

Pupillometry shows the effort of auditory attention switching^{a)}

DANIEL R. McCLOY, BONNIE K. LAU, ERIC LARSON, KATHERINE A. I. PRATT, AND
ADRIAN K. C. LEE^{b)}

*Institute for Learning and Brain Sciences, University of Washington, 1715 NE Columbia Rd.,
Box 357988, Seattle, WA, 98195-7988*

March 6, 2017

Running title: Pupillometry and attention switching

^{a)}Portions of the research described here were previously presented at the 37th Annual MidWinter Meeting of the Association for Research in Otolaryngology, and published in McCloy et al (2016), Temporal alignment of pupillary response with stimulus events via deconvolution, J. Acoust. Soc. Am. **139**(3), EL57-EL62.

^{b)}Author to whom correspondence should be addressed. Electronic mail: akclee@uw.edu

ABSTRACT

Successful speech communication often requires selective attention to a target stream amidst competing sounds, as well as the ability to switch attention among multiple interlocutors. However, auditory attention switching negatively affects both target detection accuracy and reaction time, suggesting that attention switches carry a cognitive cost. Pupillometry is one method of assessing mental effort or cognitive load. Two experiments were conducted to determine whether the effort associated with attention switches is detectable in the pupillary response. In both experiments, pupil dilation, target detection sensitivity, and reaction time were measured; the task required listeners to either maintain or switch attention between two concurrent speech streams. Secondary manipulations explored whether switch-related effort would increase when auditory streaming was harder. In Experiment 1, spatially distinct stimuli were degraded by simulating reverberation (compromising across-time streaming cues), and target-masker talker gender match was also varied. In Experiment 2, diotic streams separable by talker voice quality and pitch were degraded by noise vocoding, and the time allotted for mid-trial attention switching was varied. All trial manipulations had some effect on target detection sensitivity and/or reaction time; however, only the attention-switching manipulation affected the pupillary response: greater dilation was observed in trials requiring switching attention between talkers.

© 2017 Acoustical Society of America

Keywords: auditory attention, attention switching, listening effort, pupillometry

I. INTRODUCTION

The ability to selectively attend to a target speech stream in the presence of competing sounds is required to communicate in everyday listening environments. Evidence suggests that listener attention influences auditory stream formation;¹ for listeners with peripheral hearing deficits, changes in the encoding of stimuli often result in impaired stream selection and consequent difficulty communicating in noisy environments.² In many situations (e.g., a debate around the dinner table), it is also necessary to rapidly switch attention among multiple interlocutors — in other words, listeners must be able to continuously update what counts as foreground in their auditory scene, in order to keep up with a lively conversation.

Prior results show that when cueing listeners in a target detection task to either maintain attention to one stream or switch attention to another stream mid-trial, switching attention both reduced accuracy and led to longer response latency *even on targets prior to the attentional switch*.³ This suggests that the act of preparing or remembering to switch imposes some degree of mental effort or cognitive load that can compromise the success of the listening task. Given that listeners are aware of linguistic cues to conversational turn-taking,⁴ the pre-planning of attention switches (and associated hypothesized load) may be part of ordinary listening behavior in everyday conditions, not just an artifact of laboratory experimentation.

Pupillometry, the tracking of pupil diameter, has been used for over five decades to measure cognitive load in a variety of task types.^{5,6} Pupil dilation is an involuntary, time-locked, physiological response that is present from infancy in humans and other animal species. In general, as the cognitive demands of a task increase, pupil dilation of up to about 5-6 mm can be observed up to 1 second after onset of relevant stimuli.⁵⁻⁷ While this task-evoked pupillary response is slow (~ 1 Hz), recent results show that it is possible to track attention and cognitive processes with higher temporal resolution (~ 10 Hz) with deconvolution of the pupillary response.^{8,9}

47 Prior work has shown that the pupillary response co-varies with differences in memory
 48 demands,¹⁰ sentence complexity,¹¹ lexical frequency of isolated written words,¹² or difficulty
 49 of mathematical operations.¹³ In the auditory domain, larger pupil dilations have been
 50 reported in response to decreased speech intelligibility due to background noise,¹⁴ speech
 51 maskers versus fluctuating noise maskers,¹⁵ and severity of spectral degradation of spoken
 52 sentences.¹⁶ The pupillary response has also emerged as a measure of listening effort, which
 53 has been defined as “the mental exertion required to attend to, and understand, an auditory
 54 message,”¹⁷ or, more broadly, as “the deliberate allocation of mental resources to overcome
 55 obstacles in goal pursuit when carrying out a task” involving listening.¹⁸ In this guise,
 56 pupillometry has been used in several studies to investigate the effects of age and hearing
 57 loss on listening effort.^{16,19,20}

58 Recent evidence suggests that the pupillary response is also sensitive to auditory attention.
 59 Dividing attention between two auditory streams is known to negatively affect performance
 60 in psychoacoustic tasks;^{21,22} greater pupil dilation and later peak pupil-size latency have also
 61 been reported for tasks in which listeners must divide their attention between both speech
 62 streams present in the stimulus instead of attending only one of the two,²² or when the
 63 expected location or talker of a speech stream were unknown as opposed to predictable.²³

64 However, it is unknown whether the greater pupil dilation in divided attention tasks is due
 65 to the demands of processing more information, or the effort of switching attention back and
 66 forth between streams (or both). The present study was designed to test whether auditory
 67 attention switches in a strictly selective attention task would elicit mental effort that was
 68 detectable using pupillometry. Both experiments involve selective attention to one of two
 69 auditory streams (spoken alphabet letters), and a pre-trial cue indicating (1) which stream
 70 to attend to and (2) whether to maintain attention on that stream throughout the trial, or
 71 switch attention to the other stream at a designated mid-trial gap. In this way, there is no
 72 need or advantage for listeners to try to attend both streams throughout the trial, so any

73 increase in pupil dilation seen in the switch attention trials should index the effort due to
 74 attention switching, rather than effort due to processing two streams' worth of information.
 75 On the assumption that the divided attention results of Koelewijn and colleagues²² were
 76 at least partially due to listeners switching back and forth between streams, we predicted
 77 greater pupil dilation on trials that required attention switching.

78 Additionally, the two experiments include manipulations of the stimuli designed to compromise
 79 auditory streaming, and thereby make the task of maintaining or switching attention more
 80 difficult. We thus expected that the pupillary response would be larger in trials with more
 81 degraded stimuli, trials where target and masker streams were harder to distinguish, or
 82 trials where the time allocated for switching between streams was shorter. Secondly, these
 83 manipulations provide a test of whether the kind of pupillary response seen in previous
 84 studies that required semantic processing of meaningful sentences might also be seen in
 85 a simpler, closed-set target detection task. Based on findings showing that harder pitch
 86 discrimination trials elicit larger dilations than easier trials,²⁴ and based on findings from
 87 Winn and colleagues that differences in dilation to sentences with different degrees of spectral
 88 degradation occurred *during* sentential stimuli as well as in the post-stimulus delay and
 89 response period,¹⁶ we expected that the stimulus degradations in and of themselves might
 90 also yield larger dilations (in addition to any effect the degradations might have on auditory
 91 stream selection).

92 II. EXPERIMENT 1

93 Experiment 1 involved target detection in one of two spatially separated speech streams.
 94 In addition to the maintain- versus switch-attention manipulation, there was a stimulus
 95 manipulation previously shown²⁵ to cause variation in task performance: degradation of
 96 binaural cues to talker location (implemented as presence/absence of simulated reverberation).

97 Reduced task performance and greater pupil dilation were predicted for the reverberant
 98 condition. This manipulation was incorporated into the pre-trial cue (i.e., on reverberant
 99 trials, the cue was also reverberant). Additionally, the voice of the competing talker was
 100 varied (either the same male voice as the target talker, or a female voice); this manipulation
 101 was not signalled in the pre-trial cue. The same-voice condition was expected to degrade
 102 the separability of the talkers²⁶ and therefore decrease task performance and increase pupil
 103 dilation.

104 **A. Methods**

105 **1. *Participants***

106 Sixteen adults (ten female, aged 21 to 35 years, mean 25.1) participated in Experiment 1. All
 107 participants had normal audiometric thresholds (20 dB HL or better at octave frequencies
 108 from 250 Hz to 8 kHz), were compensated at an hourly rate, and gave informed consent to
 109 participate as overseen by the University of Washington Institutional Review Board.

110 **2. *Stimuli***

111 Stimuli comprised spoken English alphabet letters from the ISOLET v1.3 corpus²⁷ from one
 112 female and one male talker. Mean fundamental frequencies of the unprocessed recordings
 113 were 103 Hz (male talker) and 193 Hz (female talker). Letter durations ranged from 351 to
 114 478 ms, and were silence-padded to a uniform duration of 500 ms, RMS normalized, and
 115 windowed at the edges with a 5 ms cosine-squared envelope. Two streams of four letters each
 116 were generated for each trial, with a gap of 600 ms between the second and third letters
 117 of each stream. The letters “A” and “B” were used only in the pre-trial cues (described
 118 below); the target letter was “O” and letters “IJKMQRUXY” were non-target items. To

allow unambiguous attribution of button presses, the letter “O” was always separated from another “O” (in either stream) by at least 1 second; thus there were between zero and two “O” tokens per trial. The position of “O” tokens in the letter sequence was balanced across trials and conditions, with approximately 40% of all “O” tokens occurring in the third letter slot (just after the switch gap, since that slot is most likely to be affected by attention switches), and approximately 20% in each of the other three timing slots.

Reverberation was implemented using binaural room impulse responses (BRIRs) recorded by Shinn-Cunningham and colleagues.²⁸ Briefly, an “anechoic” condition was created by processing the stimuli with BRIRs truncated to include only the direct impulse response and exclude reverberant energy, while stimuli for the “reverberant” condition were processed with the full BRIRs. In both conditions, the BRIRs recorded at $\pm 45^\circ$ for each stream were used, simulating a separation of 90° azimuth between target and masker streams.

3. Procedure

All procedures were performed in a sound-treated booth; illumination was provided only by the LCD monitor that presented instructions and fixation points. Auditory stimuli were delivered via a TDT RP2 real-time processor (Tucker Davis Technologies, Alachula, FL) to Etymotic ER-2 insert earphones at a level of 65 dB SPL. A white-noise masker with π -interaural-phase was played continuously during experimental blocks at a level of 45 dB SPL, yielding a stimulus-to-noise ratio of 20 dB. The additional noise was included to provide masking of environmental sounds (e.g., friction between subject clothing and earphone tubes) and to provide consistency with follow-up neuroimaging experiments (required due to the acoustic conditions in the neuroimaging suite).

Pupil size was measured continuously during each block of trials at a 1000 Hz sampling frequency using an EyeLink1000 infra-red eye tracker (SR Research, Kanata, ON). Participants’

heads were stabilized by a chin rest and forehead bar, fixing their eyes at a distance of 50 cm from the EyeLink camera. Target detection accuracy and response time were also recorded for comparison with pupillometry data and the results of past studies.

Participants were instructed to fixate on a white dot centered on a black screen and maintain this gaze throughout test blocks. Each trial began with a 1 s auditory cue (spoken letters “AA” or “AB”); the cue was always in a male voice, and its spatial location prompted the listener to attend first to the male talker at that location. The letters spoken in the cue indicated whether to maintain attention to the cue talker’s location throughout the trial (“AA” cue) or to switch attention to the talker at the other spatial location at the mid-trial gap (“AB” cue). The cue was followed by 0.5 s of silence, followed by the main portion of the trial: two concurrent 4-letter streams with simulated spatial separation and varying talker gender (either the same male voice in both streams, or one male and one female voice), with a 600 ms gap between the second and third letters. The task was to respond by button press to the letter “O” spoken by the target talker while ignoring “O” tokens spoken by the competing talker (Figure 1).

Before starting the experimental task, participants heard 2 blocks of 10 trials for familiarization with anechoic and reverberant speech (one with a single talker, one with two simultaneous talkers). Next, listeners did 3 training blocks of 10 trials each (one block of “maintain” trials, one block of “switch” trials, and one block of randomly mixed “maintain” and “switch” trials). Training blocks were repeated until participants achieved $\geq 50\%$ of trials correct on the homogenous blocks and $\geq 40\%$ of trials correct on the mixed block. During testing, the three experimental conditions (maintain/switch, anechoic/reverberant speech, and male-male versus male-female talker combinations) were counterbalanced, intermixed within each block, and presented in 10 blocks of 32 trials each, for a total of 320 trials.

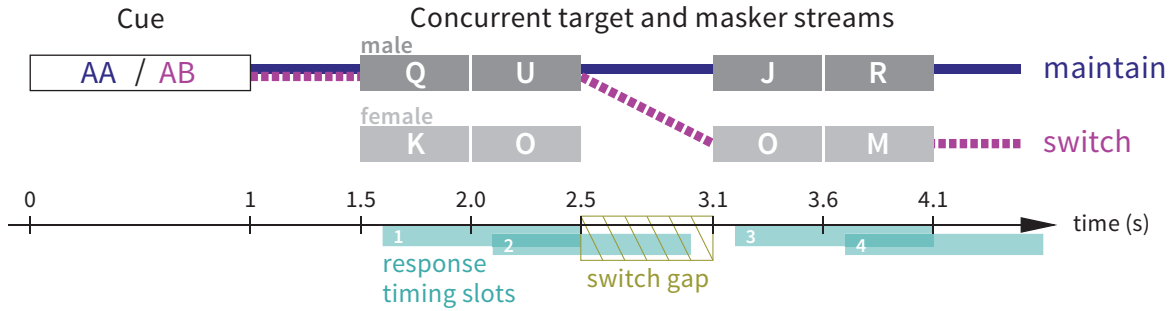


Figure 1: (Color online) Illustration of “maintain” and “switch” trial types in Experiment 1. In the depicted “switch” trial (heavy dashed line), listeners would hear cue “AB” in a male voice, attend to the male voice (“QU”) for the first half of the trial, switch to the female voice (“OM”) for the second half of the trial, and respond once (to the “O” occurring at 3.1–3.6 s). In the depicted “maintain” trial (heavy solid line), listeners would hear cue “AA” in a male voice, maintain attention to the male voice (“QUJR”) throughout the trial, and not respond at all. In the depicted trials, a button press anytime during timing slot 2 would be counted as response to the “O” at 2–2.5 s, which is a “foil” in both trial types illustrated; a button press during slot 3 would be counted as response to the “O” at 3.1–3.6 s (which is considered a target in the switch-attention trial and a foil in the maintain-attention trial), and button presses at any other time would be counted as non-foil false alarms. Note that “O” tokens never occurred in immediately adjacent timing slots (unless separated by the switch gap) so response attribution to targets or foils was unambiguous.

167 4. Behavioral analysis

168 Listener responses were labeled as “hits” if the button press occurred between 100 and
 169 1000 ms after the onset of “O” stimuli in the target stream. Responses at any other time
 170 during the trial were considered “false alarms.” False alarm responses occurring between 100
 171 and 1000 ms following the onset of “O” stimuli *in the masker stream* were additionally labeled
 172 as “responses to foils” to aid in assessing failures to selectively attend to the target stream.
 173 As illustrated in Figure 1, the response windows for adjacent letters partially overlap in time;
 174 responses that occurred during these overlap periods were attributed to an “O” stimulus if
 175 possible (e.g., given the trial depicted in Figure 1, a button press at 3.8 s was assumed to be
 176 in response to the “O” at 3.1–3.6 s, and not to the “M”). If no “O” tokens had occurred in
 177 that period of time, the response was coded as a false alarm for the purpose of calculating
 178 sensitivity, but no reaction time was computed (in other words, only responses to targets and
 179 foils were considered in the reaction time analyses).

180 Listener sensitivity and reaction time were analyzed with (generalized) linear mixed-effects
 181 regression models. A model for listener sensitivity was constructed to predict probability
 182 of button press at each timing slot (four timing slots per trial, see Figure 1) from the
 183 interaction among the fixed-effect predictors specifying trial parameters (maintain/switch,
 184 anechoic/reverberant, and talker gender match/mismatch) and an indicator variable encoding
 185 whether a target, foil, or neither was present in the timing slot. A random intercept was also
 186 estimated for each listener. An inverse probit link function was used to transform button
 187 press probabilities (bounded between 0 and 1) into unbounded continuous values suitable
 188 for linear modeling. This model has the convenient advantage that coefficient estimates are
 189 interpretable as differences in bias and sensitivity on a d' scale resulting from the various
 190 experimental manipulations.^{29–31} Full model specifications are given in the supplementary
 191 material, Equations 1 and 3; the general form of this model is given in Equation 1, where
 192 Φ^{-1} is the inverse probit link function, $Pr(Y = 1)$ is the probability of button press, X is the

design matrix of trial parameters and indicator variables, and β is the vector of parameter coefficients to be estimated.

$$(1) \Phi^{-1}(Pr(Y = 1 | X)) = X'\beta$$

Reaction time was analyzed using linear mixed-effects regression (i.e., with identity link function) but was otherwise analyzed similarly to listener sensitivity. Significance of predictors in the reaction time model was computed via F-tests using the Kenward-Roger approximation for degrees of freedom; significance in the sensitivity model was determined by likelihood ratio tests between models with and without the predictor of interest (as the Kenward-Roger approximation has not been demonstrated to work with non-normally-distributed response variables, i.e., when modeling probabilities). See supplementary material, Sections III.A and III.B and Tables I-III, for full details.

5. *Analysis of pupil diameter*

Recordings of pupil diameter for each trial were epoched from -0.5 to 6 s, with 0 s defined as the onset of the pre-trial cue. Periods where eye blinks were detected by the EyeLink software were linearly interpolated from 25 ms before blink onset to 100 ms after blink offset. Epochs were normalized by subtracting the mean pupil size between -0.5 and 0 s on each trial, and dividing by the standard deviation of pupil size across all trials (to allow pooling across subjects). Normalized pupil size data were then deconvolved with a pupil impulse response kernel.^{8,9} Briefly, the pupil response kernel represents the stereotypical time course of a pupillary response to an isolated stimulus, modeled as an Erlang gamma function with empirically-determined parameters t_{\max} (latency of response maximum) and n (Erlang shape parameter).⁷ The parameters used here were $t_{\max} = 0.512s$ and $n = 10.1$, following previous literature.^{7,9}

Fourier analysis of the subject-level mean pupil size data and the deconvolution kernel

indicated virtually no energy at frequencies above 3 Hz, so for computational efficiency the deconvolution was realized as a best-fit linear sum of kernels spaced at 100 ms intervals (similar to downsampling both signal and kernel to 10 Hz prior to deconvolution), as implemented in the `pyeparse` software.³² After deconvolution, the resulting time series can be thought of as an indicator of mental effort that is time-aligned to the stimulus (i.e., the response latency of the pupil has been effectively removed). Statistical comparison of deconvolved pupil dilation time series (i.e., “effort” in Figures 4 and 8) was performed using a non-parametric cluster-level one-sample t -test on the within-subject differences in deconvolved pupil size between experimental conditions (clustering across time only),³³ as implemented in `mne-python`.³⁴

B. Results

1. Sensitivity

Over all trials, sensitivity (d') ranged across subjects from 1.7 to 4.2 (first quartile 1.9, median 2.4, third quartile 3.0). Box-and-swarm plots displaying quartile and individual differences in d' values between experimental conditions are shown in Figure 2. Note that d' is an aggregate measure of sensitivity that does not distinguish between responses to foil items versus other types of false alarms; however, the statistical model does separately estimate significant differences between experimental conditions for both target response rate and foil response rate, and also estimates a bias term for each condition that captures non-foil false alarm response rates.

The model indicated significant main effects for all three trial type manipulations, as seen in Figure 2a, with effect sizes around 0.2 to 0.3 on a d' scale. Model results indicate that the attentional manipulation led to more responses to both targets (Wald $z=5.23$, $p<0.001$) and foils (Wald $z=2.82$, $p=0.005$) in maintain- versus switch-attention trials, though the net effect was an increase in d' in the maintain attention condition for nearly all listeners. The

241 model also showed a significant difference in response bias in the attentional contrast (Wald
 242 $z=-2.57$, $p=0.01$), with responses more likely in the switch- than the maintain-attention
 243 condition. In fact, there were slightly *fewer* total button presses in the switch-attention trials,
 244 but there were more non-foil false alarm responses in those trials. This suggests that the bias
 245 term is in fact capturing a difference in non-foil false alarm responses (i.e., presses that are
 246 not captured by terms in the model equation encoding responses to targets and foils).

247 Regarding reverberation, listeners were better at detecting targets in the anechoic trials
 248 (Wald $z=3.08$, $p=0.002$), but there was no significant difference in response to foils between
 249 anechoic and reverberant trials. Regarding talker gender (mis)match, the model indicated
 250 both better target detection (Wald $z=2.43$, $p=0.015$) and fewer responses to foils (Wald
 251 $z=-2.31$, $p=0.021$) when the target and masker talkers were different genders. The model
 252 also indicated a two-way interaction for target detection between reverberation and talker
 253 gender (Wald $z=-2.09$, $p=0.036$); this can be seen in Figure 2b: the difference between
 254 anechoic and reverberant trials was smaller when the target and masker talkers were of
 255 different genders. The three-way interaction among attention, reverberation, and talker
 256 gender was not significant.

257 To address the concern that listeners might have attempted to monitor both streams, and
 258 especially that they might do so differently in maintain- versus switch-attention trials, the
 259 rate of listener response to foil items was examined separately for each timing slot. Foil
 260 response rates ranged from 1–4% for slots 1 and 2 (before the switch gap), and from 9–15%
 261 for slots 3 and 4 (after the switch gap), but showed no statistically reliable difference between
 262 maintain- and switch-attention trials for any of the four slots (see supplementary material,
 263 Section III.D.1, for details).

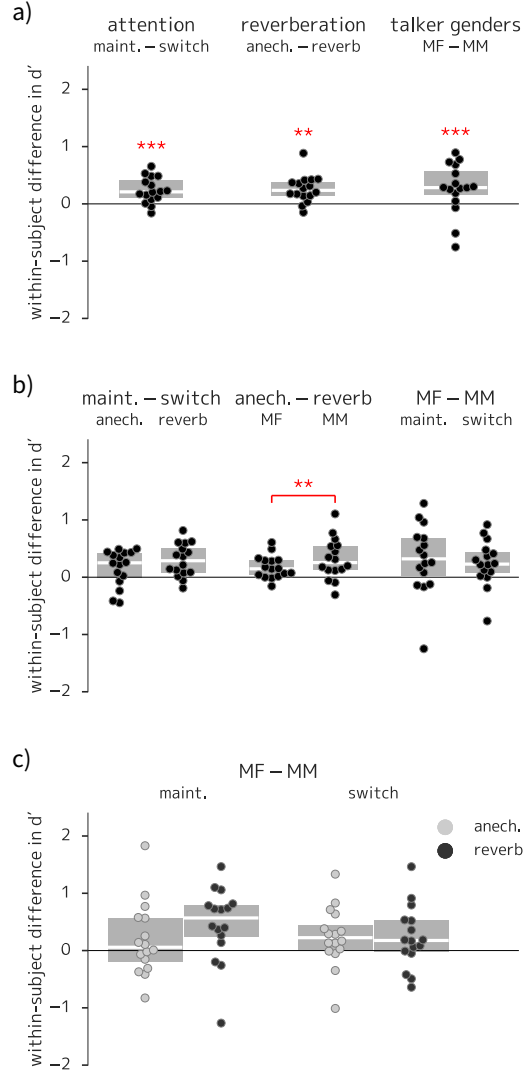


Figure 2: (Color online) Box-and-swarm plots of between-condition differences in listener sensitivity for Experiment 1. Boxes show first & third quartiles and median values; individual data points correspond to each listener; asterisks indicate comparisons with corresponding coefficients in the statistical model that were significantly different from zero. (a) Main effects of attention (higher sensitivity in maintain than switch trials), reverberation (higher sensitivity in anechoic than reverberant trials), and talker gender (mis)match (higher sensitivity in trials with different-gendered target and masker talkers). (b) Two-way interactions; the difference between anechoic and reverberant trials was significantly larger in the gender-match (MM) than in the gender-mismatch (MF) condition. (c) Three-way interaction (no statistically significant differences). ** = $p < 0.01$; *** = $p < 0.001$.

2. *Reaction time*

Over all correct responses, median reaction time for each subject ranged from 434 ms to 692 ms after the onset of the target letter. Box-and-swarm plots showing quartile and individual differences in reaction time values between experimental conditions are shown in Figure 3. The statistical model indicated a significant main effects of attentional condition, reverberation, and talker gender mismatch. Faster response times were seen for targets in maintain-attention trials (9 ms faster on average, $F(1, 5868.1)=4.45$, $p=0.035$), anechoic trials (13 ms faster, $F(1, 5868.1)=9.35$, $p=0.002$), and trials with mismatched talker gender (25 ms faster, $F(1, 5868.2)=35.74$, $p<0.001$). The model showed no significant interactions in reaction time among these trial parameters.

Post-hoc analysis of reaction time by response slot showed no significant differences for the reverberation contrast. For the talker gender (mis)match contrast and the maintain- versus switch-attention contrasts, there were significant differences only in slot 3 (see supplementary material, Section III.D.2 for details). This is consistent with a view that the act of attention switching creates a lag or slow-down in auditory perception.³

3. *Pupillometry*

Mean deconvolved pupil diameter as a function of time for the three stimulus manipulations (reverberant/anechoic trials, talker gender match/mismatch trials, and maintain/switch attention trials) are shown in Figure 4. Only the attentional manipulation shows a significant difference between conditions, with “switch attention” trials showing greater pupillary response than “maintain attention” trials in the time range from 1.0 to 5.5 seconds ($t_{crit} = 2.13$, $p<0.001$; see supplementary material, Section III.C, and supplementary Table IV, for full statistical details). The time courses diverge as soon as listeners have heard the cue, and the response remains significantly higher in the switch-attention condition throughout the remainder of

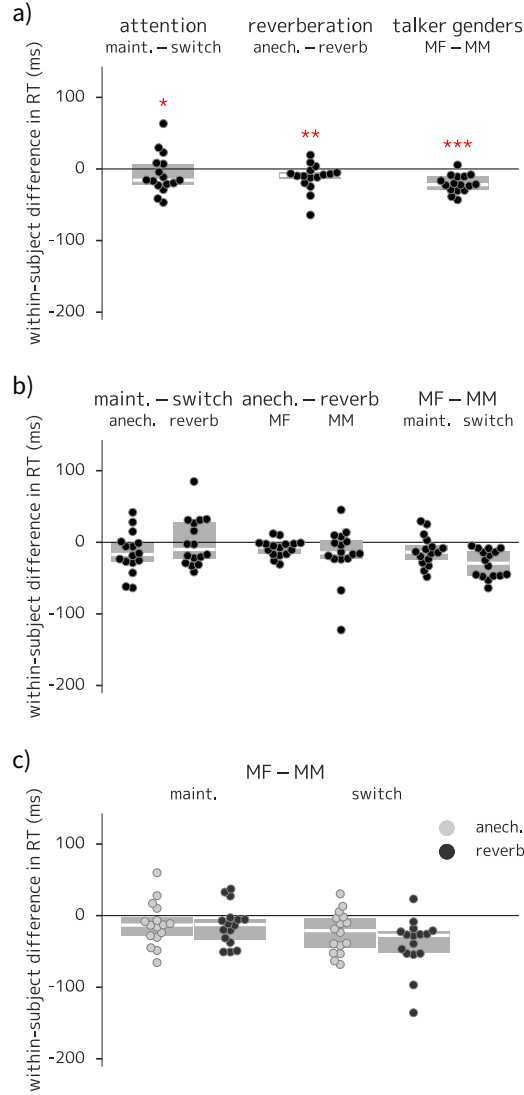


Figure 3: (Color online) Box-and-swarm plots of between-condition differences in reaction time for Experiment 1. Boxes show first & third quartiles and median values; individual data points correspond to each listener; asterisks indicate comparisons with corresponding coefficients in the statistical model that were significantly different from zero. (a) Main effects of attention (faster reaction time in maintain than switch trials), reverberation (faster reaction time in anechoic than reverberant trials), and talker gender (mis)match (faster reaction time in trials with trials with different-gendered target and masker talkers). (b) Two-way interactions (no statistically significant differences). (c) Three-way interaction (no statistically significant difference). * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; MM = matching talker genders; MF = mismatched talker genders.

288 the trial.

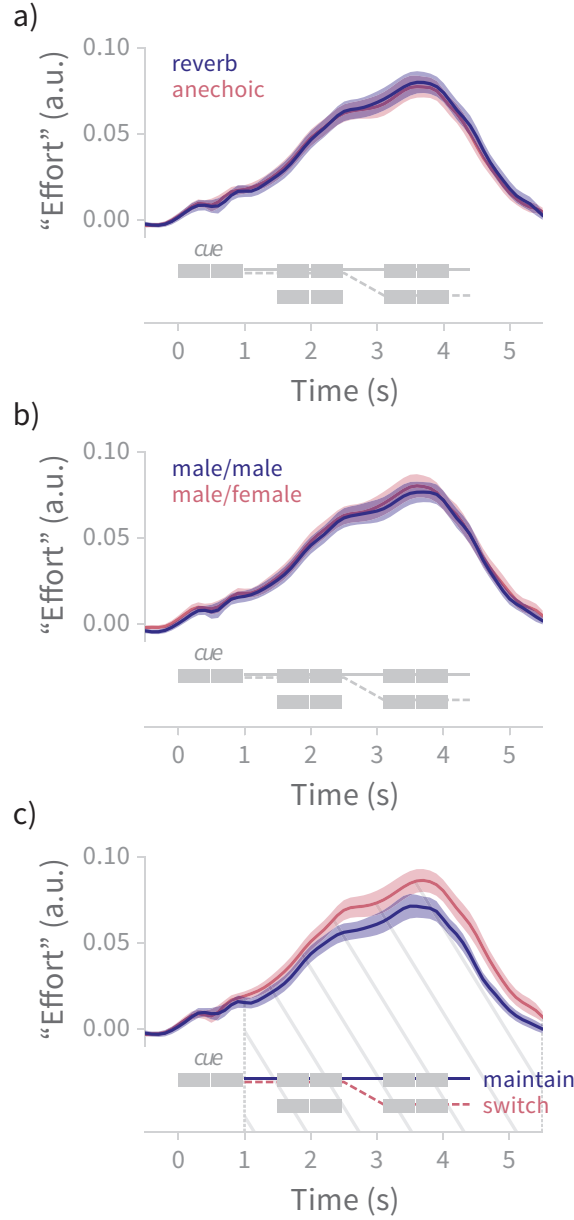


Figure 4: (Color online) Deconvolved pupil size (mean ± 1 standard error across subjects) for (a) reverberant versus anechoic trials, (b) talker gender-match versus -mismatch trials, and (c) maintain- versus switch-attention trials, with trial schematics showing the timecourse of stimulus events (compare to Figure 1). Hatched region shows temporal span of statistically significant differences between time series. The onset of statistically significant divergence (vertical dotted line) of the maintain/switch conditions is in close agreement with the end of the cue. a.u. = arbitrary units (see Section II.A.5 for explanation of “effort”).

C. Discussion

The models of listener sensitivity and reaction time showed main effects in the expected directions for all three manipulations: put simply, listener sensitivity was better and responses were faster when the talkers had different voices, when there was no reverberation, and when mid-trial switching of attention was not required. The difference between anechoic and reverberant trials was *smaller* in trials where the talkers had different voices, suggesting that the advantage of anechoic conditions and the advantage due to talker voice differences are not strictly additive. A possible explanation for this finding is that *either* talker voice difference *or* anechoic conditions are sufficient to support auditory source separation and streaming,^{25,26} but the presence of both conditions cannot overcome difficulty arising from other aspects of the task. Conversely, one might say that *both* segregating two talkers with the same voice *and* segregating two talkers in highly reverberant conditions are hard tasks, which when combined make for a task even more difficult than would be expected if the manipulations were additive (i.e., reverberation hurt performance more when both talkers were male).

Unlike listener sensitivity and reaction time, the pupillary response differed only in response to the attentional manipulation. Interestingly, the difference in pupillary response was seen across the entire trial, whereas the reaction time difference for the maintain-versus-switch contrast was restricted to slot 3 (the immediately post-switch time slot). The fact that patterns of pupillary response do not recapitulate patterns of listener behavior would make sense if, for normal hearing listeners, reverberation and talker gender mismatch are not severe enough degradations to cause sufficient extra mental effort or cognitive load to be observable in the pupil (in other words, the pupillary response may reflect the same processes as the behavioral signal, but may not be as sensitive). However, the magnitude of the effect size in d' is roughly equal for all three trial parameters (see Figure 2a); if behavioral effect size reflects degree of effort or load, then the explanation that pupillometry is just “not sensitive enough” seems unlikely. Another possibility is that the elevated pupil response is simply due

to a higher number of button presses in the switch trials: motor planning and execution are known to cause pupillary dilations.³⁵ However, as mentioned in Section II.B.1, the total number of button presses is in fact higher in the maintain-attention condition. A third possibility is that the pupil dilation only reflects certain kinds of effort or load, and that stimulus degradations that mainly affect listener ability to form and select auditory streams are not reflected in the pupillary response, whereas differences in listener attentional state, such as preparing for a mid-trial attention switch, are reflected by the pupil. Experiment 2 tests this latter explanation, by repeating the maintain/switch manipulation while increasing stimulus degradation, to further impair formation and selection of auditory streams.

III. EXPERIMENT 2

Since no effect of talker gender on pupil dilation was seen in Experiment 1, in Experiment 2 the target and masker talkers were always of opposite gender, and their status as initial target or masker was counterbalanced across trials. Since no effect of reverberation on pupillary response was seen in Experiment 1, Experiment 2 also removed the simulated spatial separation of talkers and involved a more severe cued stimulus degradation known to cause variation in task demand: spectral degradation implemented as variation in number of noise-vocoder channels, 10 or 20. Based on results from Winn and colleagues showing increased dilation for low versus high numbers of vocoder channels with full-sentence stimuli,¹⁶ greater pupil dilation was expected here in the (more difficult, lower-intelligibility) 10-channel condition. As in Experiment 1, a pre-trial cue indicated whether to maintain or switch attention between talkers at the mid-trial gap; here the cue also indicated whether spectral degradation was mild or severe (i.e., the cue underwent the same noise vocoding procedure as the main portion of the trial).

Additionally, in Experiment 2 the duration of the mid-trial temporal gap provided for attention

switching was varied (either 200 ms or 600 ms). Behavioral and neuroimaging research suggest that the time course of attention switching in the auditory domain is around 300-400 ms;^{3,36} accordingly, we expected the short gap trials to be challenging and thus predicted greater pupil dilation in short-gap trials (though only in the post-gap portion of the trial). The duration of the gap was not predictable from the pre-trial cue.

A. Methods

1. *Participants*

Sixteen adults (eight female, aged 19 to 35 years, mean 25.5) participated in Experiment 2. All participants had normal audiometric thresholds (20 dB HL or better at octave frequencies from 250 Hz to 8 kHz), were compensated at an hourly rate, and gave informed consent to participate as overseen by the University of Washington Institutional Review Board.

2. *Stimuli*

Stimuli were based on spoken English alphabet letters from the ISOLET v1.3 corpus²⁷ from the same female and male talkers used in Experiment 1, with the same stimulus preprocessing steps (padding, amplitude normalization, and edge windowing). Two streams of four letters each were generated for each trial, with a gap of either 200 or 600 ms between the second and third letters of each stream. The letters “A” and “U” were used only in the pre-trial cues (described below); the target letter was “O” and letters “DEGPV” were non-target items. The cue and non-target letters differed from those used in Experiment 1 in order to maintain discriminability of cue, target, and non-target letters even under the most degraded (10-channel vocoder) condition. Specifically, the letters were chosen so that the vowel nuclei differed between the cue, target, and non-target letters: representations of the vowel nuclei in

the International Phonetic Alphabet are /e/ and /u/ (cues “A” and “U”), /o/ (target “O”) and /i/ (non-target letters “DEGPV”).

Spectral degradation was implemented following a conventional noise vocoding strategy.³⁷ The stimuli were fourth-order Butterworth bandpass filtered into 10 or 20 spectral bands of equal equivalent rectangular bandwidths.³⁸ This filterbank ranged from 200 to 8000 Hz (low cutoff of lowest filter to high cutoff of highest filter). Each band was then half-wave rectified and filtered with a 160 Hz low-pass fourth-order Butterworth filter to extract the amplitude envelope. The resulting envelopes were used to modulate corresponding noise bands (created from white noise filtered with the same filterbank used to extract the speech bands). These modulated noise bands were then summed, and presented diotically at 65 dB SPL. As in Experiment 1, a simultaneous white-noise masker was also presented (see Section II.A.3).

3. Procedure

Participants were instructed to fixate on a white dot centered on a black screen and maintain such gaze throughout test blocks. Each trial began with a 1 s auditory cue (spoken letters “AA” or “AU”); the cue talker’s gender indicated whether to attend first to the male or female voice, and additionally indicated whether to maintain attention to that talker throughout the trial (“AA” cue) or to switch attention to the other talker at the mid-trial gap (“AU” cue). The cue was followed by 0.5 s of silence, followed by the main portion of the trial: two concurrent, diotic 4-letter streams (one male voice, one female voice), with a variable-duration gap between the second and third letters. The task was to respond by button press to the letter “O” spoken by the target talker (Figure 5). To allow unambiguous attribution of button presses, the letter “O” was always separated from another “O” (in either stream) by at least 1 second, and its position in the letter sequence was balanced across trials and conditions. Distribution of targets and foils across timing slots was equivalent to Experiment 1.

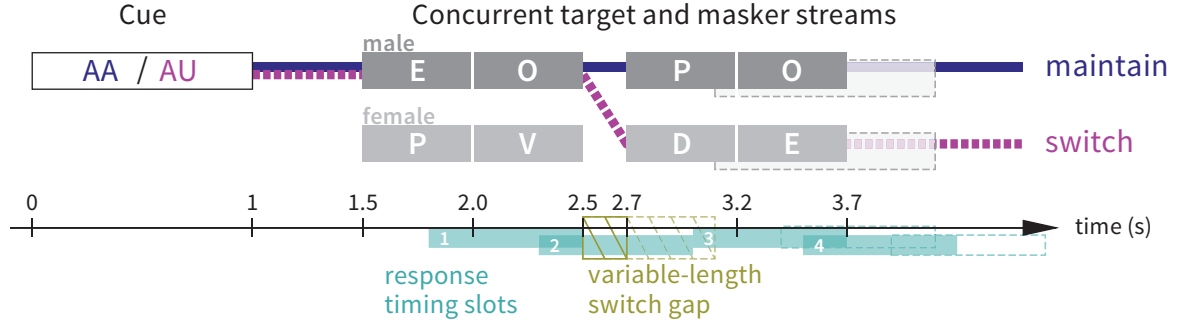


Figure 5: (Color online) Illustration of “maintain” and “switch” trial types in Experiment 2. The short-gap version is depicted; timing of long-gap trial elements (where different) are shown with faint dashed lines. In the depicted “switch” trial (heavy dashed line), listeners would hear cue “AU” in a male voice, attend to the male voice (“EO”) for the first half of the trial and the female voice (“DE”) for the second half of the trial, and respond once (to the “O” occurring at 2–2.5 seconds). In the depicted “maintain” trial (heavy solid line), listeners would hear cue “AA” in a male voice, attend to the male voice (“EOPO”) throughout the trial, and respond twice (once for each “O”).

Before starting the experimental task, participants heard 2 blocks of 10 trials for familiarization with noise-vocoded speech (one with a single talker, one with the two simultaneous talkers). Next, they did 3 training blocks of 10 trials each (one block of “maintain” trials, one block of “switch” trials, and one block of randomly mixed “maintain” and “switch” trials). Training blocks were repeated until participants achieved $\geq 50\%$ of trials correct on the homogenous blocks and $\geq 40\%$ of trials correct on the mixed block. During testing, the three experimental conditions (maintain/switch, 10/20 channel vocoder, and 200/600 ms gap duration) were counterbalanced, intermixed within each block, and presented in 10 blocks of 32 trials each, for a total of 320 trials.

4. Behavioral analysis

As in Experiment 1, listener responses were labeled as “hits” if the button press occurred within a defined temporal response window after the onset of “O” stimuli in the target stream, and all other responses were considered “false alarms.” However, unlike Experiment 1, the designated response window for targets and foil items ran from 300 to 1000 ms after the

onset of “O” stimuli (in Experiment 1 the window ranged from 100 to 1000 ms). This change resulted from a design oversight, in which the placement of target or foil items in both of slots 2 and 3 (on either side of the switch gap) yielded a period of overlap of the response windows for slots 2 and 3 in the short gap trials, in which presses could not be unambiguously attributed. However, in Experiment 1 (where response times as fast as 100 ms were allowed) the fastest response time across all subjects was 296 ms, and was the sole instance of a sub-300 ms response. Therefore, raising the lower bound on the response time window to 300 ms for Experiment 2 is unlikely to have disqualified any legitimate responses (especially given the more severe signal degradation, which is likely to increase response times relative to Experiment 1), and eliminates the overlap between response slots 2 and 3 on short-gap trials. Statistical modeling of sensitivity used the same approach as was employed in Experiment 1: predicting probability of button press in each timing slot based on fixed-effect predictors (maintain/switch, 10- or 20-channel vocoder, and short/long mid-trial gap duration), a target/foil/neither indicator variable, and a subject-level random intercept. Statistical modeling of response time also mirrored Experiment 1, in omitting the indicator variable and considering only responses to targets and foils. See supplementary material, Sections IV.A and IV.B and Tables VI-VIII, for full details.

5. *Analysis of pupil diameter*

Analysis of pupil diameter was carried out as in Experiment 1: trials epoched from -0.5 to 6 s, linear interpolation of eye blinks, per-trial baseline subtraction and per-subject division by standard deviation of pupil size. Deconvolution and statistical analysis of normalized pupil size data was also carried out identically to Experiment 1.

B. Results

1. Sensitivity

Over all trials, sensitivity (d') ranged across subjects from 1.4 to 4.2 (first quartile 1.8, median 2.2, third quartile 2.7). Box-and-swarm plots displaying quartile and individual differences in d' values between experimental conditions are shown in Figure 6. Again, note that d' is an aggregate measure of sensitivity that does not distinguish between responses to foil items versus other types of false alarms, but the statistical model does estimate separate coefficients for target response rate, foil response rate, and a bias term capturing non-foil false alarm responses. The model indicated significant main effects for all three trial type manipulations, as seen in Figure 6a. Specifically, model results indicate no significant difference in target detection between maintain- and switch-attention trials (Wald $z=1.07$, $p=0.284$), but did show fewer responses to foils in maintain-attention trials (Wald $z=-2.54$, $p=0.011$; estimated effect size 0.15 d'); a corresponding increase in d' in the maintain attention condition is seen for nearly all listeners in Figure 6a, left column. Regarding spectral degradation, listeners were better at detecting targets in 20-channel trials (Wald $z=4.09$, $p<0.001$; estimated effect size 0.19 d'), but there was no significant difference in response to foils for the spectral degradation manipulation (Wald $z=0.69$, $p=0.489$). For the switch gap length manipulation, the model indicated much lower response to target items (Wald $z=-7.51$, $p<0.001$; estimated effect size 0.35 d') and much greater response to foil items (Wald $z=9.24$, $p<0.001$; estimated effect size 0.56 d') in the long gap trials.

The model also showed two-way interactions between gap duration and spectral degradation (lower sensitivity in 10-channel long-gap trials; Figure 6b, middle column), and between gap duration and the attentional manipulation (lower sensitivity in maintain-attention long-gap trials; Figure 6b, right column). The interaction between gap duration and the attentional manipulation showed increased responses to foil items in maintain-attention long-gap trials

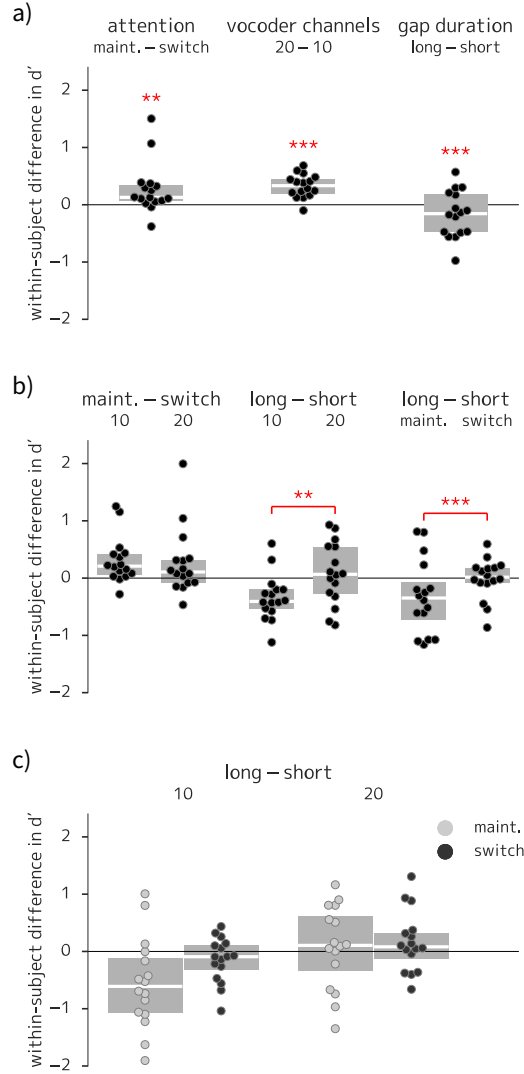


Figure 6: (Color online) Box-and-swarm plots of between-condition differences in listener sensitivity for Experiment 2. Boxes show first & third quartiles and median values; individual data points correspond to each listener; asterisks indicate comparisons with corresponding coefficients in the statistical model that were significantly different from zero. (a) Main effects of attention (higher sensitivity in maintain than switch trials), spectral degradation (higher sensitivity in 20-channel than 10-channel vocoded trials), and switch gap duration (higher sensitivity in trials with a short gap). (b) Two-way interactions: the difference between long- and short-gap trials was greater (more negative) in the 10-channel-vocoded trials and in the maintain-attention trials. (c) Three-way interaction (not significant). * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$.

(Wald $z=2.98$, $p=0.003$). The terms modeling interaction between gap duration and spectral degradation were not significantly different from zero at the $p<0.05$ level when targets and foils are modeled separately (Wald $z=1.66$, $p=0.097$ for targets; Wald $z=-1.92$, $p=0.055$ for foils), but the exclusion of these terms from the model did significantly decrease model fit according to a likelihood ratio test ($\chi^2(2)=11.38$, $p=0.003$).

Post-hoc analysis of target detection accuracy showed no significant differences by slot when correcting for multiple comparisons, but the trend suggested that the two-way interaction between gap duration and spectral degradation was driven by the *first* time slot, while the two-way interaction between gap duration and attentional condition was predominantly driven by the *last* time slot (paired t -tests by slot on logit-transformed hit rates all $p>0.04$; Bonferroni-corrected significance level 0.00625).

2. *Reaction time*

Over all correct responses, median reaction time for each subject ranged from 493 ms to 689 ms after the onset of the target letter. Box-and-swarm plots showing quartile and individual differences in reaction time values between experimental conditions are shown in Figure 7. The statistical model indicated a significant main effects of spectral degradation and switch gap length. Faster response times were seen for targets in trials processed with 20-channel vocoding (35 ms faster on average, $F(1, 4605.0)=21.79$, $p<0.001$), and trials with a long switch gap (66 ms faster, $F(1, 4606.9)=77.52$, $p<0.001$). The model also showed a significant interaction between spectral degradation and switch gap length (44 ms faster with 20-channel vocoding and long gaps, $F(1, 4604.4)=8.57$, $p=0.003$).

As in Experiment 1, post-hoc tests of reaction time difference between maintain- and switch-attention trials by slot showed a significant difference localized to slot 3 (the immediately post-gap slot), with faster reaction times in maintain-attention trials (28 ms faster on average).

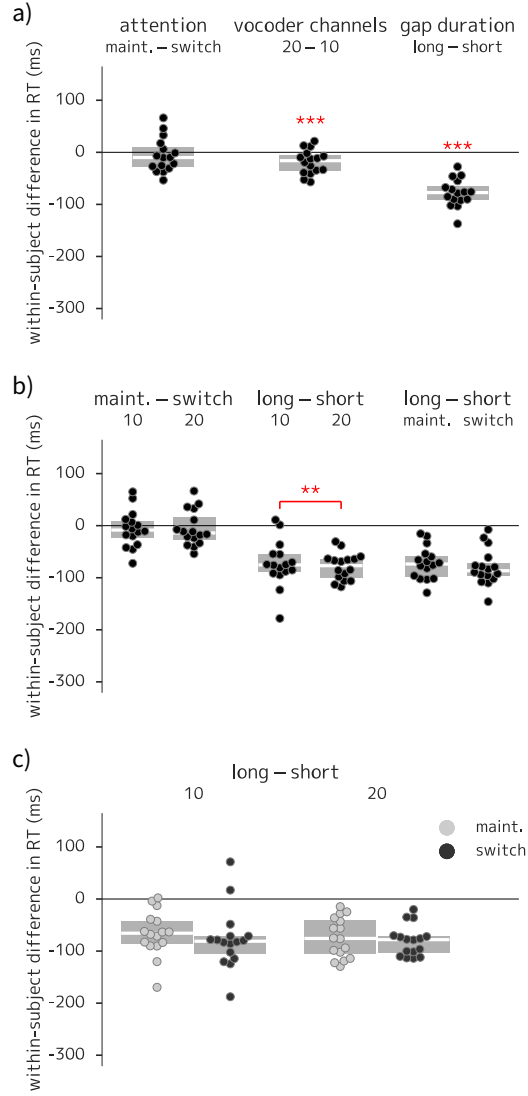


Figure 7: (Color online) Box-and-swarm plots of between-condition differences in reaction time for Experiment 2. Boxes show first & third quartiles and median values; individual data points correspond to each listener; asterisks indicate comparisons with corresponding coefficients in the statistical model that were significantly different from zero. (a) Main effects of attention, spectral degradation, and gap duration (faster response time in trials with 20-channel vocoding, and in long-gap trials). (b) Two-way interactions (larger difference in reaction times between long- and short-gap trials in the 10- versus the 20-channel condition). (c) Three-way interaction (no statistically significant difference). *** = $p < 0.001$.

For the spectral degradation contrast, a significant difference was seen only in slot 1, with faster reaction times in the 20-channel trials (68 ms faster on average); this pattern of results could arise if listener adaptation to the level of degradation was incomplete when the trial started, but was in place by the end of slot 1. For the gap length manipulation, significantly faster reaction times were seen in the long-gap trials for slot 3 (155 ms faster on average) and slot 4 (135 ms faster on average), and significantly *slower* reaction times in the long-gap trials for slot 1 (261 ms slower on average). The faster reaction times in the long-gap trials in slots 3 and 4 are expected given that listeners had additional time to process the first half of the trial and/or prepare for the second half in the long-gap condition. However, the difference in reaction time in slot 1 is unexpected and inexplicable given that the gap length manipulation was uncued. See supplementary material, Section IV.D.1 for details.

3. *Pupillometry*

Mean deconvolved pupil diameter as a function of time for the three stimulus manipulations (10/20 vocoder channels, gap duration, and maintain/switch attention trials) is shown in Figure 8. Similar to Experiment 1, the attentional manipulation shows a significant difference between conditions, with switch-attention trials showing greater pupillary response than maintain-attention trials in the time range from 0.9 to 5.6 s ($t_{crit} = 2.13$, $p < 0.001$); in Experiment 1, the significant difference spanned 1.0 - 5.5 s. Also as in Experiment 1, the time courses diverge as soon as listeners have heard the cue, and the response remains higher in the switch-attention condition throughout the rest of the trial. There is also a significant difference in the time course of the pupillary response between long- and short-gap trials in the time range 3.9 - 5.0 s ($t_{crit} = 2.13$, $p < 0.01$), with the signals diverging around the onset of the mid-trial gap (though only differing statistically in the final ~1 s of the trial). See supplementary material, Section IV.C and Table IX, for full details.

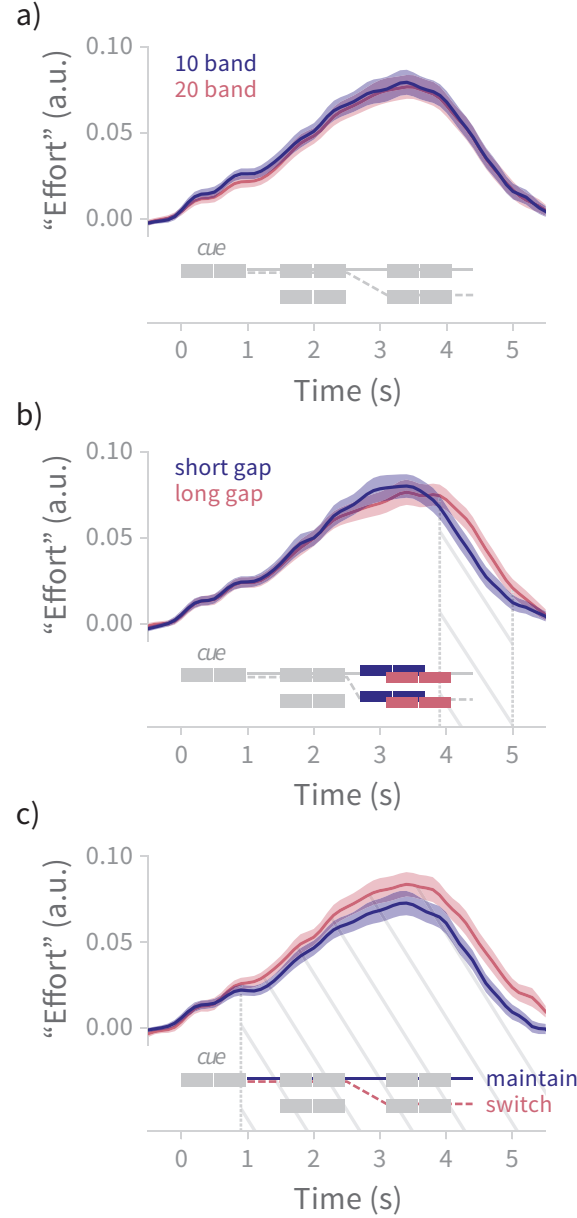


Figure 8: (Color online) Deconvolved pupil size (mean ± 1 standard error across subjects) for (a) 10- versus 20-band vocoded stimuli, (b) 200 versus 600 ms mid-trial switch gap durations, and (c) maintain- versus switch-attention trials, with trial schematics showing the timecourse of stimulus events (compare to Figure 5). Hatched region shows temporal span of statistically significant differences between time series. The late-trial divergence in (b) is attributable to the delay of stimulus presentation in the long-gap condition; the onset of divergence in (c) aligns with the end of the cue, as in Experiment 1 (see Figure 4c). a.u. = arbitrary units (see Section II.A.5 for explanation of “effort”).

C. Discussion

The model of listener sensitivity for Experiment 2 showed main effects of the spectral degradation and attentional manipulations in the expected directions (based on past literature^{16,22} and the results of Experiment 1): listener sensitivity was better when there were more vocoder channels (better spectral resolution) and when mid-trial switching of attention was not required. However, the results of the gap duration manipulation were unexpected; based on past findings that auditory attention switches take between 300 and 400 ms,^{3,36} we hypothesized that a gap duration of 200 ms would cause listeners to fail to detect targets in the immediate post-gap position (i.e., timing slot 3). We did see slower reaction time in the short-gap trials, but sensitivity was actually *better* in the short-gap trials than in the long-gap ones for most listeners (Figure 6a, right column). However, according to the statistical model this effect appears to be restricted to the 10-channel and maintain-attention trials (see Figure 6b, middle and right columns, and 6c, left column). Interestingly, the model coefficient estimates indicated that the interactions were more strongly driven by a difference in responses to foil items, not targets.

A possible explanation for the elevated response to foils in the long-gap condition is that the long-gap condition interfered with auditory streaming, the 10-channel condition also interfered with streaming, and when both conditions occurred simultaneously there was a strong effect on listener ability to group the pre- and post-gap letters into a single stream (i.e., to preserve stream identity across the gap). Using minimally processed stimuli (monotonized, but without intentional degradation), Larson and Lee showed a similar “drop off” in performance in their maintain-attention trials when the gap duration reached 800 ms;³ perhaps the spectral degradation in our stimuli decreased listeners’ tolerance for gaps in the stream, causing performance to drop off at shorter (600 ms) gap lengths. However, this explanation still does not account for the finding that the 10-channel plus long-gap difficulty seems to occur only in the maintain-attention trials. One might speculate that the act of switching attention

520 at the mid-trial gap effectively “fills in” the gap, making the temporal disconnect between
 521 pre- and post-gap letters less noticeable, and thereby preserving attended stream identity
 522 across a longer gap duration than would be possible if attention were maintained on a single
 523 source. In other words, if listeners must conceive of the “stream of interest” as a source
 524 that undergoes a change in voice quality partway through the trial, the additional mental
 525 effort required to make the switch might result in *more accurate* post-gap stream selection,
 526 whereas the putatively less effortful task of maintaining attention to a consistent source could
 527 lead to *less accurate* post-gap stream selection when stream formation is already difficult
 528 (due to strong spectral degradation) and stream interruptions are long. Further study of the
 529 temporal dynamics of auditory attention switching is needed to clarify how listeners’ intended
 530 behavior affects stream stability across temporal caesuras of varying lengths, and how this
 531 process interacts with signal degradation or quality.

532 If this speculation is correct — that signal degradation reduces listener tolerance of gaps
 533 in auditory stream formation and preservation — then this finding may have important
 534 implications for listeners experiencing both hearing loss and cognitive decline. Specifically,
 535 poor signal quality due to degradation of the auditory periphery could lead to greater difficulty
 536 in stream preservation across long gaps, but cognitive decline may make rapid switching
 537 difficult. In other words, the cognitive abilities of older listeners might require longer pauses
 538 to switch attention among multiple interlocutors, but the longer pauses may in fact make it
 539 harder to preserve focus in the face of degraded auditory input.

540 It is also interesting that the post-hoc analyses suggested possibly different temporal loci for
 541 the effects of different stimulus manipulations (i.e., affecting pre- versus post-gap time slots).
 542 This might indicate that differences in the strength of sensory memory traces of the stimuli
 543 played a role. However, it is important to note that we attempted to include time slot as
 544 an additional (interacting) term in the statistical model, but those more complex models
 545 were non-convergent; therefore we hesitate to draw any strong conclusions from the post-hoc

546 t -tests.

547 Regarding the pupillary response, we again saw a difference between maintain- and switch-
 548 attention trials, with the divergence beginning as soon as listeners heard the attentional cue.
 549 We also saw a significant difference in the pupillary response to long- versus short-gap trials,
 550 though the difference appears to be a post-gap delay in the long-gap trials (mirroring the
 551 stimulus time course), rather than a vertical shift indicating increased effort. Contrary to our
 552 hypothesis, there was no apparent effect of spectral degradation on the pupillary response.

553 IV. GENERAL DISCUSSION

554 The main goal of these experiments was to see whether the pupillary response would reflect
 555 the mental effort of switching attention between talkers who were spatially separated (Exper-
 556 iment 1), or talkers separable only by talker voice quality and pitch (Experiment 2). The
 557 overall finding was that attention switching is clearly reflected in the pupillary signal as an
 558 increase in dilation that begins either as soon as listeners are aware that a switch will be
 559 required, or perhaps as soon as they begin planning the switch; since we did not manipulate
 560 the latency between the cue and the onset of the switch gap these two possibilities cannot be
 561 disambiguated.

562 A secondary goal of these experiments was to reproduce past findings regarding the pupillary
 563 response to degraded *sentential* stimuli, but using a simpler stimulus paradigm (spoken letter
 564 sequences) and (in Experiment 1) relatively mild stimulus degradations like reverberation. In
 565 fact, we failed to see any effect of stimulus degradation in the pupillary response, neither
 566 when degrading the temporal cues for spatial separation through simulated reverberation,
 567 nor with more severe degradation of the signal’s spectral resolution through noise vocoding
 568 (Experiment 2). We believe the key difference lies in our choice of stimuli: detecting a target
 569 letter in a sequence of spoken letters is not the same kind of task as computing the meaning

570 of a well-formed sentence, and our results suggest that simply detecting targets among a
 571 small set of possible stimulus tokens does not engage the same neural circuits or invoke the
 572 same kind of mental effort or cognitive load that is responsible for pupillary dilations seen
 573 in the sentence comprehension tasks of Zekveld and colleagues (showing greater dilation
 574 to sentences with lower signal-to-noise ratios [SNRs])^{14,19} or Winn and colleagues (showing
 575 greater dilation to sentences with more severe spectral degradation).¹⁶ Taking those findings
 576 together with the results of the present study, one might say that signal degradation itself
 577 was not the proximal cause of pupil dilation in those sentence comprehension experiments;
 578 rather, it was the additional cogitation or effort needed to construct a coherent linguistic
 579 meaning from degraded speech that led to the pupillary responses they observed.

580 Notably, Winn and colleagues showed a sustained pupillary response in cases where listeners
 581 failed to answer correctly, suggesting that continued deliberation about how to respond may
 582 be reflected by pupil size. Similarly, Kuchinsky and colleagues²⁰ showed greater pupillary
 583 response in word-identification tasks involving lower SNRs when lexical competitors were
 584 present among response choices; their results show a sustained elevation in the time course
 585 of the pupillary response in the harder conditions (as well as a parallel increase in reaction
 586 time). Both sets of findings suggest that the pupillary response reflects effort exerted by
 587 the listener, as do the sustained large dilations seen in Koelewijn and colleagues' divided
 588 attention trials (where listeners heard two talkers presented dichotically, and had to report
 589 both sentences).²³

590 The present study, on the other hand, shows that for an experimental manipulation to elicit
 591 a larger pupillary response than other tasks, it is not enough that the task simply be made
 592 harder. Rather, there is an important distinction between *a task being harder* and *a listener*
 593 *trying harder*; or what, in the terms of a recent consensus paper from a workshop on hearing
 594 impairment and cognitive energy, might be described as the difference between “demands”
 595 and “motivation.”¹⁸ In this light, we can understand why our stimulus manipulations yielded

596 no change in pupillary response: our task required rapid-response target identification, in
597 which listeners had little opportunity to ponder a distorted or partial percept, nor could they
598 later reconstruct whether a target had been present based on surrounding context. Thus, the
599 listener has no recourse by which to overcome the increased task demands, and consequently
600 there should be no difference in motivation, no difference in effort, and no difference in the
601 pupillary response. In contrast, our behavioral “maintain/switch” manipulation did provide
602 an opportunity for the listener to exert effort (in the form of a well-timed mid-trial attention
603 switch) to achieve task success, and the difference in pupillary responses between maintain-
604 and switch-attention trials reflects this difference.

605 **ACKNOWLEDGMENTS**

606 Portions of this work were supported by NIH grants R01-DC013260 to AKCL, F32-DC012456
607 to EL, T32-DC000018 to the University of Washington, and NIH LRP awards to EL and
608 DRM. The authors are grateful to Susan McLaughlin and two anonymous reviewers for
609 helpful suggestions on an earlier draft of this paper, and to Maria Chait for suggesting certain
610 useful post-hoc analyses.

REFERENCES

- [1] S. A. Shamma, M. Elhilali, and C. Micheyl, “Temporal coherence and attention in auditory scene analysis,” *Trends Neurosci.* **34**(3), 114–123 (2011), doi:10.1016/j.tins.2010.11.002.
- [2] B. G. Shinn-Cunningham and V. Best, “Selective attention in normal and impaired hearing,” *Trends in Amplif.* **12**(4), 283–299 (2008), doi:10.1177/1084713808325306.
- [3] E. D. Larson and A. K. C. Lee, “Influence of preparation time and pitch separation in switching of auditory attention between streams,” *J. Acoust. Soc. Am.* **134**(2), EL165 (2013), doi:10.1121/1.4812439.
- [4] J.-P. de Ruiter, H. Mitterer, and N. J. Enfield, “Projecting the end of a speaker’s turn: A cognitive cornerstone of conversation,” *Language* **82**(3), 515–535 (2006), doi:10.1353/lan.2006.0130.
- [5] D. Kahneman and J. Beatty, “Pupil diameter and load on memory,” *Science* **154**(3756), 1583–1585 (1966), doi:10.1126/science.154.3756.1583.
- [6] J. Beatty, “Task-evoked pupillary responses, processing load, and the structure of processing resources,” *Psychol. Bull.* **91**(2), 276–292 (1982), doi:10.1037/0033-2909.91.2.276.
- [7] B. Hoeks and W. J. M. Levelt, “Pupillary dilation as a measure of attention: A quantitative system analysis,” *Beh. Res. Meth. Ins. C.* **25**(1), 16–26 (1993), doi:10.3758/BF03204445.
- [8] S. M. Wierda, H. van Rijn, N. A. Taatgen, and S. Martens, “Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution,” *P. Natl. Acad. Sci. USA* **109**(22), 8456–8460 (2012), doi:10.1073/pnas.1201858109.

- [9] D. McCloy, E. Larson, B. Lau, and A. K. C. Lee, “Temporal alignment of pupillary response with stimulus events via deconvolution,” *J. Acoust. Soc. Am.* **139**(3), EL57–EL62 (2016).
- [10] J. S. Taylor, “Pupillary response to auditory versus visual mental loading: A pilot study using super 8-mm photography,” *Percept. Motor Skill.* **52**(2), 425–426 (1981), doi:10.2466/pms.1981.52.2.425.
- [11] S. Ahern and J. Beatty, “Physiological evidence that demand for processing capacity varies with intelligence,” in M. P. Friedman, J. P. Das, and N. O’Connor (Eds.), *Intelligence and Learning* (Springer, Boston, 1981), no. 14 in NATO Conference Series, p. 121–128, doi:10.1007/978-1-4684-1083-9_9.
- [12] M. H. Papesh and S. D. Goldinger, “Pupil-BLAH-metry: Cognitive effort in speech planning reflected by pupil dilation,” *Atten. Percept. Psychophys.* **74**(4), 754–765 (2012), doi:10.3758/s13414-011-0263-y.
- [13] E. H. Hess and J. M. Polt, “Pupil size in relation to mental activity during simple problem-solving,” *Science* **143**(3611), 1190–1192 (1964), doi:10.1126/science.143.3611.1190.
- [14] A. A. Zekveld, S. E. Kramer, and J. M. Festen, “Pupil response as an indication of effortful listening: The influence of sentence intelligibility,” *Ear Hear.* **31**(4), 480–490 (2010), doi:10.1097/AUD.0b013e3181d4f251.
- [15] T. Koelewijn, A. A. Zekveld, J. M. Festen, J. Rönnberg, and S. E. Kramer, “Processing load induced by informational masking is related to linguistic abilities,” *Int. J. Otolaryngol.* **2012**, article ID 865731 (2012), doi:10.1155/2012/865731.
- [16] M. B. Winn, J. R. Edwards, and R. Y. Litovsky, “The impact of auditory spectral resolution on listening effort revealed by pupil dilation,” *Ear Hear.* **36**(4), e153–e165 (2015), doi:10.1097/AUD.0000000000000145.

- [17] R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay, “Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group ‘white paper’,” *Int. J. Audiol.* **53**(7), 433–445 (2014), doi:10.3109/14992027.2014.890296.
- [18] M. K. Pichora-Fuller, S. E. Kramer, M. A. Eckert, B. Edwards, B. W. Hornsby, L. E. Humes, U. Lemke, T. Lunner, M. Matthen, C. L. Mackersie, G. Naylor, N. A. Phillips, M. Richter, M. Rudner, M. S. Sommers, K. L. Tremblay, and A. Wingfield, “Hearing impairment and cognitive energy: The Framework for Understanding Effortful Listening (FUEL),” *Ear and Hearing* **37**, 5S–27S (2016), doi:10.1097/AUD.0000000000000312.
- [19] A. A. Zekveld, S. E. Kramer, and J. M. Festen, “Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response,” *Ear Hear.* **32**(4), 498–510 (2011), doi:10.1097/AUD.0b013e31820512bb.
- [20] S. E. Kuchinsky, J. B. Ahlstrom, K. I. Vaden, S. L. Cute, L. E. Humes, J. R. Dubno, and M. A. Eckert, “Pupil size varies with word listening and response selection difficulty in older adults with hearing loss,” *Psychophysiology* **50**(1), 23–34 (2013), doi:10.1111/j.1469-8986.2012.01477.x.
- [21] V. Best, F. J. Gallun, C. R. Mason, G. D. Kidd, Jr., and B. G. Shinn-Cunningham, “The impact of noise and hearing loss on the processing of simultaneous sentences,” *Ear Hear.* **31**(2), 213–220 (2010), doi:10.1097/AUD.0b013e3181c34ba6.
- [22] T. Koelewijn, B. G. Shinn-Cunningham, A. A. Zekveld, and S. E. Kramer, “The pupil response is sensitive to divided attention during speech processing,” *Hearing Res.* **312**, 114–120 (2014), doi:10.1016/j.heares.2014.03.010.
- [23] T. Koelewijn, H. de Kluiver, B. G. Shinn-Cunningham, A. A. Zekveld, and S. E. Kramer, “The pupil response reveals increased listening effort when it is difficult to focus attention,” *Hearing Res.* **323**, 81–90 (2015), doi:10.1016/j.heares.2015.02.004.

- [24] D. Kahneman and J. Beatty, “Pupillary responses in a pitch-discrimination task,” *Percept. Psychophys.* **2**(3), 101–105 (1967), doi:10.3758/BF03210302.
- [25] A. K. Nábelek and P. K. Robinson, “Monaural and binaural speech perception in reverberation for listeners of various ages,” *J. Acoust. Soc. Am.* **71**(5), 1242–1248 (1982), doi:10.1121/1.387773.
- [26] D. S. Brungart, “Informational and energetic masking effects in the perception of two simultaneous talkers,” *J. Acoust. Soc. Am.* **109**(3), 1101–1109 (2001), doi:10.1121/1.1345696.
- [27] R. A. Cole, Y. Muthusamy, and M. Fanty, “The ISOLET spoken letter database,” Technical Report 90-004, Oregon Graduate Institute, Hillsboro, OR (1990), paper 205.
- [28] B. G. Shinn-Cunningham, N. Kopco, and T. J. Martin, “Localizing nearby sound sources in a classroom: Binaural room impulse responses,” *J. Acoust. Soc. Am.* **117**(5), 3100–3115 (2005), doi:10.1121/1.1872572.
- [29] L. T. DeCarlo, “Signal detection theory and generalized linear models,” *Psychol. Methods* **3**(2), 186–205 (1998), doi:10.1037/1082-989X.3.2.186.
- [30] C.-F. Sheu, Y.-S. Lee, and P.-Y. Shih, “Analyzing recognition performance with sparse data,” *Behav. Res. Meth.* **40**(3), 722–727 (2008), doi:10.3758/BRM.40.3.722.
- [31] D. R. McCloy and A. K. C. Lee, “Auditory attention strategy depends on target linguistic properties and spatial configuration,” *J. Acoust. Soc. Am.* **138**(1), 97–114 (2015), doi:10.1121/1.4922328.
- [32] E. D. Larson and D. A. Engemann, “pyeparse,” (2015), doi:10.5281/zenodo.14566, version 0.1.0.
- [33] E. Maris and R. Oostenveld, “Nonparametric statistical testing of EEG- and MEG-data,” *J. Neurosci. Meth.* **164**(1), 177–190 (2007), doi:10.1016/j.jneumeth.2007.03.024.

- 707 [34] A. Gramfort, M. Luessi, E. D. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck,
 708 R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. S. Hämäläinen, “MEG and EEG data
 709 analysis with MNE-Python,” *Front. Neurosci.* **7**, paper 267 (2013), doi:10.3389/fnins.
 710 2013.00267.
- 711 [35] J.-M. Hupé, C. Lamirel, and J. Lorenceau, “Pupil dynamics during bistable motion
 712 perception,” *J. Vision* **9**(7), paper 10 (2009), doi:10.1167/9.7.10.
- 713 [36] E. D. Larson and A. K. C. Lee, “The cortical dynamics underlying effective switching of
 714 auditory spatial attention,” *NeuroImage* **64**, 365–370 (2013), doi:10.1016/j.neuroimage.
 715 2012.09.006.
- 716 [37] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech recognition
 717 with primarily temporal cues,” *Science* **270**(5234), 303–304 (1995), doi:10.1126/science.
 718 270.5234.303.
- 719 [38] B. C. J. Moore and B. R. Glasberg, “Formulae describing frequency selectivity as a
 720 function of frequency and level, and their use in calculating excitation patterns,” *Hearing*
 721 *Res.* **28**(2-3), 209–225 (1987), doi:10.1016/0378-5955(87)90050-5.

722 **LIST OF FIGURES**

723 **1** (Color online) Illustration of “maintain” and “switch” trial types in Experiment 1.
 724 In the depicted “switch” trial (heavy dashed line), listeners would hear cue “AB” in
 725 a male voice, attend to the male voice (“QU”) for the first half of the trial, switch to
 726 the female voice (“OM”) for the second half of the trial, and respond once (to the “O”
 727 occurring at 3.1–3.6 s). In the depicted “maintain” trial (heavy solid line), listeners
 728 would hear cue “AA” in a male voice, maintain attention to the male voice (“QUJR”)
 729 throughout the trial, and not respond at all. In the depicted trials, a button press
 730 anytime during timing slot 2 would be counted as response to the “O” at 2–2.5 s,
 731 which is a “foil” in both trial types illustrated; a button press during slot 3 would
 732 be counted as response to the “O” at 3.1–3.6 s (which is considered a target in the
 733 switch-attention trial and a foil in the maintain-attention trial), and button presses
 734 at any other time would be counted as non-foil false alarms. Note that “O” tokens
 735 never occurred in immediately adjacent timing slots (unless separated by the switch
 736 gap) so response attribution to targets or foils was unambiguous.

2 (Color online) Box-and-swarm plots of between-condition differences in listener sensitivity for Experiment 1. Boxes show first & third quartiles and median values; individual data points correspond to each listener; asterisks indicate comparisons with corresponding coefficients in the statistical model that were significantly different from zero. (a) Main effects of attention (higher sensitivity in maintain than switch trials), reverberation (higher sensitivity in anechoic than reverberant trials), and talker gender (mis)match (higher sensitivity in trials with different-gendered target and masker talkers). (b) Two-way interactions; the difference between anechoic and reverberant trials was significantly larger in the gender-match (MM) than in the gender-mismatch (MF) condition. (c) Three-way interaction (no statistically significant differences). ** = $p < 0.01$; *** = $p < 0.001$.

3 (Color online) Box-and-swarm plots of between-condition differences in reaction time for Experiment 1. Boxes show first & third quartiles and median values; individual data points correspond to each listener; asterisks indicate comparisons with corresponding coefficients in the statistical model that were significantly different from zero. (a) Main effects of attention (faster reaction time in maintain than switch trials), reverberation (faster reaction time in anechoic than reverberant trials), and talker gender (mis)match (faster reaction time in trials with trials with different-gendered target and masker talkers). (b) Two-way interactions (no statistically significant differences). (c) Three-way interaction (no statistically significant difference). * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; MM = matching talker genders; MF = mismatched talker genders.

4 (Color online) Deconvolved pupil size (mean ± 1 standard error across subjects) for (a) reverberant versus anechoic trials, (b) talker gender-match versus -mismatch trials, and (c) maintain- versus switch-attention trials, with trial schematics showing the timecourse of stimulus events (compare to Figure 1). Hatched region shows temporal span of statistically significant differences between time series. The onset of statistically significant divergence (vertical dotted line) of the maintain/switch conditions is in close agreement with the end of the cue. a.u. = arbitrary units (see Section II.A.5 for explanation of “effort”).

5 (Color online) Illustration of “maintain” and “switch” trial types in Experiment 2. The short-gap version is depicted; timing of long-gap trial elements (where different) are shown with faint dashed lines. In the depicted “switch” trial (heavy dashed line), listeners would hear cue “AU” in a male voice, attend to the male voice (“EO”) for the first half of the trial and the female voice (“DE”) for the second half of the trial, and respond once (to the “O” occurring at 2–2.5 seconds). In the depicted “maintain” trial (heavy solid line), listeners would hear cue “AA” in a male voice, attend to the male voice (“EOPO”) throughout the trial, and respond twice (once for each “O”).

6 (Color online) Box-and-swarm plots of between-condition differences in listener sensitivity for Experiment 2. Boxes show first & third quartiles and median values; individual data points correspond to each listener; asterisks indicate comparisons with corresponding coefficients in the statistical model that were significantly different from zero. (a) Main effects of attention (higher sensitivity in maintain than switch trials), spectral degradation (higher sensitivity in 20-channel than 10-channel vocoded trials), and switch gap duration (higher sensitivity in trials with a short gap). (b) Two-way interactions: the difference between long- and short-gap trials was greater (more negative) in the 10-channel-vocoded trials and in the maintain-attention trials. (c) Three-way interaction (not significant). * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$.

7 (Color online) Box-and-swarm plots of between-condition differences in reaction time for Experiment 2. Boxes show first & third quartiles and median values; individual data points correspond to each listener; asterisks indicate comparisons with corresponding coefficients in the statistical model that were significantly different from zero. (a) Main effects of attention, spectral degradation, and gap duration (faster response time in trials with 20-channel vocoding, and in long-gap trials). (b) Two-way interactions (larger difference in reaction times between long- and short-gap trials in the 10- versus the 20-channel condition). (c) Three-way interaction (no statistically significant difference). *** = $p < 0.001$.

8 (Color online) Deconvolved pupil size (mean ± 1 standard error across subjects) for (a) 10- versus 20-band vocoded stimuli, (b) 200 versus 600 ms mid-trial switch gap durations, and (c) maintain- versus switch-attention trials, with trial schematics showing the timecourse of stimulus events (compare to Figure 5). Hatched region shows temporal span of statistically significant differences between time series. The late-trial divergence in (b) is attributable to the delay of stimulus presentation in the long-gap condition; the onset of divergence in (c) aligns with the end of the cue, as in Experiment 1 (see Figure 4c). a.u. = arbitrary units (see Section II.A.5 for explanation of “effort”).