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

Alex Mepham1, Sarah Knight2, Ronan McGarrigle3, Lyndon Rakusen4, & Sven Mattys1

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

1. **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. (2020) and Paulus et al. (2020). A key question, however, was whether the decrease in effort would differ between the time-forward and the time-reversed conditions. A less pronounced decrease (i.e., more stable pattern) for the time-forward condition would reflect the sustained cognitive demands imposed by linguistic interference and would be in line with the expected performance pattern. Alternatively, the time-forward masker may show a steeper decrease over time (i.e., reduction in effort) because the intelligibility of the masker, while initially a disadvantage, could make it easier to isolate as a to-be-ignored stream through a process of ‘object formation’ (Shinn-Cunningham, 2008). Third, we predicted that subjective effort ratings will be higher in the time-forward condition because of its linguistic content, but will not show significant changes over time in either condition, whereas fatigue ratings will increase in both conditions reflecting similarly-adverse perceptual demands, as per McGarrigle et al. (2021a).

1. **EXPERIMENT 1**

# **METHOD**

## ***Participants***

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

## ***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.

1. ***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.

1. ***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.

1. **RESULTS**
2. ***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.

1. ***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, *Χ²*(1)= 0.59, *p* = .443. A significant effect of Time, *B* = 0.790, *SE* = 0.085, *Χ²*(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, *Χ²*(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, *Χ²*(1)= 89.65, *p* < .001, and *B* = 0.405, *SE* = 0.082, *Χ²*(1)= 24.25, *p* < .001, respectively.

1. ***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, *Χ²*(1)= 0.001, *p* = .973, or Time, *B* = -0.388, *SE* = 0.265, *Χ²*(1)= 2.137, *p* = .144. There was no significant interaction between Masker and Time, *B* = -0.012, *SE* = 0.376, *Χ²*(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).

1. ***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 *F*s< 0.93, all *p*s > .34). For fatigue, there was a significant effect of Time, *F*(2.35, 89.40) = 14.42, *p* < .001, *ηp*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.

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

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

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

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

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

1. **EXPERIMENT 2**
2. **METHOD**
3. ***Participants***

Forty native British English listeners (11 male, 26 female, three non-binary) aged between 18 and 31 years (*M* = 20.52, *SD* = 2.86) with no known history of hearing 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.

1. ***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.

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

1. ***Analysis***

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

1. **RESULTS**
2. ***Speech recognition performance***

Figure 5 shows speech recognition performance as a function of Masker and Time. There was a significant effect of Masker, *B* = 1.024, *SE* = 0.096, *Χ²*(1)= 66.78, *p* < .001, with better performance in the time-reversed (*M* = 0.685, *SD* = 0.307) than time-forward condition (*M* = 0.466, *SD* = 0.364). A significant effect of Time, *B* = 0.407, *SE* = 0.080, *Χ²*(1)= 25.99, *p* < .001, indicated an improvement in performance over the course of a block. The interaction between Masker and Time was not significant, *B* = 0.109, *SE* = 0.116, *Χ²*(1)= 0.89, *p* = .346, suggesting that the rate at which listeners’ performance improved did not differ between the time-reversed and time-forward conditions.

1. ***Pupillometry***

Figure 6 shows mean TEPR as a function of Masker and Time. There was no significant effect of Masker, *B* = -15.794, *SE* = 15.599, *Χ²*(1)= 1.020, *p* = .312, but an effect of Time showed that TEPR decreased significantly over the course of a block, *B* = -1.966, *SE* = 0.272, *Χ²*(1)= 51.841, *p* < .001. An interaction between Masker and Time, *B* = -0.100, *SE* = 0.385, *Χ²*(1)= 6.744, *p* = .009, indicated that this decrease was steeper for the time-reversed than time-forward condition. This can also be seen in Figure 7.

***Subjective Measures***

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

1. **DISCUSSION**

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

This interpretation was supported by the pupillometry data, where TEPRs decreased more slowly in the intelligible masker condition (the “hard” condition) than in the unintelligible masker condition (the “easy” condition). Therefore, learning to ignore a meaningful masker might be more effortful than learning to ignore a meaningless one. This finding broadly supports the finding of Koelewijn et al (2012) that intelligible maskers require more cognitive effort to ignore than unintelligible ones. However, this result is in contrast with Brown et al.’s (2020) observation that the effort associated with adaptation to a native accent (the “easy” condition) decreased more slowly than to a non-native accent (the “hard” condition), which suggests that radically different adaptation mechanisms might be at play for speech degraded by an external masker, as in our study, and speech degraded at the source, as in the Brown et al. study (cf. Mattys et al., 2012).

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

1. **GENERAL DISCUSSION**

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

**A. SPEECH RECOGNITION PERFORMANCE**

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

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

These results and previous ones (e.g., Versfeld et al., 2021) suggest that the effect of masker intelligibility on adaptive learning is fragile and highly sensitive to methodological considerations. Specifically, they show that masker intelligibility is not the only factor determining the ease with which listeners learn to stream a target from a masker. Target audibility, i.e., opportunities for glimpses through the masker, also plays a role. Thus, the benefit of masker familiarity for gradual object formation and inhibition (Shinn-Cunningham, 2008) seen in Experiment 1 was observed only if the target was sufficiently audible to be perceptually foregrounded and for clear differentiation between the target and the masker to occur. In other words, maskers with a linguistic content can be inhibited successfully over time, but only when the audibility of the target (relative to the masker) is favorable. On the 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 condition. The fact that it did not suggests that the faster adaptation to the intelligible masker might have been cognitively costly and required sustained attention (e.g., Huyck & Johnsrude, 2012).

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

**C. SUBJECTIVE MEASURES**

Self-reported effort did not vary as a function of masker type or time in either experiment. These results are in stark contrast with the pupillometric data, but consistent with studies that have failed to show common patterning between pupil data and subjective effort ratings (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.

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

**References**

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

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

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

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

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

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

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

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

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lie, S., Zekveld, A. A., Smits, C., Kramer, S. E., & Versfeld, N. J. (2024). Learning effects in speech-in-noise tasks: Effect of masker modulation and masking release. *Journal of 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.

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

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

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

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

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

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

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

Picou, E. M., & Ricketts, T. A. (2014). The effect of changing the secondary task in dual-task 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.

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

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

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

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

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

**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).

Figure 1



Figure 2



Figure 3



Figure 4



Figure 5



Figure 6



Figure 7



Figure 8



# **APPENDIX A: Anglicized-Modernized Harvard/IEEE Target Sentences (IEEE, 1969)**

**List 1**

1. A large size in shoes is hard to sell.
2. The boy’s canoe slid on the smooth planks.
3. Glue the paper to the dark blue background.
4. It’s quite easy to tell the depth of a well.
5. These days a chicken leg is not a rare dish.
6. Rice is often served in round bowls.
7. The juice of lemons makes fine punch.
8. The box was thrown beside the parked lorry.
9. The pigs were fed chopped corn and cabbage.
10. We faced four hours of steady work.

**List 2**

1. The girl at the booth sold fifty tickets.
2. The boy was there when the sun rose.
3. A rod is used to catch pink salmon.
4. The source of the huge river is a clear spring.
5. Kick the ball straight and follow through.
6. Help the woman get back on her feet.
7. A pot of tea helps to pass the evening.
8. Smoky fires lack flame and heat.
9. A soft cushion broke the man’s fall.
10. The salt breeze came across the sea.

**List 3**

1. It is a pleasure to read verse out loud.
2. The small pup chewed a hole in the sock.
3. The fish twisted and turned on the bent hook.
4. Press the trousers and sew on the shirt button.
5. The swan dive was far short from perfect.
6. The beauty of the view stunned the young boy.
7. Two blue fish swam in a tank.
8. Her purse was full of useless rubbish.
9. The horse reared and threw the tall rider.
10. It snowed, rained and hailed all in the same morning.

**List 4**

1. What joy there is in living.
2. Hoist the load onto your left shoulder.
3. Take the winding path to reach the lack.
4. Note closely the size of the fuel tank.
5. Wipe the grease of his dirty face.
6. Mend the coat before you go out.
7. The wrist was badly sprained and hung limp.
8. The stray cat gave birth to kittens.
9. The young girl gave no clear response.
10. The meal was cooked before the bell rang.

**List 5**

1. Mesh wire keeps the chicks inside.
2. In the early days a king ruled the state.
3. The ship was torn apart by the sharp reef.
4. Sickness kept him home in the third week.
5. The wide road shimmered in the hot sun.
6. The lazy cow lay in the cool grass.
7. Lift the square stone over the fence.
8. A rope will bind seven books together.
9. Hop over the fence and plunge in.
10. The friendly group left the local store.

**List 6**

1. Both lost their lives in the raging storm.
2. The frosty air passed through her coat.
3. The crooked maze failed to fool the mouse.
4. Counting quickly leads to wrong answers.
5. The show was a flop from the very beginning.
6. A saw is a tool used for cutting wood.
7. The wagon moved on well oiled wheels.
8. March the soldiers past the next hill.
9. A lot of sugar makes tea sweet.
10. Place the rosebush near the porch steps

**List 7**

1. Those last words were a strong statement.
2. We talked about the side show at the circus.
3. Use a pencil to write the first draft.
4. He ran halfway to the hardware store.
5. The bell rang to end third period.
6. A small stream cut across the field.
7. Cars and busses were stuck in the snow drifts.
8. A set of china hit the floor with a crash.
9. This is a great time for walks on the moors.
10. The dune rose up from the water’s edge.

**List 8**

1. The fruit peel was cut into thick slices.
2. The yacht slid around the point into the bay.
3. The two met while playing in the sand.
4. The ink stain dried on the finished page.
5. The walled town was seized without a fight.
6. The lease runs out in sixteen weeks.
7. A tame squirrel makes a nice pet.
8. The horn of the car woke the sleeping man.
9. The heat strongly with firm strokes.
10. The pearl was worn in a thin silver ring.

**List 9**

1. He lay still and hardly moved a muscle.
2. The navy attacked the big task force
3. See the cat glaring at the scared mouse.
4. There are more than two factors here.
5. The hat brim was wide and too floppy.
6. The lawyer tried to lose his case.
7. The grass curled around the fence post.
8. Cut the pie into large pieces.
9. Many try but seldom get rich.
10. Always close the barn door tightly.

**List 10**

1. The bill was paid every third week.
2. The slush lay deep along the street.
3. A wisp of cloud hung in the blue air.
4. A pound of sugar costs more than eggs.
5. The fin was sharp and cut the clear water.
6. The play seems dull and quite stupid.
7. Bail the boat to stop it from sinking.
8. The term ended in late June that year.
9. A tusk is used to make costly gifts.
10. Ten cards were laid in order.

**List 11**

1. Move the pot over the hot fire.
2. An oak tree is strong and gives shade.
3. Cats and dogs hate each other
4. The pipe began to rust from new.
5. Open the crate but don’t break the glass.
6. Add the sum to the product of these three.
7. Thieves who rob friends deserve jail.
8. A ripe taste in cheese improves with age.
9. Act on these orders with great speed.
10. The dog crawled under the high fence.

**List 12**

1. We find joy in the simplest things.
2. The bark of a pine tree is shiny and dark.
3. Leaves turn brown and yellow in the autumn.
4. The flag waves when the wind blew.
5. Split the log with a quick sharp blow.
6. Burn the peat after the logs are finished
7. He ordered apple with ice cream.
8. Clean the carpet on the right hand side only.
9. Hemp is a weed found in parts of the tropics.
10. A bad back kept his score low.

**List 13**

1. The tiny girl took off her hat.
2. Type out three lists of orders.
3. The harder he tried the less he got done.
4. The boss ran the firm with a watchful eye.
5. The cup cracked and spilled its contents.
6. Effort can cleanse the most dirty dishes.
7. The slang word for all alcohol is booze.
8. It caught its hind paw in a rusty trap.
9. The wharf could be seen from the opposite shore.
10. Feel the heat from the weak dying flame.

**List 14**

1. Port is a strong with a smoky taste.
2. A cramp is no small danger on a swim.
3. He said the same phrase thirty times.
4. Pick the bright rose without leaves.
5. Two plus seven is less than ten.
6. The glow deepened in the eyes of the sweet girl.
7. Bring your problems to the wise woman.
8. Write a nice note to a friend you cherish.
9. Clothes and lodgings are nothing to a free man.
10. We frown when events take a bad turn.

**List 15**

1. The cigar burned a hole in the desk top.
2. The young child jumped the rusty gate.
3. Guess the results from the first scores.
4. A sweet pickle tastes good with ham.
5. The just case got the right verdict.
6. These thistles bend in high wind.
7. Pure bred poodles have curls.
8. The tree top waved in a graceful way.
9. The blot on the book was made with green ink.
10. Mud splattered on his white shirt.

**List 16**

1. The sofa cushion was red and light weight.
2. The empty flask stood on the tin tray.
3. The speedy man beat his track record.
4. He broke a new racket that day.
5. The coffee table was too high for the sofa.
6. The urge to write short stories is not rare.
7. The pencils have all been used up.
8. The pirates seized the crew of the lost ship.
9. He tried to replace the coin but failed.
10. She sewed the torn coat quite neatly.

**List 17**

1. A shower of dirt fell from the hot pipes.
2. The jacket hung on the back of the wide chair.
3. At that high level the air is pure.
4. Drop the two when you add the figures.
5. It is hard to buy a filing case now.
6. An abrupt start will not with the prize.
7. Wood is best for making toys and blocks.
8. The office pain was a dull sad tan.
9. He knew the skill of the great young actress.
10. A rag will soak up spilled water.

**List 18**

1. Add the store’s account to the last penny.
2. The steam hissed from the broken valve.
3. The dog almost hurt the small child.
4. There is the sound of dry leaves outside.
5. The sky that morning was clear and bright blue.
6. Scraps of torn paper littered the floor.
7. Sunday is the best part of the week.
8. The doctor cured him with these pills.
9. The new girl was fired today at noon.
10. She felt happy when the ship arrived in the port.

**List 19**

1. She has a way of wearing smart clothes.
2. Acids burns holes in wool cloth.
3. Fairy tales should be fun to write.
4. Eight miles of woodland burned to waste.
5. The third act was dull and tired the players.
6. A young child should not suffer fright.
7. Add the column and put the sum here.
8. We admire and love a good cook.
9. Over there the flood mark is ten inches.
10. He carved a head from the round block of marble.

**List 20**

1. The farmers came in to thresh the oat crop.
2. The fruit of the fig tree is apple shaped.
3. Corn cobs can be used to kindle fire.
4. Where were they when the noise started.
5. The paper box was full of thumb tacks.
6. Sell your gift to a buyer for good gain.
7. The tongs lay beside the ice pail.
8. The petals will fall with the next puff of wind.
9. Bring your best compass to the third class.
10. They could laugh although they were sad.

**List 21 (practice trials)**

1. Even the worst will beat his low score.
2. The attic of the brown house was on fire.
3. A rod is used to catch trout and eels.
4. Float the soap on top of the bath water.
5. A blue crane is a tall wading bird.

# **APPENDIX B: BKB-R Masker sentences (Bench et al., 1979)**

**List 1**

The clown had a funny face.

The car engine’s running.

She cut with her knife.

Children like strawberries.

The house had nine rooms.

They’re buying some bread.

The green tomatoes are small.

He played with his train.

The postman shut the gate.

They’re looking at the clock.

The bag bumps on the ground.

The boy did a handstand.

A cat sits on the bed.

The truck carried fruit.

The rain came down.

The ice cream was pink.

**List 2**

The ladder’s near the door.

They had a lovely day.

The ball went into the net.

The old gloves are dirty.

He cut his finger.

The thin dog was hungry.

The boy knew the game.

The grass grows in summer.

She’s taking her coat.

The police chased the car.

A mouse ran down the hole.

The lady’s making a toy.

Some sticks were under the tree.

The little baby sleeps.

They’re watching the train.

School finished early.

**List 3**

The glass bowl broke.

The dog played with a stick.

The kettle’s quite hot.

The farmer feeds a lamb.

They say some silly things.

The lady wore a coat.

The children are walking home.

He needed his holiday.

The milk came in a bottle.

The man cleaned his shoes.

They ate the lemon jelly.

The boy’s running away.

Father looked at the book.

She drinks from her cup.

The room’s getting cold.

A girl kicked the table.

**List 4**

The wife helped her husband.

The music was very loud.

The old man worries.

A boy ran down the path.

The house had a nice garden.

She spoke to her son.

They’re crossing the street.

Lemons grow on trees.

He found his brother.

Some animals sleep on straw.

The jam jar was full.

They’re kneeling down.

The girl lost her doll.

The cook’s making a cake.

The child drops the toy.

The mud stuck on his shoe**.**