Article:
Mepham, A., Knight, S., McGarrigle, R. orcid.org/0000-0003-1704-1135 et al. (2 more authors) (2025) Pupillometry Reveals the Role of Signal-to-Noise Ratio in Adaption to Linguistic Interference Over Time. Journal of Speech, Language and Hearing Research. ISSN 1092-4388

https://doi.org/10.1044/2025_JSLHR-24-00658

Pupillometry reveals the role of SNR in adaption to linguistic interference over time

Alex Mepham[1], Sarah Knight[2], Ronan McGarrigle[3], Lyndon Rakusen[4], & Sven Mattys[1]

[1] University of York, UK

[2] Newcastle University, UK

[3] University of Leeds, UK

[4] University of Arizona, USA

Corresponding author: Sven Mattys, sven.mattys@york.ac.uk

Conflict of interest statement: There are no conflicts of interest for this study

Abstract

Purpose. Studies of speech-in-speech listening show that intelligible maskers are more detrimental to target perception than unintelligible maskers, an effect we refer to as linguistic interference. Research also shows that performance improves over time through adaptation. The extent to which the speed of adaptation differs for intelligible and unintelligible maskers and whether this pattern is reflected in changes in listening effort are open questions.

Method. In this pre-registered study, native English listeners transcribed English sentences against an intelligible masker (time-forward English talkers) vs. an unintelligible masker (time-reversed English talkers). Over 50 trials, transcription accuracy and task-evoked pupil response (TEPR) were recorded, along with self-reported effort and fatigue ratings. In Experiment 1, we used an adaptive procedure to ensure a starting performance of ~50% correct in both conditions. In Experiment 2, we used a fixed signal-to-noise ratio (SNR: -1.5 dB) for both conditions.

Results. Both experiments showed performance patterns consistent with linguistic interference. The speed of adaptation depended on the SNR. When the SNR was higher for the intelligible masker condition as a result of the 50% starting performance across conditions (Experiment 1), adaptation was faster for that condition; TEPRs were not affected by trial number or condition. When the SNR was fixed (Experiment 2), adaptation was similar in both conditions but TEPRs decreased faster in the unintelligible than intelligible masker condition. Self-reported ratings of effort and fatigue were not affected by masker conditions in either experiment.

Conclusions. Learning to segregate target speech from maskers depends on both the intelligibility of the maskers and the SNR. We discuss ways in which auditory stream formation is automatic or requires cognitive resources.

Keywords: pupillometry; linguistic interference; perceptual adaptation

I. INTRODUCTION

Listeners face various challenges when listening to speech in a background of competing talkers. The target signal can be degraded due to spectro-temporal overlap with the competing speech, creating interference at the cochlear level (*energetic masking*, e.g., Culling & Stone, 2017). In this case, performance is determined primarily by the extent to which the target signal can be "glimpsed" through the masker (Barker & Cooke, 2007) in regions of reduced spectro-temporal overlap, which in turn depends in part on the signal-to-noise ratio (SNR) between the target and the masker. Listening can also be compromised by non-energetic properties of the competing signal (*informational masking,* Cooke et al., 2008; Kidd & Colburn, 2017). Informational masking can take various forms, including misallocation of masker components to the target speaker (phonetic features, segments, words) and attentional capture due to phonological or semantic familiarity with the masker (Cooke et al., 2008; Summers & Roberts, 2020). In other words, even in the absence of spectro-temporal overlap, listeners must successfully isolate the masker as a to-be-ignored stream through a process of auditory 'object formation' and direct their attention accordingly (Shinn-Cunningham, 2008). Attentional capture by the masker is often illustrated by the fact that it is more difficult to understand a target speaker when the language of the competing talkers is known to the listener than when it is unknown (Brouwer et al., 2012; Calandruccio et al., 2013; Cooke et al., 2008; Garcia-Lecumberri & Cooke, 2006; Kilman et al., 2014; Mepham et al., 2022; Van Engen & Bradlow, 2007). We refer to this specific type of informational masking as "linguistic interference," which is the topic of the present experiments.

While the above studies have investigated defining characteristics of linguistic interference, little is known about how linguistic interference changes over time. Studies show that the perception of distorted or masked speech tends to improve between the start and the end of a block of trials. For instance, Cooke et al. (2022) showed rapid adaptation at

the beginning of an experiment across a wide range of degradations, and plateauing performance subsequently. Importantly, Bent et al. (2009) found that the point at which a learning plateau is reached depends on the type of adverse condition. In their experiment, perception of eight-channel noise-vocoded speech started at 70% correct and plateaued around ~83% after 60 sentences, whereas speech perception in six-talker babble at 0 dB SNR started at 67% correct and plateaued around 74% after 40 sentences. Not all distortions benefit from repeated exposure, however. Lie et al. (2024) found learning effects in temporally- and spectrally-modulated noise but less so in stationary noise and for degradations with a low speech reception threshold (see also Rhebergen et al., 2006; Versfeld et al., 2021).

Critical for the question of linguistic interference, Mepham et al. (2022) showed that improvements in sentence transcription were slower when the competing talkers were intelligible to the listeners (time-forward speech in a known language) than when they were unintelligible (time-reversed speech or speech in an unknown language). Thus, learning to ignore an intelligible masker was harder than learning to ignore an unintelligible masker, presumably because of the sustained informational masking caused by familiar aspects of the intelligible masker and the listener's inability to inhibit those familiar aspects through practice. Note, however, that Versfeld et al. (2021), who measured changes in speech reception threshold (SRT) over 87 sentences, did not find substantial differences in SRT improvement between time-forward and time-reversed maskers, suggesting that Mepham et al.'s effect might be sensitive to methodological considerations (transcription performance vs. SRT) and listener engagement with the task.

It is unclear whether Mepham et al.'s (2022) differences in adaptation as a function of masker intelligibility are reflected in a change in effort. Since performance and effort do not necessarily pattern together, measures of effort can provide complementary information

about cognitive resource allocation that is not reflected in the accuracy score (Kuchinsky et al., 2013; Winn & Teece, 2021). Of particular interest is whether Mepham et al.'s faster improvement for the unintelligible-masker condition might have come at the cost of increased effort or, alternatively, whether the growing familiarity with the unintelligible masker might have made the task less (rather than more) effortful over time. The former would suggest that effort is a compensatory mechanism, with higher performance achieved at the expense of higher effort, while the latter would suggest that effort shows a simple inverse relationship to performance, with higher performance requiring lower effort (e.g., Ohlenforst et al., 2017; Wendt et al., 2018; Winn et al., 2018; Wu et al., 2016; Zekveld et al., 2018). A better understanding of the performance/effort relationship could thus also provide some insight into whether adaptation is conscious and attention-driven (i.e., compensatory; Huyck & Johnsrude, 2012) or relatively automatic (cf. segregation as a 'primitive' process, Bregman, 1990; Sussman, 2017).

Listening effort can be assessed using pupillometry (for reviews, see McGarrigle et al., 2014; Van Engen & McLaughlin, 2018; Winn et al., 2018; Zekveld et al., 2018). The task-evoked pupil response (TEPR) has been used to assess cognitive effort when listening to speech in modulated noise (Koelewijn et al., 2012, 2014a; McLaughlin et al., 2021; Ohlenforst et al., 2018; Paulus et al., 2020; Wendt et al., 2018), time-compressed and noise-vocoded speech (Paulus et al., 2020), accented speech (Brown et al., 2020; McLaughlin & Van Engen, 2020), multi-talker babble (Koelewijn et al., 2012, 2014a; Ohlenforst et al., 2018; Wendt et al., 2018), non-native speech (Borghini & Hazan, 2018), and trained vs. untrained voices (Biçer et al., 2023). In each case, increased listening demands (caused by, e.g., more adverse SNRs or accented speech) were reflected in greater TEPR, at least when intelligibility was moderate to good.

Most pupillometry experiments average pupil responses across trials to capture sensitivity to a particular signal degradation. However, recent studies have examined how pupil responses change across trials within a block (e.g., Brown et al., 2020; Paulus et al., 2020; Versfeld et al., 2021; McGarrigle et al., 2021a). For instance, Brown et al. (2020) measured TEPRs over 50 trials while participants listened to native English or Mandarin-accented English sentences. TEPRs in the early trials were larger in the non-native-accented than native-accented condition, which suggests that non-native-accented speech was initially more effortful to understand. However, TEPRs also decreased faster in that condition, indicating that listeners quickly adapted to the new mapping between accented speech and native phonemes. This effect can be thought of as a form of "levelling-out" between the two conditions, with the easy and hard conditions eventually involving comparable levels of effort. Paulus et al. (2020) also measured changes in adaptation and TEPRs over time during processing of quiet, noise-vocoded, masked, and time-compressed speech. Listeners adapted to noise-vocoded and time-compressed speech only. Mean TEPRs generally declined over time across all conditions, with no difference in this linear trend between conditions.

Although the above studies provide evidence for a general decrease in listening effort during adaptation to degraded speech, their contribution to our understanding of how linguistic interference changes over time is limited. Indeed, in the Brown et al. (2020) study, transcription performance was not a factor of interest, with performance intentionally kept over 90% in both the native- and the non-native-accented conditions. Therefore, it is impossible to assess any potential performance/effort trade-off. Likewise, while Paulus et al. (2020) compared a range of degraded conditions, none of them allowed an interpretation specific to the linguistic content of the masker independent of its energetic content. As an exception, Versfeld et al. (2021) compared changes in pupil response for time-forward and time-reversed maskers. Although they reported an improvement in performance, as

mentioned earlier, pupil response did not change significantly across trials and there was no difference between the two masker types. Therefore, the extent to which changes in linguistic interference over time are supported by changes in effort is an open question.

In the present study, we followed Mepham et al.'s (2022) approach and investigated how listeners learn to ignore the content of competing speech over time. Across 50 trials, we assessed speech recognition performance and TEPRs in native English speakers listening to target English sentences in English two-talker babble (time-forward intelligible masker) compared to the same English two-talker babble played backward (time-reversed unintelligible masker). The difference between time-forward and time-reversed maskers allows linguistic interference to be assessed while controlling for the long-term average frequency spectra of the two maskers, i.e., their average energetic masking. Of interest is whether listeners' effort tracks the ease of speech perception over time (i.e., increasing performance associated with decreasing effort) or, instead, reflects compensation mechanisms (i.e., increasing performance associated with increasingly high effort). The number of trials we chose (50) was based on similar studies that reported learning effects plateauing between 40 and 50 sentences (Bent et al., 2009), around 40 sentences (Lie et al., 2024), and between 30 and 60 sentences (Versfeld et al., 2021).

We also investigated whether the levelling-out pattern observed by Brown et al. (2020) between easy and hard conditions generalizes to time-reversed vs. time-forward maskers. Specifically, we asked whether effort would start higher in the time-forward than time-reversed condition, but drop to comparable levels after 50 trials. As well as measuring TEPRs, we were also interested in assessing the extent to which physiological changes in effort may be reflected in changes in perceived effort and/or fatigue over time. Previous research has shown that subjective and physiological indices of effort may be tapping into related, but separate, dimensions and may therefore provide complementary information

(Alhanbali et al., 2019; McGarrigle et al., 2021a; Strand et al., 2018). These self-report measures may therefore shed light on the perceived costs of adaptation to linguistic interference. For example, McGarrigle et al. (2021a) found that, although subjective ratings of effort did not change over time, subjective ratings of fatigue increased.

Finally, Mepham et al. (2022) and Paulus et al. (2020, masking condition) used fixed SNRs across participants and conditions. Although this approach means that long-term energetic masking is controlled across conditions, differences in performance between conditions are likely to be present at the start of each block, which makes the effect of time difficult to compare between conditions. To address this limitation, our study included an initial adaptive procedure which established participants' individual 50% SRT in the time-forward and time-reversed conditions. This meant that participants started the two conditions at the same performance level. We chose 50% because it is the performance level at which effort has been shown to peak (Wendt et al., 2018), in addition to helping to mitigate the risk of floor or ceiling effects. Unlike Versfeld et al. (2021), who measured performance change in terms of the SNR needed to maintain 50% accuracy, we elected not to manipulate the SNR beyond the initial 50% calibration and simply measure changes in accuracy from that starting point. We reasoned that keeping the same SNR across all the trials of a condition would allow us to interpret any changes in performance and effort independent of signal quality.

Our hypotheses were as follows: First, we predicted that sentence transcription would improve over time, reflecting listeners' ability to better stream targets from maskers as familiarity with the task and stimuli increased (Bent et al., 2009; Cooke et al., 2022; Erb et al., 2012, 2013; Lie et al., 2024). In particular, following Mepham et al. (2022), we expected that the improvement would be more pronounced in the time-reversed condition because this condition does not elicit linguistic interference. Second, we predicted that TEPRs would decrease in both conditions, indicating reductions in effort over time in line with Brown et al.

194 (2020) and Paulus et al. (2020). A key question, however, was whether the decrease in effort
195 would differ between the time-forward and the time-reversed conditions. A less pronounced
196 decrease (i.e., more stable pattern) for the time-forward condition would reflect the sustained
197 cognitive demands imposed by linguistic interference and would be in line with the expected
198 performance pattern. Alternatively, the time-forward masker may show a steeper decrease
199 over time (i.e., reduction in effort) because the intelligibility of the masker, while initially a
200 disadvantage, could make it easier to isolate as a to-be-ignored stream through a process of
201 'object formation' (Shinn-Cunningham, 2008). Third, we predicted that subjective effort
202 ratings will be higher in the time-forward condition because of its linguistic content, but will
203 not show significant changes over time in either condition, whereas fatigue ratings will
204 increase in both conditions reflecting similarly-adverse perceptual demands, as per
205 McGarrigle et al. (2021a).

206 ## II. EXPERIMENT 1

207 ### A. METHOD

208 #### *1. Participants*

209 Forty native British English listeners (10 male, 29 female, one non-binary) aged between
210 18 and 30 years (M = 21.10, SD = 3.48) with no known history of hearing impairments
211 participated in the experiment. Four of them listeners described their language status as
212 bilingual from birth, speaking British English and one other language. One of the 40
213 participant was excluded from the pupil analyses due to a high proportion of missing pupil
214 data. Using the Westfall et al. (2014) power analysis approach, it was determined that 39
215 participants were required to achieve statistical power of 0.9 to reach an effect size d = .4,
216 with 100 stimuli in a counterbalanced design (n = 50 stimuli in each condition). Details can
217 be found in the pre-registration documents referenced in the Data Availability Statement. All

participants had pure-tone audiometry (PTA) measures ≤ 20 dB HL at 0.5, 1, 2, and 4 kHz (*M* = 4.63, *SD* = 3.33). The University of York Department of Psychology ethics committee approved all procedures for the two experiments in this study (ethics reference number: 747). Listeners participated in this experiment either for course credit or were compensated at a rate of 6.00 GBP per hour. All participants provided written informed consent to take part in this study.

2. Materials

a. Target stimuli. The target stimuli were taken from Mepham et al. (2022). These were two-hundred sentences from the first 20 British-adapted and modernized Harvard/IEEE sentence lists (IEEE, 1969), spoken by a female native British English speaker (see Appendix A). Each target sentence had five keywords (e.g., "The PLAY SEEMS DULL and QUITE STUPID", keywords capitalized). All sentences were recorded in a single-walled sound-attenuated room at a 44.1 kHz sampling rate with 16-bit resolution using Audacity© using a Shure SM58 microphone and a RME Fireface UFX II built-in soundcard. Sentence duration ranged from 1.59s to 3.16s (*M* = 2.20s, *SD* = .24s). The mean fundamental frequency (F0) of the target sentences was 203Hz—see the Masker Stimuli section for further details. The F0 and vocal tract length (VTL) of all target sentences were then adjusted to an F0 of 210Hz. This value was 15 Hz below and above the high-F0 and low-F0 maskers, respectively. The F0 and VTL were edited following the procedure described in Darwin et al. (2003; see also Smith et al., 2007; Gaudrain et al., 2009). VTL was manipulated alongside F0 to improve the naturalness of the two streams, as both indices have been shown to contribute to the perception of voice identity (e.g., Skuk & Schweinberger, 2014).

b. Masker stimuli. The masker stimuli were also taken from Mepham et al. (2022). They consisted of 64 sentences from Lists 1-4 of the English BKB-R corpus (Bench et al., 1979) spoken by a female native British English speaker. This speaker was different from the

speaker who recorded the target sentences. The BKB-R sentences are simple sentences with three to four keywords (e.g., “The POSTMAN SHUT the GATE”, keywords capitalized). A full list of the BKB-R sentences used in this study is available in Appendix B. Each sentence was recorded a minimum of four times, of which the two best exemplars were kept. All sentences were manually edited using Praat (Boersma & Weenik, 2019) to remove silences at the beginning and end of the sentences. This was done through visual inspection of the spectrogram. One of the exemplars was used to create Set A, and the other exemplar was used to create Set B. The Set A sentences were concatenated into a continuous stream, henceforth Stream A. The same was done with the Set B sentences, henceforth Stream B. The mean F0 of the Stream A sentences was 208Hz and the mean F0 of the Stream B sentences was 205Hz. Sentence order within each BKB-R list was the same in both streams, but the order of the lists differed in each stream.

Following the same procedure as the one used for the target sentences, the F0 and VTL of each sentence within the streams were edited to create a high-F0 version (mean of 225 Hz) and a low-F0 version (mean of 195 Hz) of each stream. These two values are approximately 15 Hz above and below the average F0 of the target sentences (210Hz), respectively. The high-F0 version of Stream A was combined with the low-F0 version of Stream B to constitute the two-talker masker. The use of a single voice to create the two masker speakers was designed to avoid idiosyncratic dominance of one masker voice over the other, as was done in Mepham et al. (2022; see also Smith et al., 2024, for a similar procedure).

c. Experimental mixtures. The target and masker stimuli were mixed online during the experiment. For each trial, the two-talker masker speech stream was randomly sampled for the duration of the target sentence, plus two seconds preceding it and two seconds following it. The masker level was fixed at 65 dB SPL and the level of the target sentence was determined on a participant-by-participant basis by the adaptive procedure (see the Procedure

section). The masker speech stream was sampled randomly without replacement, resulting in a different masker speech stream for each trial and for each participant.

d. Subjective effort and fatigue measures

Two questions were used to examine subjective effort and fatigue. For effort, we adapted a question from the NASA task load index assessing mental demands (Hart & Staveland, 1988), a commonly used subjective measure of effort (e.g., Dimitrijevic et al., 2019; McGarrigle et al., 2021a; Pals et al., 2019; Peng & Wang, 2019; Strand et al., 2018): "How hard did you have to work to understand what was said for the previous ten sentences? (0: Very low; 20: Very high.)". For fatigue, we used the Oncology Nursing Society (ONS) Brief Fatigue Inventory: English (Burke & Naylor, 2020; Picou & Ricketts, 2014): "Please rate your fatigue (weariness, tiredness) by choosing the one number that best describes your fatigue right NOW. (0: No fatigue; 10: As bad as you can imagine.)".

3. Procedure

At the beginning of the experiment, listeners underwent audiometric threshold testing. The main experiment then comprised two blocks, one for each listening condition (time-forward masker and time-reversed masker). Each block comprised two parts. The first part was an adaptive procedure to obtain the listener's 50% SRT for the target-masker mixtures in that condition. The second part was the main speech recognition task, which contained 50 trials played at the individual 50% SRT established by the adaptive procedure. The order of the two blocks was counterbalanced across participants. In both the adaptive procedure and the speech recognition task, listeners were asked to repeat aloud as much of the target talker as they could. To familiarize the listeners with the target voice and minimize the chances that they would accidentally track one of the masker voices instead, three practice trials were played before the adaptive procedure began. These were three sentences from the

Harvard/IEEE corpus, unused in the adaptive or recognition tasks, spoken by the target speaker. They were played at 65 dB SPL with no masker talkers.

For the adaptive procedure, the intensity of the masker was fixed at 65 dB SPL and the intensity of the target talker changed from trial to trial. All listeners started with an SNR of +10 dB, with step sizes of 6 dB at the start, 4 dB after the first reversal, then 2 dB after the second reversal and the remaining reversals. The adaptive procedure followed a one-up one-down staircase for eight reversals. The 50% SRT was calculated by fitting a logistic function to the performance and corresponding SNRs for each reversal during the adaptive procedure. If the logistic function failed to fit or returned an infinite value, an approximate 50% accuracy SNR was calculated by taking the mean SNR value over all eight reversals. This back-up method was used on 15 occasions, i.e., 18.75% of the time.

At the beginning of each block of the main speech recognition task and after every 10 trials thereafter, listeners were presented with the subjective effort and fatigue rating questions (see above). The questions were presented on a monitor and participants scored their responses on a sliding scale using a computer mouse.

For both the adaptive procedure and the speech recognition task, listeners were instructed to focus on a fixation cross that appeared on the monitor for the duration of the trial. They were cued to respond at the end of the masker, which itself finished 2s after the end of the target sentence (see Experimental Mixtures section). There was a 4s gap between the end of their answer and the fixation cross for the next trial. The listeners' responses were scored online by the experimenter against the five keywords for each sentence. Listeners were offered a break between the first and the second blocks. The eye-tracking equipment was always recalibrated at the beginning of the second block. The entire experiment lasted under an hour.

4. Equipment

Listeners completed the experiment in a single-walled sound-attenuated room. PTA testing was conducted using a Kamplex Diagnostic Audiometer AD 25. During the main listening task, listeners were positioned 65 cm away from a 24″ LCD flat monitor, which displayed a fixation cross. The listener's head was stabilized on a head-and-chin rest which was secured to the edge of a table. Stimulus presentation was programmed using a bespoke Python script in PsychoPy (Pierce et al., 2019). Auditory stimuli were presented via Denon DJ DN-HP700 headphones. A microphone was positioned inside the test booth so that verbal responses could be heard and scored online by the experimenter who listened via headphones. Pupil size was recorded using the EyeLink 1000 Plus at a sampling rate of 500 Hz.

Pupil size was recorded for the right eye only. It was captured as a continuous recording for each trial and recorded as an integer number corresponding to the number of thresholded pixels in the camera's pupil image. Typical pupil area can range between 100 and 10,000 units, with a precision of 1 unit, corresponding to a resolution of 0.01 mm for a 5 mm pupil diameter (SR Research Ltd). The arbitrary pupil-size unit provided by the EyeLink 1000 Plus system was used for data analysis and in the figures. The desktop-mounted eye-tracker camera was positioned between the listener and the computer monitor at a distance of 55 cm from the listener at 0° azimuth angle. The camera was aligned to the center of the monitor and positioned just below the bottom of the monitor to maximize the trackable range without obscuring the listener's view of the monitor.

5. Analysis

a. Speech recognition performance

Speech recognition performance was calculated as the proportion of keywords correctly reported (out of 5) for each target sentence. Generalized linear mixed-effect models with a

logit link and binomial distribution were run in R (version 4.4.1) via RStudio (version 2024.4.2) using the *glmer* function from the *lme4* package (Bates et al., 2015). The models assessed mean differences in proportion of keywords correctly reported as a function of Masker (time-forward, time-reversed) and Time (trials 1 to 50). Listener and Sentence were used as random intercepts, with Masker|Listener and Masker|Sentence as random slopes, following Barr et al.'s (2013) recommendation to use the most complex random effects structure supported by the data. The BOBYQA optimizer was used to aid model convergence (Powell, 2009). Initially, a full model of all main effects and interaction terms was used and the contribution of each term was assessed using likelihood ratio testing (i.e., comparing the full model to a reduced model with the term of interest removed). Where models failed to converge, random slopes and intercepts that were highly correlated or where the variance could not be estimated were removed, resulting in the Masker|Sentence slope being removed from all analyses.

b. Pupillometry

Following recommendations from Winn et al. (2018), pupil data were pre-processed to remove noise. Any missing values in pupil size (caused by, e.g., blinks or pupil non-detection) were coded as *NA* and linearly interpolated using the *gazeR* package (Version 0.1, Geller et al., 2020). Trials containing more than 20% missing data were removed from the analysis. One participant had 21 trials with more than 20% missing data and was removed from the pupillometric analysis following procedures outlined in the pre-registration. Eleven trials were removed across the remaining participants (0.25% of all trials in the data set).

Baseline-correction was performed on a trial-by-trial basis. Of the 2 s of masker speech preceding the onset of the target sentence, we only used the 1s immediately preceding the onset of the target sentence to avoid pupil responses that might reflect sensory onset adaptation rather than a genuine dilation baseline. The mean pupil size value recorded during

this 1-s window was then subtracted from every sample recorded after target speech onset. We chose the use the mean pupil size instead of peak size or latency to mean size (e.g., Zekveld et al., 2010) because mean size is sensitive to masker manipulations and time (McGarrigle et al., 2021a,b) and because mean and peak size indices have been shown to converge on similar patterns (Neagu et al., 2023). Linear mixed-effect modelling using the *lmer* function in the *lme4* package (Bates et al., 2015) was conducted to examine TEPR as a function of Masker (time-forward, time-reversed) and Time (trials 1 to 50). Listener and Sentence were used as random intercepts, with Masker|Listener and Masker|Sentence as random slopes. The rest of the procedure was the same as that used for the behavioral data.

c. Subjective measures

Due to the small number of data points for the subjective ratings of effort and fatigue, a repeated-measures analysis of variance (*aov* function from the *stats* package) was run instead of linear mixed-effects models. The dependent variables were the effort and fatigue ratings, which were rescaled as a subtraction from the baseline effort/fatigue ratings at the start of the condition before participants completed any trials. The independent variables were Masker (time-forward, time-reversed) and Time (ordered factor with 5 levels corresponding to trials 10, 20, 30, 40, and 50). A by-participant error term [Error(participant/(Masker*Time))] was included to analyze the data as repeated-measures.

B. RESULTS

1. 50% Speech Reception Threshold (SRT)

The average SNR required to achieve 50% correct transcription was higher in the time-forward condition (M = -0.38 dB, SD = 2.15 dB) than in the time-reversed condition (M = -2.65 dB, SD = 1.82 dB), $t(39) = -7.88$. $p < .001$. This SNR difference (2.27 dB) illustrates the

cost of ignoring an intelligible (time-forward) masker compared to an unintelligible (time-reversed) masker; a hallmark of linguistic interference.

2. Speech recognition performance

Figure 1 displays speech recognition performance as a function of Masker and Time. As expected from the adaptive procedure, performance started around 50% correct in both masker conditions, and there was no significant effect of Masker, $B = 0.154$, $SE = 0.200$, $X^2(1) = 0.59$, $p = .443$. A significant effect of Time, $B = 0.790$, $SE = 0.085$, $X^2(1) = 87.58$, $p < .001$, indicated that performance improved over the course of the block. However, an interaction between Masker and Time, $B = -0.394$, $SE = 0.118$, $X^2(1) = 11.17$, $p < .001$, showed that the improvement was faster in the time-forward than time-reversed condition. A subsequent analysis of the two conditions separately indicated that the effect of Time was nevertheless significant in both conditions, $B = 0.804$, $SE = 0.085$, $X^2(1) = 89.65$, $p < .001$, and $B = 0.405$, $SE = 0.082$, $X^2(1) = 24.25$, $p < .001$, respectively.

3. Pupillometry

Figure 2 shows TEPRs as a function of Masker and Time. There was no significant effect of Masker, $B = -0.478$, $SE = 14.384$, $X^2(1) = 0.001$, $p = .973$, or Time, $B = -0.388$, $SE = 0.265$, $X^2(1) = 2.137$, $p = .144$. There was no significant interaction between Masker and Time, $B = -0.012$, $SE = 0.376$, $X^2(1) = 0.001$, $p = .975$. Figure 3 shows the average TEPR pattern over the course of a trial for each masker condition collapsed arbitrarily across bins of ten trials (trials 1-10, 11-20, 21-30, 31-40, 41-50) to illustrate TEPR patterns across time, following Brown et al. (2020).

4. Subjective Measures

Figure 4 shows the subjective ratings of effort and fatigue, calculated as change from the baseline ratings collected prior to the first trial of the main task. For effort, the main effects

and the interaction were non-significant (all Fs < 0.93, all ps > .34). For fatigue, there was a significant effect of Time, $F(2.35, 89.40) = 14.42$, $p < .001$, $\eta_p^2 = .27$, with fatigue ratings increasing over the course of the block. There was no effect of Masker, $F(1, 38) = 0.38$, $p = .541$, nor any Masker by Time interaction, $F(3.17, 120.32) = 0.26$, $p = .866$.

C. DISCUSSION

Our results replicate previous findings that the recognition of degraded speech improves over the course of an experiment (Bent et al., 2009; Cooke et al., 2022; Erb et al., 2012, 2013; Mepham et al., 2022). However, contrary to our expectation and Mepham et al.'s (2022) results, performance improved more, rather than less, in the time-forward than time-reversed masker condition. Mepham et al. (2022) claimed that the slower improvement for the time-forward condition in their study reflected the sustained linguistic interference in that condition relative to the easier segregation of a "neutral" masker, i.e., a masker with no linguistic content. Our present results suggest that, on the contrary, the linguistic, and hence familiar nature of the time-forward masker made it easier for listeners to identify it as a separate auditory object (Shinn-Cunningham, 2008) and therefore learn to ignore it over the course of the experiment.

An important methodological difference between the present study and Mepham et al.'s (2022) is that, in Mepham et al., the SNR was fixed at -3 dB for both the time-forward and time-reversed conditions. In the present study, the SNR was adjusted so that performance in both masker conditions was matched at ~50% at the start of the blocks. As a consequence, listeners in the present study required a higher (i.e., more favorable) SNR in the time-forward than the time-reversed condition. It is possible that the more favorable SNR in the time-forward condition made it easier to glimpse the target sentences through the masker (Barker & Cooke, 2007), and hence, to learn to segregate it from the masker over time. In contrast, learning to ignore the masker in the time-reversed condition might have been harder because

438 the masker was louder than the target, therefore acting as a consistently interfering distractor
439 throughout the experiment.

440 Although the pupillometric data showed a small numerical TEPR decrease over time (cf.
441 Brown et al., 2020; Paulus et al., 2020; but see Versfeld et al., 2021), this trend was not
442 significant. Since the adaptive procedure provided plenty of familiarization with the task and
443 stimuli, this might have made the decrease in effort in the course of the block less
444 pronounced than expected. We did not find an effect of masker type on TEPR, which is in
445 contrast with Koelewijn et al.'s (2012) finding of larger TEPRs in intelligible (single
446 competing talker) than unintelligible (noise) maskers. We also did not find a difference in
447 pupil size change over time between the time-forward and time-reversed conditions, which is
448 inconsistent with Brown et al.'s (2020) levelling-out pattern where the pupil size decrease
449 was steeper for hard than easy listening conditions.

450 The methodology of our study differed from Koelewijn et al.'s (2012) and Brown et al.'s
451 (2020) in several ways. In the Koelewijn et al. study, pupil responses were measured while
452 listeners underwent an SRT adaptive procedure, whereas, in this study, they were measured
453 during listening tasks with a fixed SNR. In the Brown et al. study, a dual-task paradigm was
454 used, whereas we used a single task. Finally, performance in the Brown et al. study was near
455 ceiling by design, whereas it was in the 50-60% range in the current study. These
456 methodological differences make the three studies difficult to compare. Our results indicate
457 that the more favorable SNR for the time-forward condition (-0.38 dB, compared to -2.38 dB
458 for the time-reversed condition) made the two masker conditions comparable in terms of both
459 performance and effort. The similar TEPR slope in the two conditions can tentatively be
460 interpreted as evidence that the inhibition of the linguistic content of the time-forward masker
461 came at no extra cognitive cost compared to the inhibition of the non-linguistic content of the
462 time-reversed masker.

463 The subjective effort ratings in our study did not show any patterning with the
464 pupillometric data, consistent with the majority of studies in which pupil dilation,
465 performance, and subjective effort are compared (Koelewijn et al., 2012; McGarrigle et al.,
466 2021a; Pichora-Fuller et al., 2016; Strand et al., 2018). In contrast, the subjective fatigue
467 ratings increased over time, alongside an increase in performance, in line with previous
468 findings (McGarrigle et al., 2021a). However, reported fatigue did not differ between the
469 masker conditions. Therefore, the faster improvement in the time-forward condition did not
470 come at the expense of increased perceived or physiological /expenditure.

471 As mentioned earlier, the faster improvement in the time-forward than the time-reversed
472 condition could be due to the more favorable SNR, and thus higher audibility of the target,
473 for the intelligible than unintelligible masker. Therefore, in Experiment 2, we used a single
474 SNR for both conditions, which is similar to the procedure in Mepham et al. (2022). If the
475 higher SNR for the time-forward condition in Experiment 1 was responsible for its faster
476 improvement over time, this advantage should disappear when the SNR is the same for both
477 maskers, as audibility and opportunities for glimpses would be identical in both conditions.
478 However, if the faster improvement for the time-forward condition truly demonstrates the
479 listener's ability to better learn to stream and inhibit an intelligible masker than an
480 unintelligible masker, the pattern in Experiment 2 should replicate that in Experiment 1.

481 ## III. EXPERIMENT 2

482 ### A. METHOD

483 #### *1. Participants*

484 Forty native British English listeners (11 male, 26 female, three non-binary) aged
485 between 18 and 31 years (M = 20.52, SD = 2.86) with no known history of hearing
486 impairments participated in the experiment. All listeners described their language status as

monolingual, speaking British English from birth, and had pure-tone audiometry (PTA) measures ≤ 20 dB HL at 0.5, 1, 2, and 4 kHz (M = 5.59, SD = 2.94). All other information was the same as in Experiment 1.

2. Materials

The target and masker stimuli were the same as in Experiment 1, except that, for both the time-forward and time-reversed conditions, the target sentences were played at 63.5 dB SPL (the average of the target levels in the time-forward and time-reversed conditions of Experiment 1) and the maskers were played at 65.0 dB SPL, as in Experiment 1, resulting in a -1.5 dB SNR throughout the experiment. The materials used to examine the subjective ratings of effort and fatigue were the same as in Experiment 1.

3. Procedure and Equipment

The procedure was the same as in Experiment 1, except that the adaptive procedure used at the start of each block in Experiment 1 was replaced with a practice session. In each practice session, participants listened without responding to five target sentences played without a masker and the same five sentences in the presence of a masker at -1.5 dB SNR. Participants were then presented with five new target and masker mixtures at that SNR and were asked to repeat as much of the target sentence as they could. Feedback was provided on how many keywords they reported correctly. Initially, as per our preregistration, we had planned to use a pseudo-adaptive procedure using randomly sampled SNRs to emulate the adaptive procedure of Experiment 1. However, we found that, if listeners were presented with the time-forward condition first, they often erroneously reported the content of the masker rather than the target. For this reason, we used the procedure described above. Additionally, if the listener consistently reported the masker rather than the target in the five practice trials, the experimenter entered the testing room and encouraged them to pay attention to the quieter

511 talker until they could reliably distinguish targets from maskers. The procedure for the
512 condition then restarted. The rest of the procedure was the same as in Experiment 1. The
513 equipment was the same as in Experiment 1.

514 ***4. Analysis***

515 Analyses of behavioral and pupillometric data were identical to those in Experiment 1.

516 **B. RESULTS**

517 ***1. Speech recognition performance***

518 Figure 5 shows speech recognition performance as a function of Masker and Time. There
519 was a significant effect of Masker, $B = 1.024$, $SE = 0.096$, $X^2(1) = 66.78$, $p < .001$, with better
520 performance in the time-reversed ($M = 0.685$, $SD = 0.307$) than time-forward condition ($M =$
521 0.466, $SD = 0.364$). A significant effect of Time, $B = 0.407$, $SE = 0.080$, $X^2(1) = 25.99$, $p <$
522 $.001$, indicated an improvement in performance over the course of a block. The interaction
523 between Masker and Time was not significant, $B = 0.109$, $SE = 0.116$, $X^2(1) = 0.89$, $p = .346$,
524 suggesting that the rate at which listeners' performance improved did not differ between the
525 time-reversed and time-forward conditions.

526 ***2. Pupillometry***

527 Figure 6 shows mean TEPR as a function of Masker and Time. There was no significant
528 effect of Masker, $B = -15.794$, $SE = 15.599$, $X^2(1) = 1.020$, $p = .312$, but an effect of Time
529 showed that TEPR decreased significantly over the course of a block, $B = -1.966$, $SE = 0.272$,
530 $X^2(1) = 51.841$, $p < .001$. An interaction between Masker and Time, $B = -0.100$, $SE = 0.385$,
531 $X^2(1) = 6.744$, $p = .009$, indicated that this decrease was steeper for the time-reversed than
532 time-forward condition. This can also be seen in Figure 7.

533

534 ***Subjective Measures***

535 Figure 8 presents the subjective ratings of effort and fatigue as a change from baseline
536 ratings. For effort, there was no significant effect of Masker, $F(1, 39) = 0.12$, $p = .728$, Time,
537 $F(3.02, 117.68) = 0.26$, $p = .859$, or interaction, $F(3.24, 126.31) = 0.61$, $p = .620$. For fatigue,
538 there was a significant effect of Time, $F(2.21, 86.25) = 25.27$, $p < .001$, $\eta_p^2 = .39$, with
539 increasing fatigue as the block progressed. There was no effect of Masker, $F(1, 39) = 0.289$, p
540 $= .594$, and no significant interaction, $F(2.72, 106.15) = 0.06$, $p = .973$.

541 **C. DISCUSSION**

542 Transcription performance showed the expected advantage for unintelligible (time-
543 reversed) over intelligible (time-forward) maskers, confirming the existence of linguistic
544 interference when the same SNR is used. However, Experiment 2 did not replicate the
545 interactive pattern in Experiment 1, where performance improved more steeply for the
546 intelligible than unintelligible masker condition. In Experiment 2, masker intelligibility did
547 not facilitate learning; in fact, the numerical pattern resembled the performance pattern in
548 Mepham et al. (2022), with faster improvement in the unintelligible than intelligible masker
549 condition. Thus, when glimpses are controlled through a fixed SNR (as in Experiment 2 and
550 Mepham et al., 2022), the intelligibility of a masker hinders, rather than facilitates, learning.

551 This interpretation was supported by the pupillometry data, where TEPRs decreased more
552 slowly in the intelligible masker condition (the "hard" condition) than in the unintelligible
553 masker condition (the "easy" condition). Therefore, learning to ignore a meaningful masker
554 might be more effortful than learning to ignore a meaningless one. This finding broadly
555 supports the finding of Koelewijn et al (2012) that intelligible maskers require more cognitive
556 effort to ignore than unintelligible ones. However, this result is in contrast with Brown et al.'s
557 (2020) observation that the effort associated with adaptation to a native accent (the "easy"

558 condition) decreased more slowly than to a non-native accent (the "hard" condition), which
559 suggests that radically different adaptation mechanisms might be at play for speech degraded
560 by an external masker, as in our study, and speech degraded at the source, as in the Brown et
561 al. study (cf. Mattys et al., 2012).

562 As in Experiment 1, the subjective ratings of effort showed no significant difference
563 between the two masker conditions, with subjective effort remaining constant over time,
564 again suggesting a lack of association between subjective and physiological measures of
565 listening effort. With respect to the fatigue ratings, these, too, were similar to those in
566 Experiment 1; subjective fatigue increased over time and it did so similarly for both masker
567 types. This suggests that the process of learning to inhibit an intelligible (versus an
568 unintelligible) competing talker during speech recognition evokes physiological changes that
569 do not appear to reach conscious awareness in terms of perceived effort or perceived fatigue.

570 **IV. GENERAL DISCUSSION**

571 The aim of this study was to investigate how listeners learn to ignore competing talkers
572 over time. We asked whether adaptation to a masker is easier or harder if the masker is
573 intelligible compared to unintelligible. We also asked whether changes in performance are
574 reflected in physiological (pupillometric) and self-reported measures of listening effort. We
575 measured native English speakers' transcription accuracy of target English sentences in the
576 presence of intelligible (time-forward English two-talker babble) vs. unintelligible (time-
577 reversed English two-talker babble) maskers over 50 trials. Trial-by-trial changes in pupil
578 responses were calculated and subjective ratings of effort and fatigue were collected every
579 ten trials. Experiment 1 used an adaptive procedure to set the starting performance at
580 approximately 50% across the two masker conditions. Experiment 2 used a fixed SNR across
581 masker conditions.

582 **A. SPEECH RECOGNITION PERFORMANCE**

583 Our results showed clear evidence of linguistic interference. The SNR needed to achieve
584 50% correct in the time-forward condition was 2.27 dB higher than in the time-reversed
585 condition (Experiment 1). Likewise, with a fixed SNR in both conditions, performance in the
586 time-forward condition was 21% lower than in the time-reversed condition (Experiment 2).
587 Both experiments showed modest but consistent performance improvement over the course
588 of 50 trials in all conditions (between 5% and 10%).

589 With respect to whether learning was faster in the intelligible or unintelligible masker
590 condition, the pattern depended on the relative SNR levels of the two conditions. If the SNR
591 favored the intelligible masker condition (a way of achieving parity of initial performance
592 across maskers; Experiment 1), learning was faster for that condition. However, if a fixed
593 SNR was used (Experiment 2), that pattern disappeared and, if anything, participants were
594 less successful and slower to acclimatize to the masker with linguistic content; a pattern
595 similar to that reported by Mepham et al. (2022).

596 These results and previous ones (e.g., Versfeld et al., 2021) suggest that the effect of
597 masker intelligibility on adaptive learning is fragile and highly sensitive to methodological
598 considerations. Specifically, they show that masker intelligibility is not the only factor
599 determining the ease with which listeners learn to stream a target from a masker. Target
600 audibility, i.e., opportunities for glimpses through the masker, also plays a role. Thus, the
601 benefit of masker familiarity for gradual object formation and inhibition (Shinn-Cunningham,
602 2008) seen in Experiment 1 was observed only if the target was sufficiently audible to be
603 perceptually foregrounded and for clear differentiation between the target and the masker to
604 occur. In other words, maskers with a linguistic content can be inhibited successfully over
605 time, but only when the audibility of the target (relative to the masker) is favorable. On the
606 other hand, when audibility was matched across intelligible and unintelligible maskers, the

benefit of masker familiarity was lost and there was numerical evidence that the linguistic content of the masker actually led to slower, rather than faster, improvements in performance over time. Note that a reason why Experiment 2 showed the interaction pattern of Mepham et al. (2022) numerically rather than statistically could be the more favorable SNR in Experiment 2 (-1.5 dB) than in Mepham et al. (-3 dB). The more favorable SNR in Experiment 2, and the resulting higher performance starting point, could mean that there was a more restricted performance range in the unintelligible condition, and hence, less potential for improvement than in Mepham et al.

B. PUPILLOMETRY

TEPRs decreased over time in both experiments, but this pattern was only significant in Experiment 2. This difference could be due in part to the more extensive training provided by the adaptive procedure in Experiment 1 compared to the shorter practice session in Experiment 2. Greater familiarity with the procedure and stimuli at the start of the main listening task in Experiment 1 could have made the usual decrease in effort over time (see Zekveld et al. 2018) less detectable.

More importantly, in Experiment 1, the greater performance improvement in the intelligible than unintelligible condition was not mirrored in the pupil data; that is, there was no difference in the TEPR change over time between the two masker types. There are two ways of interpreting this result. First, it could suggest that enhanced object formation due to the linguistic content of the time-forward masker came at no extra cognitive cost. This would indicate that masker segregation based on linguistic grouping operates as a primitive (Bregman, 1990) or automatic (Sussman, 2017) process and requires limited attention. Alternatively, given that better performance most often correlates with lower effort (e.g., Miles et al., 2017; Zekveld et al., 2010), especially in the 30%-70% accuracy range (Wendt et al., 2018), we could have expected effort to decrease more steeply for the time-forward

632 condition. The fact that it did not suggests that the faster adaptation to the intelligible masker
633 might have been cognitively costly and required sustained attention (e.g., Huyck &
634 Johnsrude, 2012).

635 When glimpsing opportunities were controlled across the intelligible and unintelligible
636 maskers through the use of a single SNR (Experiment 2), the clear linguistic interference
637 observed in the performance data was not reflected in the TEPR data at the start of the block.
638 This was surprising given prior reports of larger TEPRs for meaningful than meaningless
639 maskers (e.g., Koelewijn et al., 2012; Wendt et al., 2018; Zekveld & Kramer, 2014). Instead,
640 a linguistic interference effect emerged as the block progressed—something that would have
641 gone undetected had the data been averaged across the Time variable. Again, this result can
642 be interpreted in terms of either early adaptation to the unintelligible masker or more effortful
643 adaptation to the intelligible one. Both interpretations suggest that the cognitive cost of
644 dealing with an intelligible masker remains high over the course of an experiment relative to
645 that associated with dealing with an unintelligible masker. Unlike Experiment 1, the TEPR
646 pattern of change over time in Experiment 2 showed a clear difference as a function of
647 masker condition. Thus, whether performance, as opposed to SNR, is controlled between
648 intelligible and unintelligible masker conditions is critical to establish the automaticity vs.
649 effortfulness of adaptation, with adaptation being similarly automatic (or effortful) for both
650 masker types when initial performance is matched (Experiment 1) and intelligible maskers
651 being more effortful to segregate when a single SNR is used (Experiment 2).

652 **C. SUBJECTIVE MEASURES**

653 Self-reported effort did not vary as a function of masker type or time in either experiment.
654 These results are in stark contrast with the pupillometric data, but consistent with studies that
655 have failed to show common patterning between pupil data and subjective effort ratings
656 (McGarrigle et al., 2021a; Moore & Picou, 2018; Pichora-Fuller et al., 2016; Strand et al.,

2018; Versfeld et al., 2021; but see Koelewijn et al., 2012). These data hint at the listeners' potential lack of awareness of the changes in cognitive effort associated with both linguistic interference and the learning process.

Self-reported fatigue increased over time in both experiments, as it did in McGarrigle et al. (2021a). However, it did not show any differences between masker types, even in Experiment 2, where linguistic interference was visible in both the performance and pupil data. Taken together, the subjective responses suggest that participants might be more attuned to the cumulative fatigue induced by undertaking a cognitively demanding task than the differential demands imposed by listening in the presence of an intelligible (versus unintelligible) masker.

D. CONCLUSION

The results of these two experiments show the detrimental effect of an intelligible masker compared to an unintelligible masker (i.e., linguistic interference) on: (a) the SNR needed to reach 50% accuracy (Experiment 1), (b) sentence recognition accuracy (Experiment 2), and (c) listening effort reflected in the TEPR (Experiment 2). Across both experiments, sentence recognition accuracy increased over the course of each experimental block, showing adaptation to the maskers. Participants' subjective ratings of fatigue increased over time, but there was no difference between masker conditions or change over time in subjective ratings of effort. Whether linguistic interference affected the rate of adaptation over the course of a block depended on the relative SNR of the intelligible and unintelligible maskers. If the SNR favored the intelligible masker condition, adaptation was relatively rapid compared to the unintelligible masker condition. If a fixed SNR was used, no significant difference between conditions was observed in terms of rate of adaptation, but there was evidence that adapting to the unintelligible masker was less effortful, suggesting that linguistic interference – and its effect on effort – remained relatively high. Taken together, the results indicate that the extent

682 to which linguistic interference can be dealt with by the listener, and the ensuing cognitive
683 effort that this process demands, depends critically on the audibility of the target relative to
684 the masker.

Acknowledgements

This research was supported by research grants from the Leverhulme Trust (RPG-2018-152) and the Economic and Social research Council (ES/W010488/1).

Data Availability Statement

The experiments in this study were preregistered (Experiment 1: https://osf.io/m4b57/; Experiment 2: https://osf.io/pmhsx/). All materials, analysis scripts, data files, and appendices are available from these OSF links.

693 **References**

694 Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of listening effort
695 are multidimensional. *Ear and Hearing*, *40*, 1084-1097.

696 Barker, J., & Cooke, M. (2007). Modelling speaker intelligibility in noise. *Speech*
697 *Communication*, *49*, 402-417.

698 Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for
699 confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*,
700 *68*, 255-278.

701 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models
702 using lme4. *Journal of Statistical Software, 67,* 1-48.

703 Bench, J., Kowal, Å., & Bamford, J. (1979). The Bkb (Bamford-Kowal-Bench) sentence lists
704 for partially-hearing children. *British Journal of Audiology, 13,* 108-112.

705 Bent, T., Buchwald, A., & Pisoni, D. B. (2009). Perceptual adaptation and intelligibility of
706 multiple talkers for two types of degraded speech. *Journal of the Acoustical Society of*
707 *America, 126,* 2660-2669.

708 Biçer, A., Koelewijn, T., & Başkent, D. (2023). Short implicit voice training affects listening
709 effort during a voice cue sensitivity task with vocoder-degraded speech. *Ear and*
710 *Hearing*, *44*, 900-916.

711 Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer (version 6.1)
712 [computer program]. http://www.praat.org.

713 Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased
714 for non-native listeners: A pupillometry study. *Frontiers in Neuroscience, 12,* 1-13.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *Journal of the Acoustical Society of America, 131,* 1449-1464.

Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology, 73,* 1431-1443.

Burke, L. A., & Naylor, G. (2020). Daily-life fatigue in mild to moderate hearing impairment: An ecological momentary assessment study. *Ear and Hearing, 41,* 1518-1532.

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., & Bradlow, A. R. (2013). Masking release due to linguistic and phonetic dissimilarity between the target and masker speech. *American journal of Audiology, 22,* 157-164.

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *Journal of the Acoustical Society of America, 123,* 414-427.

Cooke, M., Scharenborg, O., & Meyer, B. T. (2022). The time course of adaptation to distorted speech. *Journal of the Acoustical Society of America, 151,* 2636-2646.

Culling, J. F., & Stone, M. A. (2017). Energetic masking and masking release. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The Auditory System at the Cocktail Party* (pp. 41-73). Springer Nature.

738 Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency
739 and vocal-tract length changes on attention to one of two simultaneous talkers.
740 *Journal of the Acoustical Society of America, 114*, 2913-2922.

741 Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2019). Neural indices of
742 listening effort in noisy environments. *Scientific Reports*, *9,* 11278.

743 Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2012). Auditory skills and brain morphology
744 predict individual differences in adaptation to degraded speech. *Neuropsycholia, 50,*
745 2154-2164.

746 Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2013). The brain dynamics of rapid perceptual
747 adaptation to adverse listening conditions. *Journal of Neuroscience, 33,* 10688-10697.

748 Garcia-Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-
749 native consonant perception in noise. *Journal of the Acoustical Society of America,*
750 *119,* 2445-2454.

751 Gaudrain, E., Li, S., Ban, V. S., & Patterson, R. (2009). The role of glottal pulse rate and
752 vocal tract length in the perception of speaker identity. *Proceedings of Interspeech*
753 *2009*, Brighton, United Kingdom.

754 Geller. J., Winn, M. B., Mahr, T., & Mirman, D. (2020). GazeR: A package for processing
755 gaze position and pupil size data. *Behavior Research Methods, 52*, 2232-2255.

756 Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index):
757 Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.),
758 *Human mental workload* (pp. 139–183). North-Holland.

759 Huyck, J. J., & Johnsrude, I. S. (2012). Rapid perceptual learning of noise-vocoded speech
760 requires attention. *Journal of the Acoustical Society of America*, *131*, EL236-EL242.

761 IEEE Subcommittee on Subjective Measurements (1969). IEEE recommended practices for
762 speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*,
763 227–246.

764 Kidd, G. Jr., & Colburn, H. S. (2017). Informational masking in speech recognition. In J. C.
765 Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The Auditory System at*
766 *the Cocktail Party* (pp. 75-109). Springer Nature.

767 Kilman, L., Zekveld, A., Hällgren, M., & Rönnberg, J. (2014). The influence of non-native
768 language proficiency on speech perception performance. *Frontiers in Psychology, 5,*
769 1-9.

770 Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2014a). The
771 pupil response is sensitive to divided attention during speech processing. *Hearing*
772 *research, 312,* 114-120.

773 Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2014b). The influence of
774 informational masking on speech perception and pupil response in adults with hearing
775 impairment. *Journal of the Acoustical Society of America, 135,* 1596-1606.

776 Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012).
777 Processing load induced by informational masking is related to linguistic abilities.
778 *International Journal of Otolaryngology, 2012,* 1-11.

779 Kuchinsky S. E., Ahlstrom J. B., Vaden K. I. Jr., Cute S. L., Humes L. E., Dubno J. R., &
780 Eckert M. A. (2013). Pupil size varies with word listening and response selection
781 difficulty in older adults with hearing loss. *Psychophysiology*, *50*, 23-34.

782 Lie, S., Zekveld, A. A., Smits, C., Kramer, S. E., & Versfeld, N. J. (2024). Learning effects in
783 speech-in-noise tasks: Effect of masker modulation and masking release. *Journal of*
784 *the Acoustical Society of America*, *156*, 341-349.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive processes*, *27*, 953-978.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International Journal of Audiology*, *53*, 433-445.

McGarrigle, R., Rakusen, L., & Mattys, S. (2021a). Effortful listening under the microscope: Examining relations between pupillometric and subjective markers of effort and tiredness from listening. *Psychophysiology, 58,* 1-22.

McGarrigle, R., Knight, S., Rakusen, L., Geller, J., & Mattys, S. (2021b). Older adults show a more sustained pattern of effortful listening than young adults. *Psychology and aging*, *36*, 504.

McLaughlin, D. J., Braver, T. S., & Peelle, J. E. (2021). Measuring the subjective cost of listening effort using a discounting task. *Journal of Speech, Language, and Hearing Research, 64,* 337-347.

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *Journal of the Acoustical Society of America*, *147*, EL151-EL156.

Mepham, A., Bi, Y., & Mattys, S. L. (2022). The time-course of linguistic interference during native and non-native speech-in-speech listening. *Journal of the Acoustical Society of America, 152,* 954-969.

Miles, K., McMahon, C., Boisvert, I., Ibrahim, R., De Lissa, P., Graham, P., & Lyxell, B. (2017). Objective assessment of listening effort: Coregistration of pupillometry and EEG. *Trends in Hearing*, *21*, 1-13.

809 Moore, T. M., & Picou, E. M. (2018). A potential bias in subjective ratings of mental
810 effort. *Journal of Speech, Language, and Hearing Research*, *61*, 2405-2421.

811 Ohlenforst, B., Wendt, D., Kramer, S. E., Naylor, G., Zekveld, A. A., & Lynner, T. (2018).
812 Impact of SNR, masker type and noise reduction processing on sentence recognition
813 performance and listening effort as indicated by the pupil dilation response. *Hearing*
814 *Research, 365,* 90-99.

815 Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J.,
816 & Kramer, S. E. (2017). Impact of stimulus-related factors and hearing impairment on
817 listening effort as indicated by pupil dilation. *Hearing Research*, *351*, 68-79.

818 Pals, C., Sarampalis, A., van Dijk, M., & Başkent, D. (2019). Effects of additional low-pass-
819 filtered speech on listening effort for noise-band-vocoded speech in quiet and in
820 noise. *Earing Research, 40,* 3-17.

821 Paulus, M., Hazan, V., & Adank, P. (2020). The relationship between talker acoustics,
822 intelligibility, and effort in degraded listening conditions. *Journal of the Acoustical*
823 *Society of America, 147,* 3348-3359.

824 Peng, Z. E., & Wang, L. M. (2019). Listening effort by native and nonnative listeners due to
825 noise, reverberation, and talker foreign accent during English speech perception.
826 *Journal of Speech, Language, and Hearing Research*, *62*, 1068-1081.

827 Pierce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E.,
828 & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavioral*
829 *Research Methods*, *51*, 195-203.

830 Picou, E. M., & Ricketts, T. A. (2014). The effect of changing the secondary task in dual-task
831 paradigms for measuring listening effort. *Ear and Hearing, 35,* 611-622.

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge, 26.

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *Journal of the Acoustical Society of America*, *120*(6), 3988-3997.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*, 182-186.

Skuk, V. G., & Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research, 57*, 285-296.

Smith, D. R. R., Walters, T. C., & Patterson, R. D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *Journal of the Acoustical Society of America, 122,* 3628-3639.

Smith, E. D., Holt, L. L., & Dick, F. (2024). A one-man bilingual cocktail party: linguistic and non-linguistic effects on bilinguals' speech recognition in Mandarin and English. *Cognitive Research: Principles and Implications*, *9*, 1-17.

SR Research Ltd. EyeLink 1000 Plus User Manual. Version 1.0.12. Mississauga, Ontario, Canada: SR Research Ltd., 2017.

Strand., J. F., Brown, V. A., Marchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring Listening Effort: Convergent Validity, Sensitivity, and Links with Cognitive and Personality Measures. *Journal of Speech, Language, and Hearing Research, 61,* 1463-1486.

Summers, R. J., & Roberts, B. (2020). Informational masking of speech by acoustically similar intelligible and unintelligible interferers. *Journal of the Acoustical Society of America, 147,* 1113-1125.

Sussman, E. S. (2017). Auditory scene analysis: An attention perspective. *Journal of Speech, Language, and Hearing Research*, *60*, 2989-3000.

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America, 121,* 519-526.

Van Engen, K. J., & McLaughlin, D, J. (2018). Eyes and ears: Using eye tracking and pupillometry to understand challenges to speech recognition. *Hearing Research, 369,* 56-66.

Versfeld, N. J., Lie, S., Kramer, S. E., & Zekveld, A. A. (2021). Informational masking with speech-on-speech intelligibility: Pupil response and time-course of learning. *Journal of the Acoustical Society of America*, *149*, 2353-2366.

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research, 369,* 67-78.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 2020-2045.

Winn, M. B. & Teece, K. H. (2021). Listening effort is not the same as speech intelligibility score. *Trends in Hearing, 25,* 1-26.

878 Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice
879 for using pupillometry to measure listening effort: An introduction for those who want
880 to get started. *Trends in Hearing, 22,* 1-32.

881 Wu, Y. H., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (2016). Psychometric functions of
882 dual-task paradigms for measuring listening effort. *Ear and Hearing*, *37*, 660-670.

883 Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to
884 auditory stimuli: Current state of knowledge. *Trends in Hearing*, *22,* 1-25.

885 Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of
886 listening conditions: Insights from pupillometry. *Psychophysiology*, *51*, 277-284.

887 Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of
888 effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*, 480-
889 490.

890

Figure Captions

Figure 1. Mean proportion of keywords correctly reported as a function of Masker (time-forward, time-reversed) and Time (trials 1 to 50). The shaded area represents 95% confidence intervals.

Figure 2. Mean task-evoked pupil-response (TEPR) as a function of Masker (time-forward, time-reversed) and Time (trials 1 to 50). The shaded area represents 95% confidence intervals. The TEPR scale is based on the arbitrary pupil-size units provided by the EyeLink 1000 Plus system.

Figure 3. Mean task-evoked pupil-response (TEPR) binned by groups of ten trials, as per Brown et al. (2020). The leftmost vertical line indicates the start of the masker period used for baseline calculation (t = -1s, i.e., 1s after the onset of the masker sentence) and the second vertical line indicates the start of the target sentence (t = 0s). The TEPR scale is based on the arbitrary pupil-size units provided by the EyeLink 1000 Plus system.

Figure 4. Mean ratings of effort and fatigue after trials 10, 20, 30, 40, and 50 relative to baseline ratings (i.e., average rating minus baseline rating, out of 20 for effort and out of 10 for fatigue).

Figure 5. Mean proportion of keywords correctly reported as a function of Masker (time-forward, time-reversed) and Time (trials 1 to 50). The shaded area represents 95% confidence intervals.

Figure 6. Mean task-evoked pupil-response (TEPR) as a function of Masker (time-forward, time-reversed) and Time (trials 1 to 50). The shaded area represents 95% confidence intervals. The TEPR scale is based on the arbitrary pupil-size units provided by the EyeLink 1000 Plus system.

Figure 7. Mean task-evoked pupil-response (TEPR) binned by groups of ten trials, as per Brown et al. (2020). The leftmost vertical line indicates the start of the masker period used for baseline calculation (t = -1s, i.e., 1s after the onset of the masker sentence) and the second vertical line indicates the start of the target sentence (t = 0s). The TEPR scale is based on the arbitrary pupil-size units provided by the EyeLink 1000 Plus system.

Figure 8. Mean ratings of effort and fatigue after trials 10, 20, 30, 40, and 50 relative to baseline ratings (i.e., average rating minus baseline rating, out of 20 for effort and out of 10 for fatigue).