013), using fMRI and MEG respectively, have studied brain correlates of top-down attention by preceding a simultaneous pair of spoken inputs differing in both location and pitch with a cue specifying the pitch or location of the input to respond to: a fronto-parietal network is activated by cues orienting attention by pitch and by location, much as in studies of visual endogenous orienting, albeit with some regions more active for pitch than for

[1] Broadbent's (1954) classical "split span" experiments indicated a similar conclusion for switching between speech streams in different locations.
[2] Holmes et al. (2018) kept the interval between trial onset (signalled by a composite visual cue) and speech onset constant, but varied the time at which the composite cue revealed the voice cue randomly from trial to trial; this cue "reveal" rather than the trial onset may have been a trigger for phasic alertness.

location selection and vice versa. But the functional benefit of this activity is unclear.

Cuing and switching voices. In the experiments we report in this article, we measure the benefit of preparation for a voice using a paradigm that combines cuing of the relevant voice with an assessment of the cost of switching voices from trial to trial. The participant hears on each trial just two simultaneous words, spoken by male and female speakers with the same apparent location; a cue to attend to one of them precedes the speech onset by a variable interval; the voice to attend to may or may not switch from trial to trial. Our experiments were prompted by a series of studies using a similar paradigm by Koch and colleagues, starting with Koch, Lawo, Fels and Vorländer (2011), which we review in detail below. Their experiments demonstrated substantial costs of switching attention between voices from trial to trial, but surprisingly, in most cases, their participants showed no sign of taking advantage of a long cue-stimulus interval to reduce the switch cost by shifting attention endogenously to the cued voice. Our experiments, in contrast, show that there are straightforward conditions under which preparation for a voice switch is effective.

The voice-cuing paradigm is modeled on the task-cuing reaction time (RT) paradigm widely used in research on task-set control (Meiran, 1996, 2000, 2014; Monsell & Mizon, 2006). In a typical task-cuing experiment, on each trial a stimulus is presented that affords two (or more) simple tasks with which the participant has been familiarized during practice. A cue preceding the stimulus tells the participant which task to perform. As one example, the stimulus might be a digit and the task is to classify it as odd/even or high/low in value (with a left/right key press). As another, the stimulus might be a colored shape and the participant must identify its color or shape with a key press (Monsell & Mizon, 2006). Variation of the cue-stimulus interval (CSI) affords more or less time for preparation following the cue. Mean RT is longer, and error rates usually higher, on trials when the task changes than when it repeats: there is a "switch cost". As the CSI increases from zero to somewhere between 0.5 and 1 s, there is a reduction in switch cost (a "RISC effect"), but a further increase in CSI almost never reduces the switch cost to zero: there is an asymptotic or "residual" cost. A standard interpretation of the RISC effect is that it indexes endogenous task-set preparation, whilst the residual switch cost represents some sort of competition

from and/or inertia of the previous task-sets that endogenous preparation cannot, or cannot always, overcome. (See Kiesel et al, 2010; Koch, Poljac, Müller, and Kiesel, 2018; Monsell, 2003, 2015, 2017; Vandierendonck et al, 2010 for reviews; we elaborate these theoretical interpretations in the General Discussion.) A merit of the paradigm is that any generic "warning signal" effect of the CSI applies equally to switch and repeat trials, so that the RISC effect partials out the effect of selective preparation for one task or, in the present case, one voice, from generic effects of temporal warning. Another effect of CSI on RT arises from the fact that at very short CSIs, cue processing necessarily overlaps with stimulus processing. But the effect of this overlap is the same for cues signaling task switches and repeats (at least if the cue changes on every trial – see below), and is partialled out by examining the effect of CSI on the switch cost – the RISC effect.

To adapt the task-cuing paradigm to explore advance preparation for a voice, one can simply present on each trial two voices simultaneously saying a different word – for example, different digit names. The cue specifies which voice to attend to, and the participant must rapidly classify its digit (e.g., as odd/even). This classification task remains constant, and only the relevant voice changes or repeats from trial to trial. This is essentially a task-cuing experiment in which the only task parameter that changes is an attentional parameter – which voice to attend to. (In the General Discussion we review cuing studies in which other attentional parameters are analogously cued.) Koch and colleagues have reported a substantial series of voice-cuing experiments, in which they preceded a dichotic simultaneous digit pair (spoken by a male voice on one ear, and a female voice on the other, with the assignment to ears random on each trial) with a visual or auditory gender or ear cue that specified which voice to respond to (Koch, Lawo, Fels, & Vorländer, 2011; Lawo, Fels, Oberem, & Koch, 2014; Lawo & Koch, 2014, 2015; Nolden, Ibrahim & Koch, 2018, Seibold, Nolden, Oberem, Fels, & Koch, 2018). In most of their experiments the task was classification by magnitude (<5 versus >5) with a vocal or manual response; in others the task was shadowing. The CSI was varied to contrast performance at a short and a long preparation interval, usually 100 ms versus 900/1000 ms., (but 400 versus 1200 ms in Nolden, Ibrahim & Koch, 2018). In some experiments the CSI and the criterion for selecting the target voice, gender or ear, varied from one

trial to another; in others the CSI and criterion were constant within a block of trials and alternated over blocks. Koch and colleagues obtained a substantial and consistent cost of a voice switch, and a reduction in RT at the longer interval, but, surprisingly, little consistent evidence for a reduction in switch cost as the preparation interval increased. As Nolden et al (2018, p.2) summarize the implication of previous results from their lab, "participants did not seem to prepare actively for shifting auditory attention."; the marked reduction in overall RT at the longer interval is attributed to other general preparatory mechanisms. The exceptions — experiments where some evidence for a reliable RISC effect was detected — include: experiments (Koch et al, 2011, Experiment 2, Lawo & Koch, 2014, Nolden et al, 2018, Experiment 2) which must be discounted because each voice was cued by only one cue, so that cue change was confounded with voice change[3]; an experiment by Lawo and Koch (2015, Experiment 2) in which there was no overall significant RISC effect, but the switch cost reduced with increasing CSI for transparent verbal responses, but increased for arbitrary verbal responses, leading to a significant 3 way interaction; Seibold et al.'s (2018) Experiment 3, in which the participant was instructed to switch voices every two trials in a sequence of eight 'alternating runs' (cf. Rogers & Monsell, 1995) with a response-stimulus interval of 100 versus 1000 ms; even then the RISC effect was small. Does the lack of a reliable RISC effect in the many other cuing experiments of Koch and colleagues mean that participants cannot tune attention to a voice in advance, or find this very difficult, except perhaps when the changes in the target speaker are completely predictable relatively far ahead?

Several features of these studies might have mitigated against participants exercising their ability to prepare for a voice:

- Except for Seibold et al. (2018), the male or female voice presented on each trial in their studies was sampled randomly from a small set of male or female voices. Hence the cue specified voice gender rather than a specific person's voice. In an EEG study similar to Holmes et al. (2018), Holmes, Kitterick and Summerfield (2016) found preparatory EEG activity when participants

[3] In task-switching experiments, cue repetition has been shown to facilitate performance even when the task repeats, and this cue priming effect may reduce as the cue-stimulus interval increases (e.g. Monsell & Mizon, 2006). As noted later, the confound is avoided by using two cues per voice, as in most of Koch et al's other experiments.

were cued by the gender of the voice, but only when a single male or female voice was used, not when the cued voice was sampled from several of the same gender. Although, as already mentioned, a difference in gender usually makes simultaneous voices easier to select, we normally deploy our attention to voices, for example in a noisy restaurant, to select a specific person's voice, not a class of voice defined by gender or the ear in which we will hear the voice. Hence in the present experiments participants were cued to attend to either of just two speakers, male and female, the same two speakers throughout the experiment.

- Koch et al.'s male and female voices were presented dichotically, randomly assigned one to each ear. Separation by location provides a strong basis for auditory selective attention. Perhaps the possibility, or indeed necessity, of attending left or right when the stimuli were heard discouraged anticipatory deployment of attention to a voice per se. In the present experiments, we presented the voices binaurally, so that they were both heard in the same central apparent location.
- The perceptual attributes associated with a particular voice are complex, and it may require considerable familiarity with a voice to tune attention efficiently in advance of the stimulus. Familiarity with the voice improves spoken word recognition (Nygaard, Sommers and Pisoni, 1994; Nygaard and Pisoni, 1998; Yonan and Sommers, 2000) and, as already mentioned, familiarity with the target or masking voice improves speech segregation. In Experiment 1 we therefore maximized the familiarity of the voices by using as speakers for each participant their parents. In Experiment 2 we checked whether the results depended on so high a level of familiarity.

Work on auditory stream segregation (or "auditory scene analysis") indicates that listeners tend initially to group as a single auditory object sounds with a simultaneous onset (Carlyon, 2004), Such grouping could impair tuning of attention to one of two speech tokens with precisely simultaneous onsets, which are relatively rare in real-world scenarios. To examine the importance of this factor, we included in Experiment 1 a comparison between a condition with simultaneous

onsets of the two digit names to a condition in which one onset lagged about a third of a second behind the other[4]. This 'successive' condition also raises the question: if the onset of the voice to which one is trying to attend is preceded by the occurrence of the other voice, does that exogenous input nullify or reduce any benefit of endogenous preparation?

One merit of adapting the well-studied task-cuing paradigm to study voice-cuing is that we take advantage of methodological refinements that have emerged from two decades of research with the paradigm (Meiran, 2014; Monsell & Mizon, 2006). First, if just one cue is used per task, then task switch/repeat is confounded with cue change/repeat. Immediate cue repetitions turned out to facilitate performance even when the task does not change. Our solution is to deploy two cues per voice and change the cue on every trial. Second, to avoid a confound between the time available for active endogenous preparation for a voice with the interval elapsed since the previous trial (which might determine passive dissipation of the previous attentional setting), we kept the response-stimulus interval constant while varying the CSI. Third, to discourage a tendency for participants to prepare for a switch before the cue when the switch probability is as high as 0.5, we used a switch probability of 0.33. Fourth, to capture not only the reduction in switch cost as CSI increases, but also its asymptote, it is advisable to use at least three CSIs, suitably spaced. Finally, endogenous preparation, if possible, is optional: participants do not need to prepare in advance, beyond encoding the cue, to accomplish the task. Hence it is important to motivate participants to minimize RT and maximize accuracy through advance preparation, and we used performance bonuses to encourage this.

Experiment 1

Student participants heard on each trial the names of two different digits, one spoken by their mother and one by their father. A visual cue – the name (e.g. "Dad") or a photo of one parent – preceded the speech onset by a CSI of 50, 800, or 1400 ms, to tell the participant which voice to respond to. They pressed a key to indicate whether the digit spoken in the cued voice was odd or

[4] Nolden Ibrahim and Koch (2018) have now reported a similar manipulation; we summarize their findings in the Experiment 1 Discussion.

even. To motivate selective attention the majority of the digit pairs were response-incongruent, so that attending to the wrong voice would lead to the wrong response. But a pilot experiment had revealed that if *all* the pairs were response-incongruent, some participants discovered the clever, if baroque, strategy of listening always to one voice and reversing the response rule when the other voice was cued. To prevent this, on 20% of trials the digit pair was response-congruent (both spoken digits were odd, or both were even).

Method

Participants. Twenty-four University of Exeter and Exeter College students, 5 men and 19 women, aged between 17 and 25 (M=20.75), participated in two sessions, for which they were paid £15 plus performance bonuses. Two were replaced due to very long mean RTs. All participants had a male and female parent who talked to them frequently and were willing to have their voices recorded. The study adhered to the guidelines of, and was approved by, the local Ethics Committee (Department of Psychology, College of Life and Environmental Sciences, University of Exeter).

To determine the adequacy of this sample size (N = 24) for detecting a RISC effect for voices with magnitude and variability similar to those generally obtained for task cuing using similar methods, we performed power analyses in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) based on effect sizes in ten task cuing experiments conducted in our laboratory over the last decade (see Supplementary Materials for details). The G*Power 'a priori' procedure (estimating N given the expected effect size, at the required α and power thresholds) yielded, for observing the range of effect sizes $0.313 \leq \eta_p^2 \leq 0.815$ (observed in our previous experiments), for $\alpha \leq 0.05$, a mean estimate of N = 9.7 (median 10, range 5-14) required to achieve power ≥ 0.8; and N = 11.9 (median 12, range 6-12) to achieve power ≥ 0.9. In addition, according to the G*Power 'post hoc' procedure (which estimates achievable power given N, the expected effect size and α) the mean achievable power based on the effect sizes in our previous 10 experiments given N = 24 (and $\alpha \leq 0.05$) was estimated to be 0.996 (median 0.999, range 0.984-1.000).

Stimulus materials. The participants' parents were digitally recorded in a quiet room saying the names of the digits 2 through 9 several times. They were played a recording illustrating how to

pronounce each digit briskly and with neutral declarative intonation. The clearest token of each name was selected, extracted, and saved as a .wav file using Praat (Boersma & Weeninck, 2014), adjusting amplitudes if necessary so that the subjective amplitudes approximately matched across the two parents' recordings. The mean duration of the spoken word varied over voice pairs between 310 and 595 ms (M = 410 ms, SD = 63 ms). The mean and standard deviation of the difference in F_0 between mother and father in the voice pairs, measured using Praat, was 8.8 ± 3.2 semitones.

Participants provided a photo of each parent, which was cropped to a head-and-shoulders shot with a width on screen of 5.3 cm and a height of between 5 and 8 cm. Verbal cues were also created for each participant, using the name they customarily use to address or refer to each parent (e.g., "Ma", "Mummy", "Dad"). These were displayed in Arial font, with the first letter in upper case; they were 3.1-7.1 cm wide and 1.2-2.4 cm tall (with a larger font for the shorter words to minimize differences in width).

Procedure. Participants were seated in a sound-attenuating booth, wearing headphones, with their eyes approximately 60 cm from the screen, and their index fingers on the *v* and *m* keys on the computer keyboard. The first session began with 32 trials of practice on the odd/even task with just one spoken digit name per trial; a photo or name cue was followed by a digit name spoken by that parent, and the participant pressed the left key for an odd digit, and the right key for an even digit. Each combination of cue-type, parent and digit was presented once.

In the rest of each session, a visual cue (photo or name) was displayed on each trial for a CSI of 50, 800 or 1400 ms before the participant heard a digit name spoken by each parent, played to both ears over earphones so as to appear to come from one central source. In one session, the speech onsets were simultaneous. In the other they were successive, with the second onset 366 ms after the first, so that there was an overlap of 60 ms on average, but with considerable variability over voice pairs and digit names. CSI remained constant within a block, and the interval between the previous response and the onset of speech was 2.2 s unless the response was incorrect, in which case the word "ERROR" was shown for an extra 2 s, or unless the participant failed to respond

within 4 sec (counted from the onset of the second voice in the Successive condition), in which case "NO RESPONSE" was displayed for an extra 2 s.

In each of the two sessions (one for the Simultaneous, one for the Successive condition, order balanced over participants) there were 3 practice blocks of 24 trials (one for each CSI, in descending order) followed by 12 experimental blocks of 61 trials (of which the first was a warmup trial excluded from analysis). In the first three experimental blocks, one of the six possible permutations of the order of three CSIs was used (balanced over participants) and then replicated three times over the remaining blocks (4-12). Participants were informed that the voice would change from trial to trial on only one third of the trials, and encouraged to respond as quickly as possible while avoiding errors. After each experimental block they saw their mean correct RT and error rate, and a score (RT/10 + 5 per error), receiving a bonus of £0.30 when the score was lower than their previous average for that CSI.

Design. On the first trial of each block, the cue type (photo, name) was chosen randomly and then alternated from trial to trial, so that there were no immediate cue repetitions. Trial sequences were randomized for every participant such that the voice switched on one third of the trials. On 80% of the trials a response-incongruent digit pair was presented (one odd, one even), and the remaining 20% trials were equally divided among the congruent even-even and the congruent odd-odd digit pairs.

In the Simultaneous condition, for each CSI, each voice was cued twice on a repeat trial and once on a switch trial for each of the 32 incongruent pairs of male-and female-voiced digits – this meant that the incongruent digit pairs were represented exactly equally over the combinations of relevant voice (male/female) x CSI x switch/repeat, but randomly combined with cue type. For the (much less numerous) congruent digit pairs, the 8 digits that could be spoken by the relevant voice were used equally in the combinations of all design factors (except cue), whereas the digit spoken by the irrelevant voice was sampled randomly among the set of 4 possible even or 4 possible odd digits.

The Successive condition required the additional factor of which voice occurred first, hence the number of trials per design cell was half of that in the Simultaneous condition. Here the 32 incongruent digit pairs were represented equally over the factor combinations (except with cue type) for the repeat trials. For switches, the eight digits spoken by the relevant voice were represented equally over the factor combinations, whereas the digit spoken by the irrelevant voice was sampled such that for every two participants the 32 incongruent digit pairs were represented equally over the factor combinations. For the congruent digit pairs, on repeat trials the 8 digits that could be spoken by the relevant speaker were used equally across the factor combinations, whereas the digit spoken by the irrelevant voice was sampled randomly; for switches both the relevant and irrelevant digits were sampled randomly.

Results

As explained in the Introduction, we included response-congruent trials merely to prevent the adoption of a strategy of attending consistently to just one voice. We thus present results only from the trials with response-incongruent digit pairs (80% of the total) for which the design was fully balanced, and for which attending to the wrong voice would result in an error. (The essential data pattern was very similar if the congruent trials were included.) We present data for mean RT for correct responses following correct responses (excluding RTs over 3000 ms: 0.39% of the trials for the Simultaneous presentation session, and 0.12% for the Successive presentation session), and error rate, averaged over cue type. Equivalent analyses for median correct RT yielded similar results. In each condition, mean correct RTs and error rates were submitted to ANOVAs with factors switch/repeat, CSI and cue-type (name versus photo). The Huyn-Feldt correction for sphericity was used but degrees of freedom are reported uncorrected. Although we included cue type (name versus photo) as a factor in the ANOVAs, this variable was manipulated only to avoid cue repetitions, and there was no obvious reason to expect interesting effects of this factor. For simplicity we therefore omit effects of and interactions involving cue type from the Results sections, but report and discuss them for completeness in Appendix B.

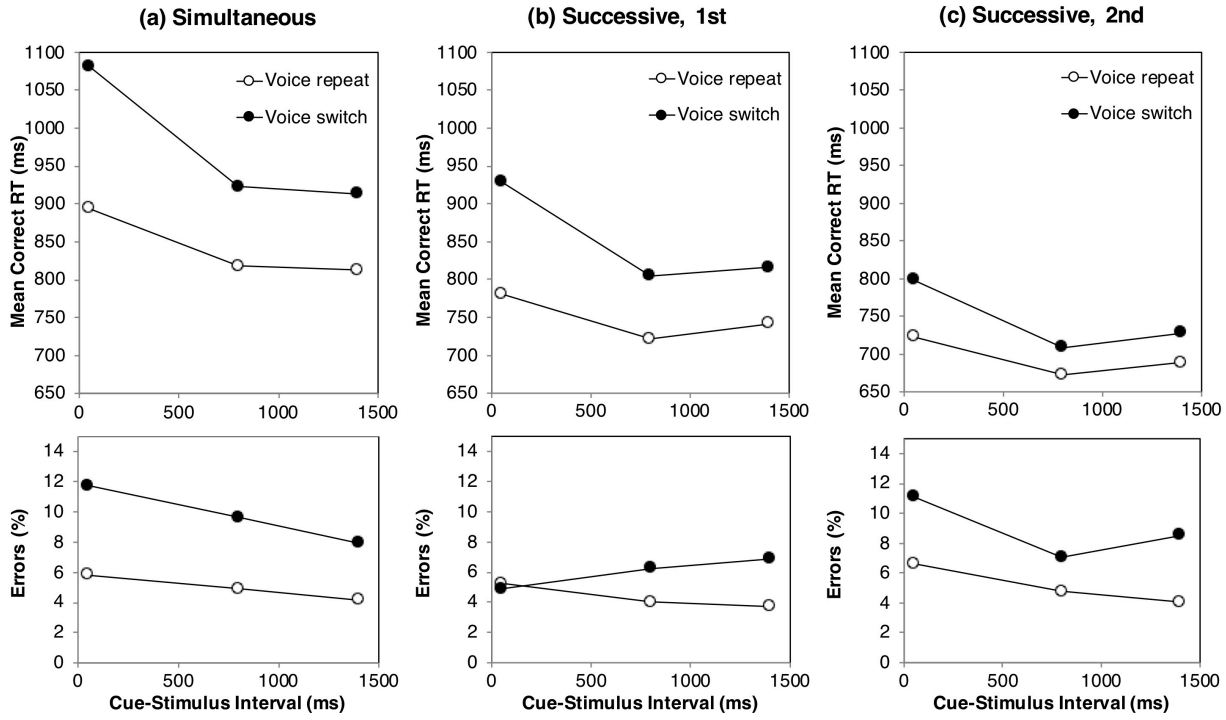

Figure 1: Mean correct RT and error rate for Experiment 1. (a) Simultaneous condition; (b) Successive condition (366 ms onset asynchrony), first voice cued; (c) Successive condition, second voice cued.

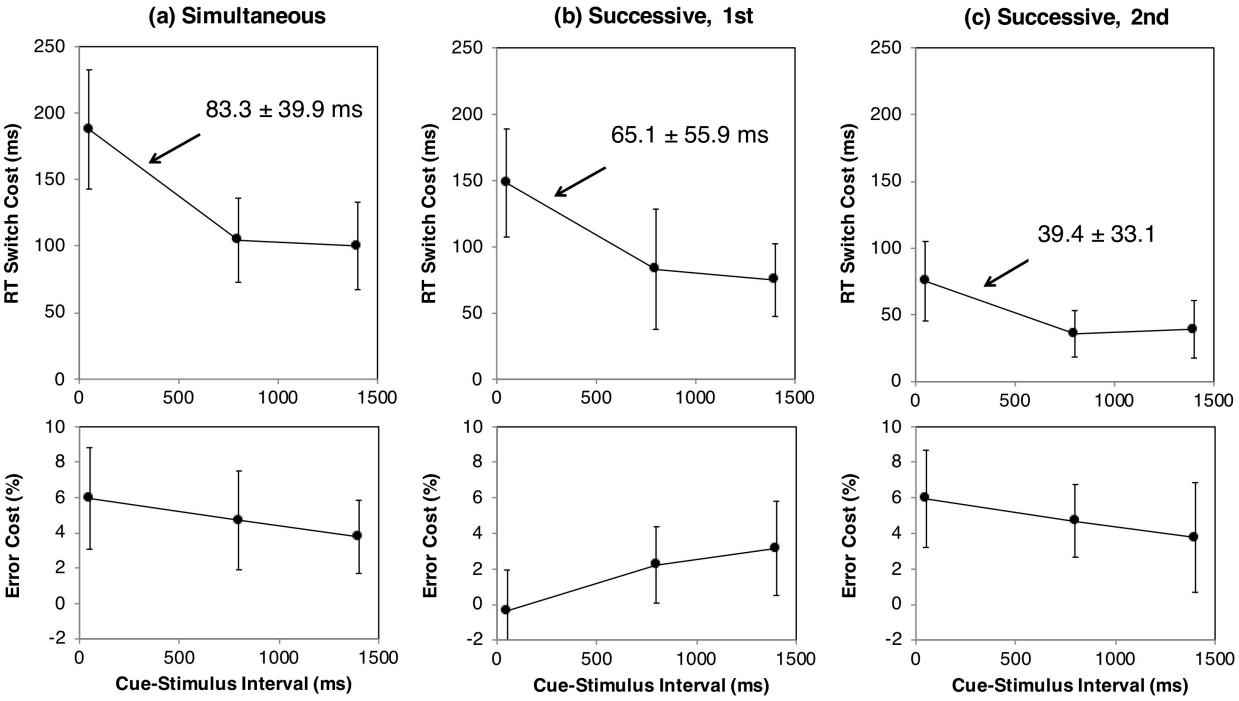

Figure 2. Switch costs (switch minus repeat) and their 95% confidence intervals, from the values plotted in Figure 1 for (a) Simultaneous condition; (b) Successive condition, first voice cued; (c) Successive condition, second voice cued. The reduction in RT switch cost from CSI=50 to CSI=800 ms and its 95% confidence interval are also shown.

Figure 1 shows performance as a function of voice switch/repeat and CSI. Figure 2 shows the switch costs obtained by subtracting repeat from switch RT and errors, and 95% confidence intervals for the effect of a voice switch at each CSI, and of the reduction in switch cost given 800 ms rather than 50 ms to prepare.

Simultaneous presentation. As may be seen in Figure 1a, with no time to prepare, mean RT on voice repeat trials was ~900 ms. When the voice switched, RT was longer by about 190 ms. This substantial voice switch cost was reduced when the participant had 800 ms to prepare. The reduction was a significant 83.3 ms (CI: ± 39.9 ms) – a reduction of 44%. Extending the preparation time to 1400 ms reduced the mean switch cost further by only 4.3 ms (CI: ± 35.8 ms), indicating that the switch cost was at or near asymptote by a CSI of 800 ms. The CIs for the longer CSIs in Figure 2a indicate a highly significant "residual" cost. The overall switch x CSI interaction was highly significant, $F(2, 46) = 14.50$, $p < 0.001$, $\eta_p^2 = 0.387$. The error rates in Fig 1a and 2a also indicate significant switch costs at each CSI, but only a modest decrease in switch cost with increasing CSI; the interaction was not reliable, $F<1$.

Successive presentation, cued voice first. Figures 1b and 2b show equivalent plots for trials from the Successive condition in which the cued voice occurred first (based on only half the number of trials available in the simultaneous condition.) We might expect responding to the cued voice to be easier when its onset is heard in isolation and indeed that was the case. In an ANOVA comparing this to the Simultaneous condition, RTs were significantly shorter, $F(1,23) = 12.68$, $p = 0.002$, $\eta_p^2 = 0.355$, and error rates were lower $F(1,23) = 6.91$, $p = 0.015$, $\eta_p^2 = 0.231$. The switch cost was also smaller for both RT, $F(1,23) = 3.64$, $p=0.069$, $\eta_p^2 = 0.137$, and error rate, $F(1,23) = 9,73$, $p = 0.005$, $\eta_p^2 = 0.297$. But the effect of increasing CSI on the switch cost was very similar with simultaneous and successive presentation: the three way interaction between condition, switch and CSI was not reliable for RT, $F<1$, and only approached significance for error rate, $p =.095$.

With successive presentation, the switch cost reduced as preparation time increased to 800 ms, and the reduction, although smaller absolutely (65.1 ms; CI: ± 55.9) than with simultaneous

presentation, was still significant, and the proportionate reduction was almost identical at 44%. The switch cost again was near-asymptotic by CSI=800 ms, with a further reduction of only 8.2 ms (CI: ±36.3) as CSI increased to 1400 ms, and the CIs in Figure 2b indicate a robust residual switch cost. In ANOVAs on this condition alone (Successive, cued voice first), the CSI x switch interaction was significant for RTs, $F(2,46) = 6.60$, $p = 0.005$, $\eta_p^2 = 0.223$, whilst the apparent increase in error cost with CSI was not significant, $F(2,46) = 1.95$.

Successive presentation, cued voice second. When the cued voice occurs second, it is less obvious what to expect. One might expect that, when one is trying to attend to cued voice A, hearing 366 ms worth of voice B before the onset of voice A would nullify or at least reduce the efficacy of preparation for voice A. On the other hand, if preparation for voice A can continue while voice B is heard (unlikely, perhaps) then there is an extra 366 ms for preparation. But in fact the critical CSI x switch interaction seen in Figures 1c and 2c, when the cued voice came second, was similar to that in Figures 1b and 2b when the cued voice came first, albeit compressed in scale. As CSI increased from 50 to 800 ms, the switch cost still reduced by 39.4 ms (CI: ±33.1) ms, a proportionate reduction of 52%, with no further reduction -3.4 (CI: ±32.2) at CSI=1400 ms. The switch x CSI interaction was reliable for RT, $F(2,46) = 3.82$, $p = 0.029$, $\eta_p^2 = 0.142$, but not errors, $F<1$. An ANOVA with first/second voice x switch/repeat x CSI shows a shorter RT, $F(1,23) = 68.3$, $p < 0.001$, $\eta_p^2 = 0.748$, and a smaller RT switch cost, $F(1,23) = 15.89$, $p = 0.001$, $\eta_p^2 = 0.409$, when the cued voice came second, but no significant three way interaction, $F<1$. The cued voice coming second rather than first also reduced the error rate, $F(1,23) = 8.60$, $p = 0.007$, $\eta_p^2 = 0.272$, and the error switch cost, $F(1,23) = 7.51$, $p = 0.012$, $\eta_p^2 = 0.246$, but the three-way interaction was not significant, $F(2,46) = 2.02$, $p = 0.144$, $\eta_p^2 = 0.081$.

Discussion

Switching voices from trial to trial incurred a substantial switch cost in RT and errors. Allowing 800 ms for preparation between cue and speech onset reliably reduced that switch cost by nearly half. Delaying the onset of one of the two voices made the task easier and reduced the switch cost somewhat, but the proportionate reduction in switch cost was very similar with simultaneous

and successive preparation, even when the cued voice was heard second, suggesting that hearing the distractor voice first did not exogenously undo the benefit of exogenous preparation for a switch indexed by the RISC effect. Preparation did not eliminate the residual switch cost whether the cued voice occurred first or second.

Nolden, Ibrahim & Koch (2018) recently reported a similar manipulation of voice onset asychnrony, with CSIs of 400 and 1200 ms, one auditory tone cue per voice, and the cued voice occurring 200 ms before or after the distractor voice. Although they again found little evidence for a RISC effect[5], the marked reduction in overall RT at the longer interval was significantly larger when the distractor was the first item heard, leading them to speculate that general preparation was more efficient for distractor suppression than for target enhancement. In our data, however, the overall reduction in RT was smaller (by 24 ms, F=1.1) when the distractor was heard first, though the error rate showed a greater overall reduction (by 2.6%, F=2.779, p= .07).

Experiment 2

The primary purpose of Experiment 1 was to address the conflict between our intuition that people can prepare to attend to a specific voice and Koch et al.'s finding of no consistent reduction in switch cost in their experiments. We found clear evidence that participants used a long preparation interval to shift attention between the voices. Among several possible reasons why their experiments did not, we speculated that preparation for a voice might require considerable familiarity with that voice. We therefore tested our participants on very familiar voices – those of their parents. But of course our experiment differed in other ways from those of Koch et al. To determine whether familiarity was indeed critical we repeated the Simultaneous condition of Experiment 1 using participants who lacked prior familiarity with the voices.

[5] In the task-switching literature, a CSI of the order of 500-700 ms is typically sufficient for the RISC effect to reach asymptote, so 400 ms. may be too long for an interval allowing little or no preparation.

Figure 3: Iconic cues for male and female speakers used in Exp 2.

Method

Two changes were necessary. With participants who did not already know the speakers or their voices, we could not use as cues photos or names such as “Dad”. Hence, we substituted the verbal cues “male” and “female” (in Arial font, 2.5 cm x 0.8 cm and 3.4 cm x 0.8 cm, respectively), and a male silhouette (3.2 cm x 2.9 cm), and a female silhouette (2.7 cm x 2.9 cm) as shown in Figure 3. Second, the stimuli recorded for Experiment 1 from the 24 parental couples varied in quality, and exhibited some idiolectal and dialectal departures from standard southern British English. Such departures from clear and/or standard pronunciation, although unproblematic for listeners used to them, might cause difficulties for listeners unfamiliar with the speakers. We therefore selected from the original set of stimuli those spoken by the 12 men and 12 women with the clearest and most standard pronunciations; eight male and female speakers were paired in the same way as in Experiment 1, and another four were re-coupled, with some adjustments in amplitude to keep the subjective amplitudes similar across a pair. Each pair of voices was used for two participants. Mean separation in F_0 of the subset of male and female voices was 9 semitones (SD = 3.25), almost identical to the whole set in Experiment 1, as was the mean duration of the words (M = 417 ms, SD = 36 ms). In all other respects, the experiment replicated the Simultaneous condition of Experiment 1 with a new set of 24 participants recruited from the same population (10 men and 14 women, aged between 17 and 34; M = 20.9) who had no prior familiarity with the speakers. The power analyses for Experiment 1 suggested, and its results confirm, that the N of 24

is more than adequate for detecting RISC effects like those in Experiment 1 and our previous task-cuing experiments.

Results

The data were analyzed as for Experiment 1, with 0.14% of correct RTs over 3 sec and therefore excluded. Figure 4 shows the mean correct RTs and error rates from Experiment 2 plotted for comparison side by side with those from the simultaneous onsets condition of Experiment 1. Figure 5 shows the switch costs.

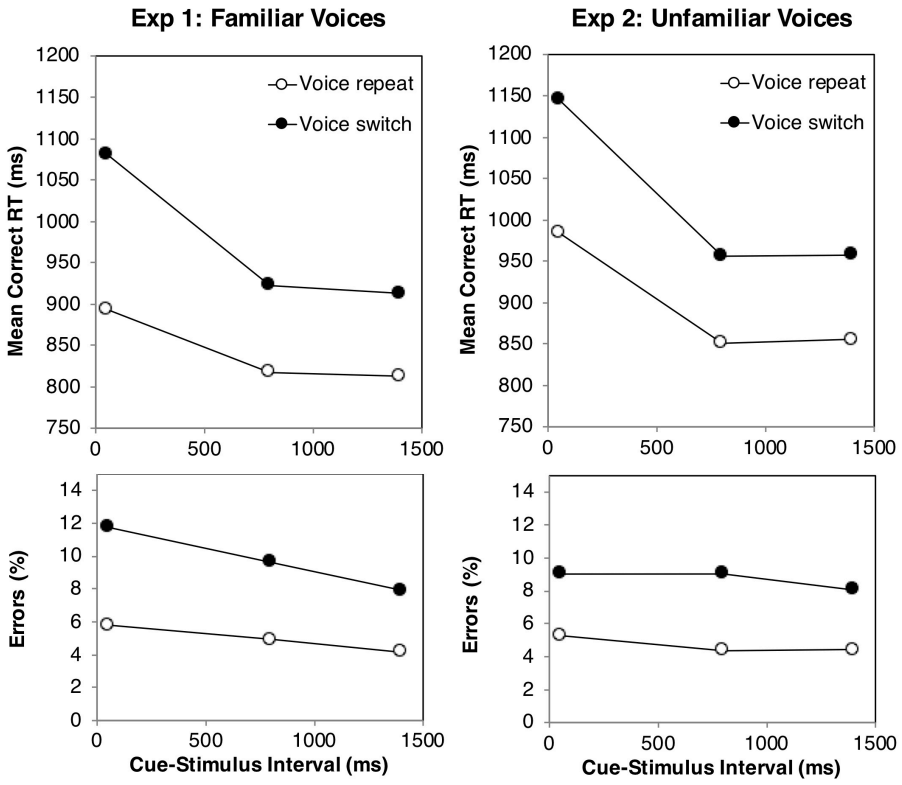

Figure 4: Mean correct RT and error rate for Experiment 1, simultaneous condition (left panels) and Experiment 2 (right panels).

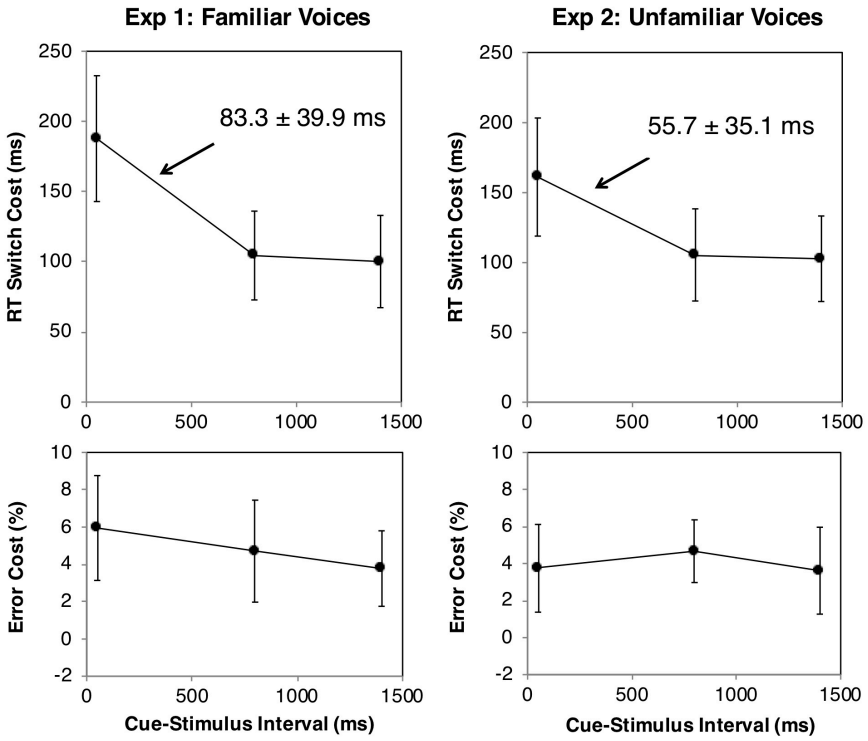

Figure 5. Switch costs (switch minus repeat) from the values plotted in Figure 4 for the simultaneous condition of Experiment 1 (left panels), and for Experiment 2 (right panels).

The data from Experiment 2 (unfamiliar voices) look very similar to those from Experiment 1 (familiar voices). The RTs were a little longer in Experiment 2, but not reliably so, $F = 1.16$, and the error rates very slightly lower, $F<1$. More important, the voice switch costs were comparable. In Experiment 2 participants showed a substantial voice switch cost of about 160 ms at CSI=50 ms, which reduced by a highly significant 55.7 ms (CI: ±35.1 ms) as the CSI increased to 800 ms, with only a trivial further reduction of 2.7 ms (CI: ± 29.0)[6] at CSI=1400 ms. The overall switch x CSI interaction was highly significant for RT, $F(2,46) = 6.50$, $p= 0.007$, $\eta_p^2 = 0.22$, but not for errors, $F<1$. As indicated by the CIs in Figure 5, both the reduction in RT switch cost with preparation and the asymptotic residual switch cost were statistically robust.

As well as no main effect of experiment, there was also no statistically reliable interaction of experiment and switch cost, nor of experiment x switch/repeat x CSI, $F<1$ in all cases. However, both the absolute reduction in RT switch cost (55.7 versus 83.3 ms) and the proportionate reduction (34.3% versus 44%) were numerically smaller in Experiment 2. To explore this further we

[6] If we pool this estimate with that from the simultaneous condition of Experiment 1, the change in switch cost from CSI=800 to CSI=1400 is 3.5 ms (CI = 22.4 ms).

examined the effect of position within a run of trials in which the same voice was cued. The repeat versus switch contrast pools over all voice repeat trials, but in the task switching literature, task-cuing experiments with unpredictable tasks often show a graded improvement in performance to asymptote over a run of repeat trials following a switch of tasks (e.g. Monsell, Sumner, & Waters, 2003). Figure 6 shows mean correct RT and errors for the switch trial and the first four voice repeats following a switch, including in the analysis only trials preceded by four correct responses. While the error data were almost identical for the two experiments, the RT data for Experiment 2 (unfamiliar voices) show both a smaller change between the first trial of a run (the switch trial) and the first repeat trial (though not significantly smaller, $t(23) = 1.2$) and a steeper improvement with successive repetitions thereafter: the improvement for unfamiliar voices between run positions 2 and 5 was 81 ms, compared to 39 ms for familiar voices, and the difference in slopes was statistically significant, $F(1,23) = 4.77$, $p = 0.039$, $\eta_p^2 = 0.172$. This observation needs replication but suggests that, during trials following the disruption due to a voice switch, participants regained their asymptotic performance level more rapidly when the voices were familiar.

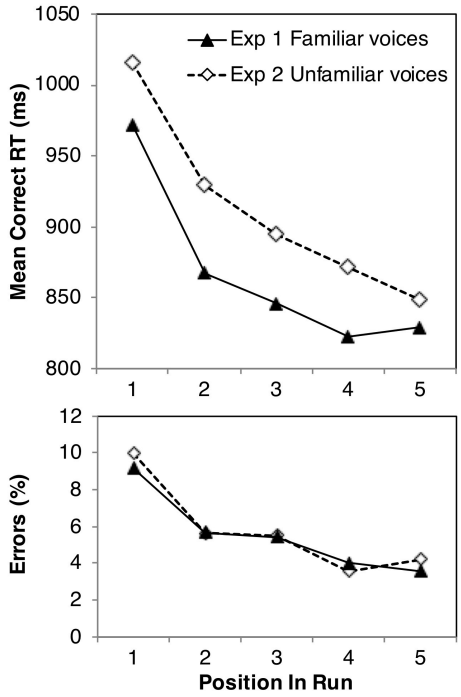

Figure 6. Mean correct RT and error rate as a function of position in a run of trials cuing the same voice, for Experiment 1 (simultaneous condition) and Experiment 2; position 1 is the switch trial.

Effect of within-experiment familiarization. Although participants in Experiment 2 started out unfamiliar with the voices, they necessarily became more familiar with them (or at least with these tokens of their speech) during the course of the experiment. Figure 7 shows the switch cost x CSI interaction for the first and second half of Experiment 2. Although there is a hint of a greater numerical reduction in switch cost in the second half, the session-half x switch x CSI interaction was nowhere near significant for either RT, $F<1$, or errors, $F=1.25$.

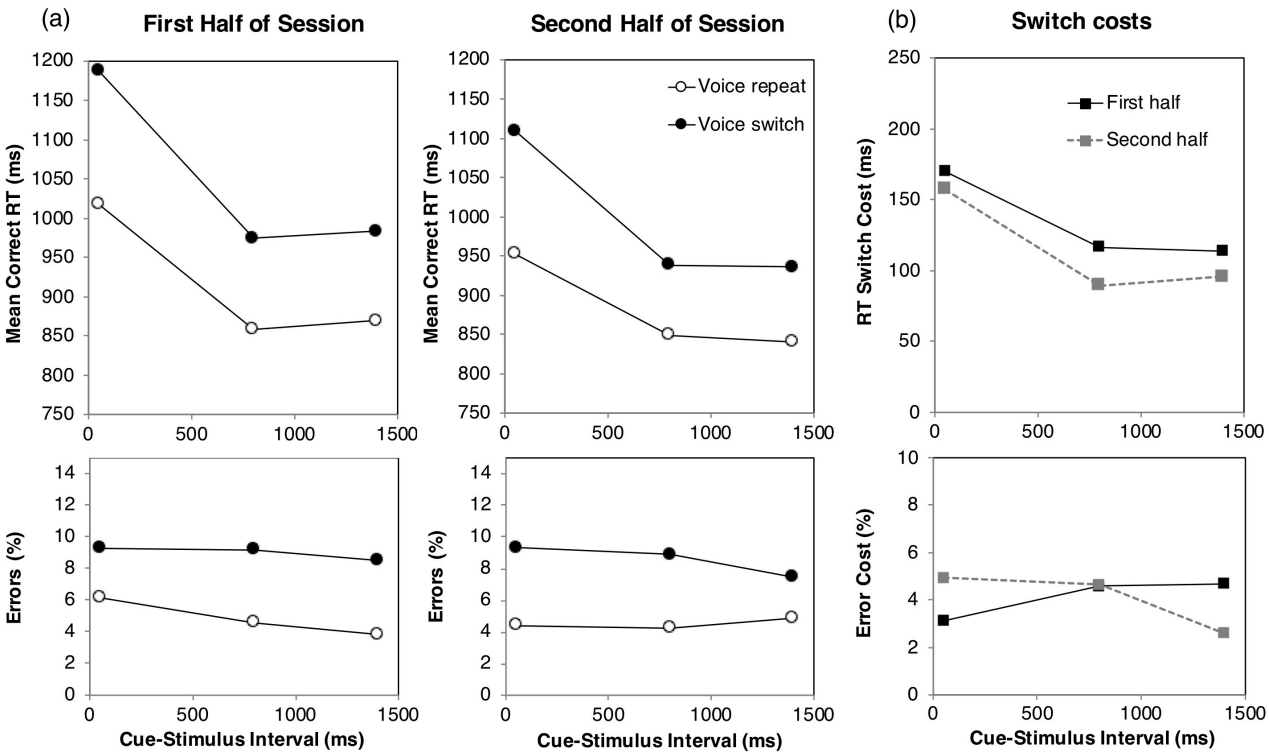

Figure 7. (a) Mean correct RT and error rate; (b) switch costs, for the two halves of Experiment 2.

Effect of F0 difference. One might expect performance and/or the ease of adjusting to a voice switch to depend on the perceptual similarity of the two voices. The only objective measure of similarity available is the pitch separation between the two voices heard, which had a range over participants of 0.9 to 13.7 semitones in Experiment 1, and 4.7 to 14.2 semitones in Experiment 2. Of course, with these natural stimuli, the voices of a pair also differed in intonation, accent, timbre and speech rate, and the pitch varied somewhat over the different digit names recorded by each speaker. An analysis of correlations between F_0 and measures of performance, and components of

the switch cost, is reported in Appendix A. No significant relation between F_0 difference and either performance or switch cost was found.

General Discussion

In Experiment 1, in which we cued attention to one of two highly familiar speakers simultaneously uttering different words, we observed a substantial cost of switching voices from trial to trial. This cost was reduced, by nearly half, when the interval between cue and stimulus onset increased from 50 to 800 ms., with no further significant reduction in cost at the longer interval of 1400 ms. Experiment 2 asked whether this demonstration of effective preparation for a voice depended on the voices being extremely familiar. The participants heard a pair of voices from Experiment 1, but had first encountered them only during the practice trials. They showed a very similar voice switch cost, a similar reduction in that cost with preparation, and a similar residual cost. Nor did we see a marked improvement in preparation as they became more familiar with the voices over the two halves of the experiment. However, a more fine-grained comparison of position-in-run effects across experiments indicated a more rapid return to optimal performance within a run of trials attending to one voice if the voice was very familiar.

Hence the answer to the question in this paper's title appears to be: yes, we can prepare to attend to one of two simultaneous voices, and this does not require extensive familiarity with the voices. But the substantial asymptotic or residual cost of a voice switch suggests a substantial "inertia" in attention to a voice that cannot be overcome by advance preparation on a switch trial. This inertia is only completely overcome – attentional tuning is re-optimized – if several trials requiring attention to the same voice follow, and this re-optimization may be faster for a highly familiar voice.

With respect to why Koch and colleagues (Koch et al., 2011; Lawo et al., 2014; Lawo & Koch, 2015; Seibold et al., 2018) obtained little evidence for effective preparatory tuning, we can now focus on factors other than familiarity. One possibility is that, in most of their experiments, they cued not an individual voice but the gender of the speaker, and the voice of that gender they

would then hear (as well as the distractor voice) could be any of two or three speakers of their gender. Indeed, one experiment in which they did observe a statistically reliable, albeit modest, reduction in the cost of switching the target voice was when there were only two speakers presented (Seibold et al., 2018, Experiment 3). In that experiment target voice switched in short, simple and fully predictable sequences (alternating runs of two trials), but a voice-cuing experiment using only one male and one female voice in the same study (Seibold et al., 2018, Experiment 2) did not find a reliable effect of preparation on the switch cost. Other parameters that may have made a RISC effect hard to detect are the relative proportions of switches vs. repeats (1:1 in their studies, 1:2 in ours) and incongruent vs. congruent trials (1:1 in their studies; 4:1 in ours). As mentioned in the Introduction, making switches relatively common (50% or more) has been shown to reduce the sensitivity to preparation for a task switch (e.g., Kikumoto, Hubbard, & Mayr, 2016; Mayr, Kuhns, & Rieter 2013; Monsell & Mizon, 2006). A relatively high proportion of response-congruent trials, (on which classification of the digit spoken by either voice results in the same response) may also demotivate preparation for one voice. That the two voices were played dichotically in Koch et al.'s studies, with random ear assignment from trial to trial, may also have made it more difficult to prepare to attend to a male/female voice per se, as might the change in the voice selection criterion from ear in some blocks to gender in others (e.g. Seibold et al., 2018).

In the Introduction we also raised the possibility that the simultaneous onset of the voices might impair segregating them as separate objects of attention, but the results of Experiment 1 did not support this speculation. When we delayed the onset of one of the voices in Experiment 1, performance improved, but the benefit of preparation for a change of voice indicated by the RISC effect was very similar. The successive condition of Experiment 1 also afforded the opportunity to ask the question: if participants can endogenously tune attention to one of two voices A and B, what happens to the resulting faster response to Voice A if, just before they hear it, the participant hears Voice B. Does this exogenously supercede the benefit of endogenous tuning to Voice A? The data from the successive condition of Experiment 1 suggest not, though it may be that the interval between the onsets of the two voices (about one third of a second) contained insufficient speech

from the distractor voice, or allowed insufficient time, for exogenous cuing to suppress the effect of endogenous cuing indexed by the RISC effect.

The reader may wonder whether perceptual segregation of the attended voice in the conditions with simultaneous voice onset in our experiments happens "on-line" from the onset of speech, or whether the listener encodes the combined speech streams unfiltered into echoic memory and then somehow scans the stored complex to extract the target voice and figure out what it said (i.e. divided attention followed by focused attention, as in the call-sign paradigm)[7] The reaction times (about 900 ms on voice repeat trials with no preparation, 800 ms with, in Experiment 1) suggest to us an on-line process. Measures of word recognition like lexical decision time are typically 100 to 150 ms longer for auditory than visual words, reflecting the distribution of auditory information required for recognition over time, yet we obtained broadly similar absolute RTs from our participant population for odd/even decisions when participants had to select from one of three simultaneous and spatially separate visual digits with or without preparation (e.g. Longman, Lavric, Munteanu, and Monsell, 2014), and in this case eye-tracking showed that attention oriented to the cued location before the stimulus for longer CSIs.

No less important than the affirmative answer to the titular question (yes, we can prepare to attend to one of two voices) is the finding that with two highly familiar voices heard throughout the experiment from just one location, the efficacy of endogenous preparation is strikingly limited: we observed substantial asymptotic “residual” costs of a switch in the voice to be listened to, of the order of 100 ms in the simultaneous voice conditions. Attention to voices is of course but one of many domains or levels of analysis within which attention can be directed. (See Treisman, 1969, for a seminal differentiation of the multiple levels of analysis at which attention can be deployed). We conclude with a brief survey of evidence suggestive of significant attentional inertia in other domains, followed by some consideration of possible explanations.

Spatial attention in vision. Generally speaking, telling participants where to look in advance of a stimulus will lead to anticipatory fixations. In addition, an extensive body of research using

[7] A reviewer's suggestion.

Posner's (1980) cuing paradigm indicate that covert attention, independent of fixation, may be efficiently cued in advance of stimulus onset, even when the cue is only 80% valid. However, there are circumstances under which even overt spatial attention can exhibit substantial inertia. In recent eyetracking studies, we have used a central word or symbol to cue attention to one of three parafoveal target digit locations in the upcoming stimulus, each location associated with a different classification task. Longman et al. (2014) used either arbitrary cues or cues transparently specifying the *task* to be performed. On switch trials, fixations from the cue to the relevant digit were substantially delayed, and there was a greater tendency to fixate the location associated with the previous task than the other irrelevant location. The latter tendency reduced with a longer CSI, but remained substantial with a CSI as long as 1.4 s. In contrast, when just one task was required throughout the experiment regardless of the target digit's location, so that the cue just indicated the digit's location, the delay in orienting to the target location and the tendency to orient to the previous location were an order of magnitude smaller, and entirely eliminated at the longer CSIs. Moreover, even when different tasks were associated with the three locations, if the cue transparently specified *location* rather than *task*, there was no evidence for attentional inertia at the longer CSIs (Longman, Lavric, & Monsell, 2016). Giving participants control over their preparation interval by triggering stimulus presentation when the participant shifted their fixation towards one of the three locations also eliminated attentional inertia (Longman, Lavric, & Monsell, 2017). It seems that when spatial orientation has high priority, or is the only thing cued, overt shifts of spatial attention can be efficient, with little sign of inertia. But when spatial attention is construed by the participant as just one of several parameters of tasks switched between, it can exhibit substantial inertia.

Spatial attention in hearing. Early experiments using dichotic presentation, such as Broadbent's (1954) "split-span" experiments, suggested that switching between speech streams in the two ears is costly in time. Auditory analogs to Posner's visual cuing paradigm (e.g. Mondor & Zatorre, 1995), as well as "call-sign" experiments in which space is precued and stimuli presented in expected or unexpected locations (Kidd, Arbogast, Mason, & Gallun, 2005), or performance is

compared with or without location cues (Kitterick et al., 2010), support the idea that a spatial filter, "spotlight", or attentional gradient can be endogenously cued in the auditory as well as in the visual domain. The general assumption seems to be that such cuing is relatively flexible. However, in their experiments on cuing of attention to dichotically presented male and female voices reviewed above, Lawo et al. (2014) and Seibold et al. (2018) also compared cuing by ear to cuing by voice. Ear-switching costs were substantial, the overall effect of CSI was greater for ear-cuing, but although the RISC effect was slightly greater for ear- than for voice cuing, the CSI by switch interaction was not reliable, suggesting non-trivial inertia for spatial attention in hearing, under at least some conditions.

Attention to modalities. Switching of attention between modalities has long been known to involve a processing delay, as evidenced by the "prior entry" effect: one of a roughly simultaneous bimodal pair of stimuli will appear to occur earlier, relative to the other, if attention is directed to its modality (Spence and Parise, 2010, for review). Experiments by Kristofferson (1967) in which RT was measured to visual and auditory signals with or without modality uncertainty indicated delays in switching modality of the order of 50 ms. Can such delays be eliminated by cuing and time for preparation? In a cuing study similar to those in Koch's lab described in the Introduction, but with modality as the target, Lukas, Philipp and Koch (2010) presented a bimodal stimulus (a tone on left or right ear combined with a visual stimulus located to left or right) and cued the modality for a left/right decision. An average switch cost of 123 ms at a CSI of 200 ms was reduced at a CSI of 1000 ms to only 22 ms. We do not know whether the latter number represents asymptotic performance, but it suggests relatively little inertia in attention to modality under conditions similar to those in which substantial inertia is found in attention to voice. Fintor, Stephan and Koch (2018) similarly had participants switch between a task mapping visual left/right stimuli to manual left/right responses and auditory left/right stimuli to vocal "left"/"right" responses, and found complete elimination of RT (though not error) costs at a CSI of 1000 ms.

Attention to perceptual dimensions. A number of studies suggest that although cuing the dimension to be attended to helps, there are also limits on how much it helps, manifest in residual

switch costs in spite of ample time to prepare between cue and stimulus. In visual search for a singleton target differing from distractor items on a single dimension (colour or form), Müller, Reimann, and Krummenacher (2003) found costs of trial-to-trial switches of the dimension, and a reduction, but not elimination, of switch costs when the trial was preceded by a verbal dimension cue. Mayr et al. (2013) and Kikumoto et al. (2016) examined compound search in which participants had to locate a cued form or color singleton in a small set of objects and discriminate its value. Eyetracking showed that although the delay induced by a dimension switch was significantly reduced by increasing the CSI, there was still a substantial delay in fixating the relevant dimension. Meiran and Marciano (2002) required same/different judgements with multi-dimensional visual stimuli; they cued on each trial the relevant dimension for the judgement, and found dimension shift costs which did not reduce at all with preparation, although parallel experiments in which other task parameters (such as response mapping) were cued did show substantial preparation effects. In our lab, ERP measures have shown that when participants must switch between judging lexical versus perceptual properties of a letter string (Elchlepp, Lavric, & Monsell, 2015), or between the color and identity of a letter (Elchlepp, Best, Lavric, & Monsell, 2017), early processes are prolonged on a switch trial, even after a long CSI, consistent with attentional inertia.

Attentional inertia. Putting these distinct domains together, it seems that for a number of the domains in which attention can be deployed, there is evidence that in many cases, if not all, endogenous preparation does not eliminate the cost of an attentional switch completely. Hence the observation of substantial limitations to the efficacy of preparation is by no means unique to attention to voices. Is there a common property linking these cases? One possibility might be complexity. A voice among voices, or one task among others, has a complex specification compared to a location in space or a modality. But complexity does not quite seem to capture the difficulty of fully preparing to attend to a dimension (e.g. colour versus form), in contrast to evidence that participants can prepare to attend to a feature value on a dimension (e.g. red) with no residual cost (Lien, Ruthruff and Johnston, 2010); the difference here seems more to do with the abstractness of the specification.

How is attentional inertia to be explained? From the perspective of the literature on task-set control, specification of attentional parameters is but one component (or set of parameters) of task-set, sometimes distinguished from other components, such as S-R mappings and effector specification with the term "stimulus set", in contrast to "response set" (Meiran, 2000), or "attentional set", in contrast to "intentional set" (Rushworth, Passingham, & Nobre, 2002). Several accounts of irreducible residual switch costs from the task-set literature can in principle be applied to attention to voices.

The term "inertia" (originally introduced to the task-switching literature by Allport, Styles and Hsieh, 1994) implies that the voice selection template activated on a trial (and/or suppression of the rival template) simply carries over to the next unless control processes step in to change it when the cue signals a voice change. (It does not follow that top down control does *nothing* during the CSI on a voice repeat trial. If one imagines a continuum of activation of templates A and B, top down control may be construed as biasing the setting one way on this continuum when A is cued, and in the opposite direction when B is cued, and this prevents attention drifting to a point of indifference, as perhaps would happen if nothing were heard for several minutes. However, a more substantial shift along the continuum must be accomplished when the target voice changes.) The general puzzle associated with this class of account is this: if top-down control can bias attention enough to give the new voice priority for processing, and can do this endogenously, in the absence of any stimulus, why it cannot complete the job without the compound stimulus being processed? (And why can it not fully optimize the bias until after attending to the new voice for several trials thereafter?)

In the task-switching literature, one answer has been to suggest that the presence of the stimulus is somehow needed to complete the "act of control" (e.g. Rogers and Monsell, 1995). An alternative answer, also appealing to the impact of the stimulus onset, is that task parameters, presumably including attentional set, become associated with stimuli, so that they are retrieved or reactivated by stimuli; if the parameters retrieved are those required on the present trial, that helps, but if they conflict, then interference results; such associative retrieval happens on all trials, but its

influence is maximal on switch trials because that is when control settings are least stable. In short, even if the appropriate top-down bias can be applied effectively before the stimulus arrives, its work is partially undone by the associative retrieval of competing task parameters. This account, proposed by Waszak, Hommel and Allport (2003, 2005), received support from their finding that when participants had to respond to one element of a compound stimulus (object picture + object name) by naming the picture or the word, RT was influenced by the prior history of association of each element with either or both the two tasks. For example, if the picture had previously been named, even many trials before, the cost of switching from the picture-naming to the word-naming task was greater than if it had not. Three points are worthy of note here. One is that this account is not mutually exclusive with (and may even require) the "inertia" account; indeed Waszak et al. (2005) appeal to the effect of inertia on a switch trial to explain why performance is much more vulnerable to associative interference on task switch than on task repeat trials in their data. Second, although Waszak et al used a long CSI, which in principle allowed preparation, they did not manipulate CSI to show that preparation was effective; experiments that manipulate CSI provide some evidence that stimulus-task associations influence switch costs occur only at short CSIs (Koch & Allport, 2006; Rubin & Koch, 2006), and hence may not be a source of the residual cost. Third, although the situation in the present experiments has similarities to that studied by Waszak et al. (a compound stimulus, one component of which must be responded to), when applied to the present case it has a somewhat paradoxical flavor if we assume that the relevant voice parameters are associatively retrieved by recognition of the two individual voices or voice tokens in the compound. We would have to assume that the elements are identified in time to retrieve associated attentional parameters that in turn have time to bias attention in such a way as to slow the processing of the relevant digit. An alternative idea might be that each compound stimulus, as a complex containing from its onset properties of both voices, becomes associated with and tends to retrieve or reactivate the attentional parameters of both voices, thus attenuating the effect of whatever bias towards one voice endogenous control has by then achieved, resulting in the residual cost. And of course we

would still need an account of how actually applying the new attentional setting over the first few trials of a run overcomes both inertia and associative interference.

In summary, we reviewed the literature to date on attending to one of several simultaneous voices in the absence of location cues. Although we can clearly switch voluntarily between two concurrent speakers, there has hitherto been little unambiguous evidence that we can improve speech segregation by endogenously shifting attention to the expected voice in advance, and the most direct interrogation of this question, using a voice-cuing paradigm, by Koch and colleagues, has suggested that we generally do not. In two experiments in which we cued attention to one of two speakers saying simultaneous digit names, we found that switching attention between voices from trial to trial is costly in RT and errors, but that participants could reduce this cost substantially given a preparation interval of 800 ms – clear evidence for advance endogenous preparation, as seen for other objects of attention. However, extending the preparation time beyond 800 ms to 1400 ms produced little further benefit. The substantial residual cost of a voice switch, as for several other attributes of task-set, suggests an "attentional inertia" – a persistence of the previous tuning of the voice template, that can only partially be reset through top-down control (enough to select the appropriate voice on most trials) , but the changed setting of the attentional template must then be exercised on several trials to re-optimize the tuning to the changed target voice. This retuning process happens faster for a familiar voice. Of course, one word per trial per speaker is clearly an artificial situation. In a multi-talker environment with continuous speech, it is likely that the equivalent optimization happens over several words of continued attention to one voice.

References

Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltá & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing* (pp. 421–452). Cambridge, MA: MIT Press.

Boersma, P & Weenink, D. (2014) Praat: doing phonetics by computer [Computer program].Version 5.4, retrieved 4 October 2014 from http://www.praat.org/

Bressler, S., Masud, S., Bharadwaj, H., Shinn-Cunningham, B. (2014) Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research*, 78, 349-360.

Broadbent D.E. (1954) The role of auditory localization on attention and memory span. *Journal of Experimental Psychology*, 47, 191-196.

Bronkhorst, A.W. (2015) The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention Perception & Psychophysics*, 77, 1465-1487.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109, 1101-1109.

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, 2527–2538.

Carlyon R.P. (2004) How the brain separates sounds. *Trends in Cognitive Sciences*, 8, 465-471

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25, 975-979.

Darwin, C. J. (2008). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 363, 1011–21.

Darwin, C.J., Bringart, D.S., & Simpson, B. D. (2003) Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 114, 2914-2922.

Darwin, C. J. & Hukin, R. W,. (2000a) Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of Americ*a, 107, 970-977.

Darwin, C. J. & Hukin, R. W,. (2000b) Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. *The Journal of the Acoustical Society of America*, 107, 970-977.

Elchlepp H., Best, M., Lavric A., & Monsell S. (2017) Shifting attention between visual dimensions as a source of switch costs. *Psychological Science*, 28, 470-481

Elchlepp H., Lavric A., & Monsell S. (2015) A change of task prolongs early processes: Evidence from ERPs in lexical tasks. *Journal of Experimental Psychology: General*, 144, 299-325.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Fintor, E., Stephan, D.N., & Koch, I. (2018) The interplay of crossmodal attentional preparation and modality compatibility in cued task switching. *Quarterly Journal of Experimental Psychology*, Advance online publication. doi: 10.1177/1747021818771836

Freyman, R.L., Balakrishnan, U., & Helfer K.S. (2004) Effect of number of masking talkers and auditory priming on informational masking in speech recognition, *The Journal of the Acoustical Society of America*, 115, 2246-2256.

Hill, K. T., and Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex*, 20, 583-590.

Holmes. E. (2014) Preparatory and selective attention during multi-talker listening in normal and impaired hearing. Ph D. Thesis, University of York.

Holmes, E., Domingo, Y., & Johnsrude, I.S. (2018) Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science*, Advance online publication. doi: 10.1177/0956797618779083

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.

Holmes, E. , Kitterick, P.T & Summerfield, A.Q. (2018) The advantage of being prepared to listen is underpinned by attentional mechanisms that develop over time. *Attention, Perception & Psychophysics*, 80, 1520-1538

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, 24, 1995–2004

Kidd, G. J., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, 118, 3804– 3815.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. & Koch, I. (2010). Control and interference in task switching – A review. *Psychological Bulletin*, 136, 849-847.

Kikumoto, A., Hubbard, J., & Mayr, U. (2016). Dynamics of task-set carry-over: evidence from eye-movement analyses. *Psychonomic Bulletin and Review*, 23, 899-906.

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, 2498–2508.

Kristofferson, A.B. (1967) Attention and psychophysical time. *Acta Psychologica*, 27, 93-100.

Koch, I. & Allport (2006) Cue-based preparation and stimulus-based priming of tasks in task switching. *Memory & Cognition*, 34, 433-444.

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, 1140–1147.

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking - An integrative review of dual-task and task-switching research. *Psychological Bulletin*, 144, 557-583.

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, EL165-71.

Larson, E. & Lee A.K.C. (2014). Switching auditory attention using spatial and non-spatial features recruits different cortical networks. *NeuroImage* 84, 681–687.

Lavric, A., Mizon, G., & Monsell S. (2008). Neurophysiological signature of effective anticipatory task-set control: a task-switching investigation. *European Journal of Neuroscience*, 28, 1016-1029

Lawo, V., Fels, J., Oberem, J., & Koch, I. (2014). Intentional attention switching in dichotic listening: Exploring the efficiency of nonspatial and spatial selection. *Quarterly Journal of Experimental Psychology*, 67, 2010-2024.

Lawo, V., & Koch, I. (2014). Examining age-related differences in auditory attention control using a task-switching procedure. *Journals of Gerontology Series B: Psychological Sciences and Social Science*s, 69, 237–244

Lawo, V. & Koch, I. (2015) Attention and action: The role of response mappings in auditory attention switching. *Journal of Cognitive Psychology*, 27, 194-206.

Lee, A. K. C, Rajaram, S., Xia, J., Bharadwaj, H., Larson, E., Hämäläinen, M.S., & Shinn-Cunningham (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.

Lien, M. C., Ruthruff, E., & Johnston, J. C. (2010). Attentional capture with rapidly changing attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, 36, 1–16.

Longman, C.S., Lavric, A., Munteanu, C. & Monsell, S. (2014) Attentional inertia and delayed orienting of spatial attention in task-switching. *Journal of Experimental Psychology: Human Perception and Performance*, 40, 1580-1602.

Longman, C.S., Lavric, A. & Monsell, S. (2016) The coupling between spatial attention and other components of task-set: A task-switching investigation. *Quarterly Journal of Experimental Psychology*, 69, 2248-2275.

Longman, C.S., Lavric, A. & Monsell, S. (2017) Self-paced preparation for a task switch eliminates attentional inertia but not the performance switch cost. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 862-873.

Lukas, S., Philipp, A.M., & Koch, I. (2010) The role of preparation and cue-modality in crossmodal task switching, *Acta Psychologica*, 134, 318-322.

Mayr, U., Kuhns, D., & Rieter, M. (2013). Eye movements reveal dynamics of task control. *Journal of Experimental Psychology: General*, 142, 489-509.

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1423-1442

Meiran, N. (2000). Reconfiguration of stimulus task sets and response task sets during task switching. In J. Driver & S. Monsell (Eds.), *Control of Cognitive Processes: Attention and Performance XVIII* (pp. 377-399). Cambridge, MA: MIT Press.

Meiran, N. (2014). The task-cuing paradigm: A user's guide. In J. A. Grange & G. Houghton (Eds.), *Task switching and cognitive control* (pp. 45-73). Oxford: Oxford University Press.

Meiran, N., & Marciano, H. (2002). Limitations in advance task preparation: Switching the relevant stimulus dimension in speeded same-different comparisons. *Memory & Cognition*, 30, 540-550.

Mondor, T. A., and Zatorre, R. J. (1995) Shifting and focusing auditory spatial attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 387–409.

Monsell S. (2003) Task switching. *Trends in Cognitive Sciences*, 7, 134-140.

Monsell, S. (2015) Task-set control and task switching. In J. Fawcett, E. F. Risko, & A. Kingstone (Eds) *The Handbook of Attention*, Ch, 7 (pp.139-172), MIT Press.

Monsell, S. (2017) Task-set regulation. In T. Egner (Ed) *The Wiley Handbook of Cognitive Control*, Ch 2 (pp. 29-49)

Monsell S. & Mizon G.A. (2006) Can the task-cuing paradigm measure an "endogenous" task-set reconfiguration process? *Journal of Experimental Psychology: Human Perception and Performance*, 32, 493-516.

Monsell, S., Sumner, P., & Waters, H. (2003). Task-set reconfiguration with predictable and unpredictable task switches. *Memory & Cognition*, 31, 327–342.

Müller, H. J., Reimann, B., & Krummenacher, J. (2003). Visual search for singleton feature targets across dimensions: Stimulus-and expectancy-driven effects in dimensional weighting. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 1021–1035.

Nolden, S., Ibrahim, C., & Koch, I. (2018). Cognitive control in the cocktail party: Preparing selective attention to dichotically presented voices supports distractor suppression. *Attention, Perception, & Psychophysics*. Advance online publication: doi: 10.3758/s13414-018-1620-x

Nobre, A.C. & Heideman, S.G. (2015) Temporal orienting of attention. In J. Fawcett, E. F. Risko, & A. Kingstone (Eds) *The Handbook of Attention*, Ch, 5 (pp. 57-78), MIT Press.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60, 355-376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5, 42-46.

Posner, M.I. (1980) Orienting of attention, *Quarterly Journal of Experimental Psychology*, 32, 3-25.

Posner, M.I. (2016) Orienting of attention: then and now. *Quarterly Journal of Experimental Psychology*, 69, 1864-1875.

Rogers, R. D., & Monsell, S. (1995). The costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124, 207–231.

Rubin O. & Koch, I. (2006) Exogenous influences on task set activation in task switching. *Quarterly Journal of Experimental Psychology*, 59, 1033–1046.

Rushworth, M. F. S., Passingham, R. E., & Nobre, A. C. (2002). Components of switching intentional set. *Journal of Cognitive Neuroscience*, 14, 1139–1150.

Samson, F., & Johnsrude, I.S. (2016) Effects of a consistent target or masker voice on target speech intelligibility in two- and three-talker mixtures. *The Journal of the Acoustical Society of America*, 139, 1037-1046.

Seibold, J. C., Nolden, S., Oberem, J., Fels, J., & Koch, I. (2018). Intentional preparation of auditory attention-switches: explicit cueing and sequential switch-predictability. *Quarterly Journal of Experimental Psychology*, 71, 1382-1395

Shafiro, V., & Gygi, B. (2007). Perceiving the speech of multiple concurrent talkers in a combined divided and selective attention task. *The Journal of the Acoustical Society of America*, 122, EL229–35.

Shinn-Cunningham, B. & Best, V. (2015) Auditory selective attention. In J. Fawcett, E. F. Risko, & A. Kingstone (Eds) *The Handbook of Attention*, Ch, 5 (pp. 99-117), MIT Press.

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013) The advantage of knowing the talker. *Journal of the American Academy of Audiology*, 24, 689-700.

Spence, C & Parise, C. (2010) Prior entry: A review. *Consciousness and Cognition*, 19, 364-369

Treisman A.M. (1969) Strategies and models of selective attention. *Psychological Review*, 76, 282-299

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin*, 136, 601–626.

Waszak, F., Hommel, B., & Allport, A. (2003). Task-switching and long- term priming: Role of episodic stimulus-task bindings in task-shift costs. *Cognitive Psychology*, 46, 361–413.

Waszak, F., Hommel, B., & Allport, A. (2005). Interaction of task readiness and automatic retrieval in task switching: Negative priming and competitor priming. *Memory & Cognition*, 33, 595–610.

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging*, 15, 88-99.

Appendix A

Effects of the F_0 difference between the voices.

One might expect poorer overall performance if the two voices are similar, and possibly greater difficulty shifting the attentional template the further apart the voices are in stimulus space. Correlations over participants between the difference in average F_0 between the male and female speaker over the eight speech tokens and overall performance were as follows. Experiment 1: for mean RT, r=0.040; for error rate, -.026; Experiment 2: for mean RT, r = 0.22; for error rate, r = 0.08 (p>.31 in all cases). Correlations between F_0 difference and the RT switch cost at the shortest CSI were r= 0.380 (p=.067) for Experiment 1, and 0.07 for Experiment 2, and for the RISC effect r = 0.357 (p=.087), and 0.08 for Experiment 2; the correlations for Experiment 1 were, however, inflated by extreme values. For the residual switch cost, r = 0.137 for Experiment 1, and -0.01 for Experiment 2. Hence no significant relationship could be detected between F_0 difference and either overall performance or any component of the switch cost. However, the voices of a pair differed also in intonation, speech rate and timbre; the pitch as well as other properties varied somewhat over the speech tokens, and the voices were not chosen to maximize the range of F0 differences. In an unpublished voice-cuing study with a large between-subjects manipulation of F_0 difference, we found slightly worse overall performance for the smaller F_0 difference, but there too no effect of the difference on switch cost.

Appendix B

Effects of cue type

We used both a verbal and a pictorial cue for each voice in order to avoid cue repetitions from trial to trial. We report here the effects of this factor in both experiments. In task switching experiments with visual cues (e.g. Lavric, Mizon & Monsell, 2008), we have generally found a transparent verbal cue (e.g. the name of the dimension to attend to, e.g. “color”) to be somewhat more effective than a transparent pictorial cue (e.g. a collage of the possible colors), in the sense of resulting in a smaller residual switch cost. What might we expect in the current case? One possibility is that a visual cue word, through automatically activating its phonology, might interfere

more with the processing of speech than a photo or iconic cue. Although one might expect such interference to be equivalent on voice switch and repeat trials, its impact might also be greater when the burden of switching voices is added. It is also likely that the time taken to extract the meaning of the cue will differ somewhat for the verbal and pictorial cues; any such difference in processing time would have more of an impact on performance at a short CSI (because the cue is still being processed when the stimulus is heard), leading possibly to interactions between the effects of cue type, voice switch and CSI short CSI. Perhaps the clearest evidence of any difference in efficacy of the two types of cue would be a difference in the residual cost.

With this in mind we report in Table B1, for the three conditions of Experiment 1 and for Experiment 2, the mean RT and error rate for the interaction of switch/repeat and CSI, for each cue type, and in Table B2 the corresponding ANOVA results for the effect of and interactions involving cue type. Table B1 also shows the residual cost, computed as the average difference between switch and repeat RT and error rate over the two longer CSIs for each cue type, and a test of the difference between the residual cost for verbal and pictorial cues.

For the Simultaneous condition of Experiment 1, and for Experiment 2, there is evidence of a larger residual cost for word cues than for pictorial cues, significant for reaction time in Experiment 1, and for errors in Experiment 2. Data from these two conditions also indicate a larger switch cost at the shortest CSI (50 ms) for verbal than for pictorial cues, thought the corresponding interaction is statistically significant only for RTs in Experiment 1. However, the Successive condition of Experiment 1 does not show the same pattern: the residual cost is larger for the pictorial cues than for the verbal cues, though not reliably so, and there is no evidence for a larger switch cost at the short CSI for verbal cues. Hence if we were to speculate that the effect for simultaneous presentation reflects some kind of interference from the phonology activated by a verbal cue, we would also have to argue also that the interference is only manifest when processing is challenged by simultaneous onsets.

Table B1

Mean RT (ms) and % Error for each combination of cue type, voice repeat/switch and CSI with residual cost (average switch cost for CSI=800 and CSI=1400) for each cue type

	Verbal cue				Pictorial cue				Difference (+ 95% CI)	*p*
	CSI=50	CSI=800	CSI=1400		CSI=50	CSI=800	CSI=1400			
Exp 1 Simultaneous										
Voice repeat	911 (6.3)	827 (4.7)	816 (4.3)		876 (5.3)	809 (5.1)	810 (4.1)			
Voice switch	1129 (12.9)	935 (10.4)	953 (8.1)		1034 (10.6)	910 (8.8)	873 (7.7)			
Switch cost	218 (6.6)	108 (5.7)	137 (3.9)		157 (5.3)	101 (3.7)	63 (3.7)			
Residual cost, RT				122				82	40.7 ± 32.3	***0.016***
Residual cost, %Error				4.8				3.7	1.1 ± 2.0	*0.27*
Exp 1 Sequential, 1st word										
Voice repeat	797 (4.8)	728 (4.2)	745 (4.7)		765 (5.7)	715 (4.9)	738 (2.8)			
Voice switch	937 (6.3)	788 (7.6)	823 (3.3)		921 (3.5)	821 (4.9)	809 (5.7)			
Switch cost	141 (1.7)	60 (3.4)	78 (3.3)		155 (-2.2)	106 (1.0)	71 (3.0)			
Residual cost, RT				69.1				88.6	–19.5 ± 34.0	*0.25*
Residual cost, %Error				3.4				2.0	1.4 ± 3.0	*0.36*
Exp 1 Sequential, 2nd word										
Voice repeat	747 (6.5))	681 (5.7)	697 (4.1)		698 (6.7)	664 (3.8)	681 (4.0)			
Voice switch	809 (10.8)	706 (8.2)	719 (9.6)		787 (11.4)	711 (5.9)	737 (7.5))			
Switch cost	62 (4.2)	25 (2.4)	22 (5.5)		89 (4.7)	47 (2.2)	56 (3.4)			
Residual cost RT				23.5				51.6	–28.0 ± 34.7	*0.11*
Residual cost %Error				3.9				2.8	1.1 ± 4.7	*0.62*
Exp 2 Simultaneous										
Voice repeat	998 (6.2)	860 (4.5)	861 (5.0)		971 (4.4)	842 (4.2)	851 (3.9)			
Voice switch	1175 (9.9)	988 (11.3)	958 (9.8)		1116 (8.1)	924 (6.8)	958 (6.3)			
Switch cost	177 (3.7)	129 (6.8)	97 (4.8)		145 (3.8)	82 (2.6)	108 (2.4)			
Residual cost RT				113.2				94.6	18.5 ± 35.7	*1.07*
Residual cost %Error				5.8				2.5	3.3 ± 2.4	***0.008***

Table B2

Effects of cue type, cue type x switch and cue type x switch x CSI, from ANOVAs

		Mean correct RT		% Error	
Exp 1 Simultaneous	df	F	p	F	p
Cue type	1,23	56.195	**<.001**	1.574	.222
Cue type x Switch	1,23	20.066	**<.001**	1.355	.256
Cue type x Switch x CSI	2,46	3.830	**.030**	.274	.762
Exp 1 Sequential, 1st word	df	F	p	F	p
Cue type	1,23	1.626	.215	5.093	**.034**
Cue type x Switch	1,23	1.019	.323	2.690	.115
Cue type x Switch x CSI	2,46	.580	.525	.725	.480
Exp 1 Sequential, 2nd word	df	F	p	F	p
Cue type	1,23	1.763	.197	1.918	.179
Cue type x Switch	1,23	4.125	**.054**	.101	.754
Cue type x Switch x CSI	2,46	.070	.928	.266	.736
Exp 2 Simultaneous	df	F	p	F	p
Cue type	1,23	14.764	**.001**	7.807	**.010**
Cue type x Switch	1,23	3.134	.090	3.303	.082
Cue type x Switch x CSI	2,46	1.229	.302	1.547	.227