# Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response

Barbara Ohlenforst[1,4], Dorothea Wendt[4,5], Sophia E. Kramer[1], Graham Naylor[6], Adriana A. Zekveld[1,2,3], Thomas Lunner[2,3,4,5]

[1]Section Ear & Hearing, Dept. of Otolaryngology-Head and Neck Surgery, VU University Medical Center and Amsterdam Public Health Research Institute, Amsterdam, The Netherlands;

[2]Department of Behavioral Sciences and Learning, Linköping University, Sweden;

[3]Linnaeus Centre HEAD, The Swedish Institute for Disability Research, Linköping and Örebro Universities, Sweden;

[4]Eriksholm Research Center, Oticon A/S, Denmark;

[5]Department of Electrical Engineering, Technical University of Denmark, Denmark;

[6]MRC/CSO Institute of Hearing Research, Scottish Section, Glasgow, United Kingdom. Part of the University of Nottingham.

Received, 2017

This work was supported by grants from the European Commission (FP7-LISTEN607373) and the Oticon Foundation.

Corresponding author: Barbara Ohlenforst,

Section Ear & Hearing, Dept. of Otolaryngology-Head and Neck Surgery, VU University Medical Center and Amsterdam Public Health Research Institute, Amsterdam, The Netherlands; P.O. Box 7057, 1007 MB, Phone: +31 20 4440952; Fax: +3120 444 2033

Eriksholm Research Centre, Rørtangvej 20, 3070 Snekkersten, Denmark;

E-mail: b.ohlenforst@vumc.nl; Phone: 0045-48298900 Fax: 0045-49223629

## Abstract

Recent studies have shown that activating the noise reduction scheme in hearing aids results in a smaller peak pupil dilation (PPD), indicating reduced listening effort, and 50% and 95% correct sentence recognition with a 4-talker masker. The objective of this study was to measure the effect of the noise reduction scheme (on or off) on PPD and sentence recognition across a wide range of signal-to-noise ratios (SNRs) from +16 dB to -12 dB and two masker types (4-talker and stationary noise). Relatively low PPDs were observed at very low (-12 dB) and very high (+16 dB to +8 dB) SNRs presumably due to 'giving up' and 'easy listening', respectively. The maximum PPD was observed with SNRs at approximately 50% correct sentence recognition. Sentence recognition with both masker types was significantly improved by the noise reduction scheme, which corresponds to the shift in performance from SNR function at approximately 5 dB toward a lower SNR. This intelligibility effect was accompanied by a corresponding effect on the PPD, shifting the peak by approximately 4 dB toward a lower SNR. In addition, with the 4-talker masker, when the noise reduction scheme was active, the PPD was smaller overall than that when the scheme was inactive. We conclude that with the 4-talker masker, noise reduction scheme processing provides a listening effort benefit in addition to any effect associated with improved intelligibility. Thus, the effect of the noise reduction scheme on listening effort incorporates more than can be explained by intelligibility alone, emphasizing the potential importance of measuring listening effort in addition to traditional speech reception measures.


**Keywords:** Hearing impairment, speech recognition, noise reduction scheme, hearing aids, pupil dilation, listening effort, signal-to-noise ratio

## 1. Introduction

Audiological evaluations and research studies investigating hearing aid signal processing have typically focused on changes or benefits in intelligibility but often failed to provide a complete picture of the processes involved in speech recognition (Dillon et al., 1993; Ricketts et al., 2001; Sarampalis et al., 2009). Traditional speech reception measures have been shown to be insensitive to the possible benefits of hearing aid algorithms due to ceiling effects or great variability (Gatehouse et al., 1990). Baer and colleagues (Baer et al., 1993) suggested that the greatest benefit of noise reduction processing in hearing aids may be reduced listening effort rather than enhanced speech intelligibility.

According to the Framework for Understanding Effortful Listening (FUEL) (Pichora-Fuller et al., 2016), listening effort depends on a range of factors, including not only individual factors, such as hearing ability and motivation to continue listening, but also external factors, such as the task demands imposed by the listening situation (Brehm, 1999). Participants may invest less effort in their task performance when the task demands are too high or allocate less cognitive resources under very easy listening conditions (Ohlenforst et al., 2017a). Recently, an increasing number of studies have sought additional methods to gain information about effortful listening as a supplement to traditional audiological measures to assess individual hearing ability (McGarrigle et al., 2014; Ohlenforst et al., 2017b; Pals et al., 2013; Wu et al., 2016). These methods include subjective assessments, such as self-reports and questionnaires (McAuliffe et al., 2012; Panico et al., 2009; Picou et al., 2011); behavioral measures, such as dual-task paradigms or reaction time measures (Fraser et al., 2010; Houben et al., 2013; Tun et al., 2009); and physiological measures, such as the pupil response and functional magnetic resonance imaging (fMRI) or EEG measures (Kuchinsky et al., 2013; Obleser et al., 2012; Petersen et al., 2015). Importantly, the listening conditions may affect listening effort even when speech intelligibility is not affected, such as when speech intelligibility is at a ceiling and hence constitutes an insensitive outcome measure (Koelewijn et al., 2014; Wendt et al., 2017). For example, Wendt et al., (2017) showed that activating the noise reduction scheme at ceiling performance reduced listening effort, but speech in noise performance was unaffected. Therefore, simultaneously assessing listening effort and speech performance may uncover challenges or changes in processing speech that may not be evident with traditional measures.

Numerous studies across different research areas have shown that pupil dilation increases as the processing load imposed by the task demands increases (Beatty, 1982; Engelhardt et al., 2010; Granholm et al., 1996; Kahneman, 1973; Van Der Meer et al., 2010). Pupillometry has repeatedly been verified as a valid measure for quantifying the effort required for speech recognition with background noise (Koelewijn et al., 2012; Koelewijn et al., 2014; Kramer et al., 1997; Ohlenforst et al., 2017a; Ohlenforst et al., 2017b; Wendt et al., 2017; Zekveld et al., 2011). For instance, the SNR (ranging from -20 dB to +16 dB) and masker type (stationary and 1-talker masker) have been shown to affect pupil dilation during listening (Ohlenforst et al., 2017a). Recent studies indicate that effort is not necessarily monotonically related to the task demands. The changes in effort follow an inverse U-shaped function, indicating that listeners may exert less effort due to ‘giving up’ under very difficult conditions and ‘taking it easy’ when listening at high SNRs (Ohlenforst et al., 2017a; Wu et al., 2016; Zekveld et al., 2014). Ohlenforst et al. (Ohlenforst et al., 2017a) investigated the peak pupil dilation (PPD) across a range of SNRs in hearing-impaired and normal-hearing listeners. These authors showed that the PPD, which is an indication of the cognitive processing load, was affected by an interaction between the masker type and hearing status of the individual. In the presence of a stationary noise masker, the hearing-impaired listeners showed relatively large PPDs across a wide range of SNRs, while the normal-hearing listeners showed a maximum PPD across a relatively narrow range of low (challenging) SNRs (Ohlenforst et al., 2017a). With a single-talker masker, the maximum PPD was in the mid-range of SNRs, while relatively smaller PPDs were observed at low and high SNRs in both groups of listeners. Interestingly, recent findings across a variety of studies in the field of listening effort suggest that the allocation of mental resources needed during listening to reach speech understanding in daily life listening situations may differ between normal-hearing and hearing-impaired listeners (Ohlenforst et al., 2017a; Ohlenforst et al., 2017b; Zekveld et al., 2011).

Hearing aids are designed to improve the audibility of sounds and facilitate the intelligibility of speech in both quiet and noisy environments. These improvements may be accompanied by reduced listening effort. The advanced signal processing in hearing aids includes a digital noise reduction scheme, which aims to reduce the level of interfering background noise by improving the SNR. Recent studies indicate that the noise reduction scheme improves the recall of words presented in a competing multi-talker background (Lunner et al., 2016; Ng et al., 2015; Ng et al., 2013). The researchers concluded that the noise reduction scheme may reduce the adverse effect of noise on memory and thereby facilitate the segregation of the target from the multi-talker masker signal. This enhanced memory of the target words was interpreted to represent reduced listening effort (Lunner et al., 2016; Ng et al., 2015; Ng et al., 2013). Moreover, Wendt et al. (2017) presented speech in a 4-talker babble masker at two SNRs (SNR50 and SNR95) corresponding to the individual 50% or 95% sentence recognition level. These authors assessed the effect of the noise reduction scheme by applying a combination of a digital noise reduction scheme and directional microphones. When the scheme was activated in the hearing aid, the speech recognition performance at SNR50 was significantly improved and accompanied by significantly smaller PPDs. Interestingly, activating the noise reduction scheme did not affect the near-ceiling speech recognition performance at SNR95. Nevertheless, significantly smaller PPDs were observed, indicating that the noise reduction scheme had a beneficial effect on listening effort. Thus, measuring listening effort by assessing PPD could provide a sensitive outcome measure of hearing aid benefit even at high performance level traditional methods of audiological assessment are not sufficiently sensitive.

The studies described above (Ng et al., 2015; Ng et al., 2013; Wendt et al., 2017) indicate that effort can be reduced with modern hearing aid signal processing. However, knowledge regarding the benefit of noise reduction processing on listening effort remains very limited as only a few listening conditions were tested in these studies. In contrast, the effect of noise reduction processing on intelligibility has been studied by several groups of researchers. In these studies, the inconsistency in the diverse noise reduction processing schemes studied renders generalization problematic, especially as processing schemes become increasingly sophisticatedion over time. Some research studies have indicated that the application of noise reduction processing may not always be beneficial for speech intelligibility (Bentler et al., 2008; Nordrum et al., 2006). Such negative effects suggest that while the background noise may be removed, the target speech might also be degraded. Stronger or more aggressive signal processing may cause more signal enhancement but could simultaneously introduce more degradation (Loizou et al., 2011). For example, in a recent study, the effect of noise reduction processing on sentence recognition was tested in the presence of a cafeteria background masker (Neher et al., 2013). Simulated hearing aid processing including coherence-based noise reduction was presented via headphones to hearing-impaired listeners. The algorithm was designed to suppress the reverberant signal components and diffuse the background noise at mid to high frequencies but did not include directionality. The results showed that sentence recognition was unaffected by the moderate noise reduction processing, but the strong noise reduction processing reduced speech recognition by approximately 5%. The effect was replicated in a follow up study in which the same acoustic test conditions were used in a group of habitual hearing aid users (Neher, 2014). Compared to the moderate or no noise reduction processing, the strong noise reduction processing reduced speech recognition at -4 dB and 0 dB SNR.

How hearing-impaired listeners invest listening effort across a broader range of listening situations and how effortful listening relates to performance measures remain unclear. The current study aimed to examine how a noise reduction scheme influences sentence recognition and listening effort. The applied noise reduction scheme preserves speech and reduces noise in complex environments by a fast-acting combination of a beam-former (Kjems et al., 2012) and a single-channel Wiener post-filter (Jensen et al., 2015) to attenuate interfering sounds. Any effect of the noise reduction processing on intelligibility likely affects the PPD in a corresponding direction as the intelligibility of speech has a strong and reliable effect on the PPD (Koelewijn et al., 2014; Ohlenforst et al., 2017a; Zekveld et al., 2014). However, in addition to this intelligibility effect, the noise reduction processing may have additional effects on the PPD, as suggested by recent studies investigating listening effort that demonstrated that hearing aid processing has a beneficial effect on listening effort due to reduced background noise and reduced cognitive effort during speech processing (Picou et al., 2013; Sarampalis et al., 2009; Wendt et al., 2017). Demonstrating the effect of noise reduction processing on listening effort combined with simultaneous knowledge regarding speech in noise performance could further substantiate the value of measuring effort as an extra dimension in addition to traditional speech reception measures.

Recent research found better SRTs in speech recognition in the presence of a single-talker masker than those in the presence of a stationary noise masker (Koelewijn et al., 2012). The envelope modulations of the multi-talker masker might allow the participants to listen in the energy dips in the spectral-temporal domain and glimpse parts of the target sentence (Festen et al., 1990; Francart et al., 2011; Koelewijn et al., 2012; Koelewijn et al., 2014; Rosen et al., 2013). Based on the characteristics of the masker types and recent findings, we hypothesize that speech recognition performance is better with the 4-talker masker than that with the stationary noise masker (Koelewijn et al., 2012; Koelewijn et al., 2014). However, recent studies suggest that the intelligibility of speech masked by additional interfering speech information may require more mental effort than that with an energetic mask (Larsby et al., 2008). Informational masking, including lexical interference or the competition for neural resources, may cause higher listening effort (Beatty, 1982; Koelewijn et al., 2012; Koelewijn et al., 2014; Scott et al., 2004; Scott et al., 2009). We hypothesized that the better speech recognition with the 4-taker masker compared to that with the stationary noise masker could be accompanied by larger PPDs. We hypothesized that sentence recognition is improved and listening effort is reduced with SNRs corresponding to approximately 50% correct or better performance with the active noise reduction compared to the inactive noise reduction scheme. This hypothesis is motivated by two arguments. First, in a previous study conducted by Wendt and colleagues (2017), SRT targeting 50% correct performance was significantly improved by the active noise reduction scheme compared to that with the inactive noise reduction scheme setting. Second, the segregation between the target and masker signal at very low SNRs might be more difficult for the algorithm, which might have an impact on the SNR improvement provided by the algorithm.

## 2. Materials and methods

### 2.1 Participants

Twenty-five experienced hearing aid users were recruited from the Eriksholm Research Centre in Denmark. On average, the participants had used hearing aids for 7.7 years (SD=3.1 years). The participants were aged between 46 and 77 years (mean age 64.3 years, SD=9.4) and native Danish speakers. The audiometric inclusion criterion for the participants was symmetrical, with mild to moderate sensorineural hearing thresholds. The average pure tone hearing thresholds ranged between 35 dB and 60 dB HL (see Figure 1), and air-bone gaps less than 10 dB between 500 Hz and 4000 Hz were required in both ears. All the participants had normal or corrected-to-normal vision and no history of neurological diseases, dyslexia or diabetes mellitus. All the participants provided written informed consent, and the study was approved by the local regional ethics committee (De Videnskabsetiske Komiteer for Region Hovedstaden).

### 2.2 Auditory stimuli

Everyday Danish sentences from the Hearing in Noise Sentence Test (HINT) (Nielsen et al., 2009) were presented in a spatial setup with five loudspeakers in a sound proof measurement booth as shown in Figure 2. The target sentences were spoken by a male and presented from a loudspeaker located at 0 degree azimuth. All the sentences contained five words, 8-9 syllables were included in each sentence, and the single words did not contain more than four syllables (Nielsen et al., 2009). The following is an example of a presented sentence: "Filmen er rigtig godt lavet" (translation: "the movie was well made"). The sentence duration was on average 1.4 seconds. The listeners were presented with a training list of 20 sentences for each masker type, followed by eight lists of 25 sentences for every SNR. To cover the large number of testing conditions, the sentence material was re-used across four experimental visits. Recent research assessed the possible learning effects due to repeated exposure to HINT sentences across three experimental visits with an interval of three weeks between visits. The results showed that the memory effects of the sentence material are not significant with limited exposure when the sentences were only presented once during each visit (Simonsen et al., 2016). The experimental visits in the current study were separated by at least three weeks, and identical sentence material was not repeated within each visit to prevent learning effects of the speech material. The speech recognition performance was measured in the presence of a stationary noise or a 4-talker masker background. The 4-talker masker was made from four single-talker maskers, including two different male voices and two different female voices. Each separate talker read a text passage from a newspaper, and one single talker was presented from one loudspeaker, each positioned at -/+ 90 and -/+ 150 degree azimuth (Wendt et al., 2017). We balanced the distribution of the talkers across loudspeakers for each SNR by switching the order of the talkers. There were never two talkers of the same gender next to each other or on the opposite position of each loudspeaker. In each trial, the masker started 3 seconds prior to the presentation of the sentence and ended 3 seconds after the sentence offset. The participants repeated the sentence aloud once the masker stopped. The same presentation procedure was applied for both masker types. The long-term average frequency spectrum of both masker types was identical to the spectrum of the target speech signal, and the masker was always presented at 70 dB SPL. The masker levels were kept constant to ensure that the noise would not become too loud at low SNRs. Changing the noise levels might also allow the listeners to estimate the upcoming task difficulty. The same SNR range was chosen for both masker types. We included a large range of positive SNRs as previous findings suggested that typical, ecologically sound environments for hearing-impaired listeners occur at SNRs of approximately +5 dB or better (Festen et al., 1990; Ohlenforst et al., 2017a; Smeds et al., 2015; Wu et al., 2014; Zekveld et al., 2014). Speech masked with a stationary masker and 4-talker masker was presented at eight SNRs between -12 dB and +16 dB and distributed in steps of 4 dB. Per the masker type, 25 sentences were presented for each SNR.

### 2.3 Noise reduction scheme

All the participants wore identical hearing aid models during the sentence recognition test and examined in the same two different settings. In one setting, the noise reduction scheme was turned off, but the hearing aid provided audibility based on each individual's hearing threshold via the Voice Aligned Compression (VAC) rationale (Le Goff, 2015). The VAC amplification rationale is based on a wide dynamic range compression scheme with compression knee points between 20 and 50 dB SPL depending on the frequency range and the individuals' hearing thresholds. The hearing aid was set to mimic the natural acoustic effect of the pinna; thus, the microphone setting was close to omnidirectional, and no actual noise reduction processing was applied. The other setting involved activating the noise reduction scheme. In this setting, a fast-acting combination of a minimum variance distortion-less response (MVDR) beam-former (Kjems et al., 2012) and a single-channel

Wiener post-filter (Jensen et al., 2015) was applied before the VAC. In the algorithm, spatial filtering and Wiener filtering were applied to attenuate interfering sounds originating behind the listener.

The output SNR method suggested by Naylor and Johannesen (2009) was used to directly measure the SNR effect of the complete noise reduction scheme. The hearing aid was placed in a sound field and exposed to running speech plus noise mixtures in SNRs ranging from -10 dB SNR to +20 dB in steps of 5 dB for the two different noise types (speech-weighted unmodulated noise and multi-talker babble noise). The output SNR method was applied to NR on and off. In the range of -10 dB SNR to +10 dB SNR, the listeners experience an articulation-index (AI) weighted SNR improvement ranging from 4.5 dB to 5.2 dB for NR on compared to that for NR off for the speech-weighted noise and an AI weighted SNR improvement ranging from 4.2 dB to 4.8 dB for the multi-talker babble noise. For SNRs above +10 dB, the SNR improvement gradually declined to a few dB because the noise estimates in the noise reduction algorithm decline at high SNRs, and thus, the noise reduction algorithm becomes less effective.

### 2.4 Pupillometry

During the experiment, the pupil location and pupil size were recorded using an eye tracking system by SensoMotoric Instruments (SMI, Berlin, Germany, 2D Video-Oculography, version 4), which applies infrared video tracking to measure the pupil diameter. The eye tracking system had a sampling frequency of 120 Hz and a spatial resolution of 0.03 mm. The pupil location and pupil size were recorded by the eye tracker and stored on a connected computer with time stamps corresponding to the start of each trial, including the masker onset, the sentence onset and the offset for the post-masker. The experimenter monitored the pupil recordings and applied corrective actions. In the case that a participant moved his/her head or upper body or the real-time pupil recordings were missing data regarding the pupil diameter, corrective actions, such as adjusting the participants' position, the distance to the eye tracker, or light, were applied.

## 2.5 Procedures

In total, 17 adults from the Eriksholm pool of participants with recent pure tone audiogram data and recently made ear impressions (less than 6-month-old) were required to participate in four experimental visits. We recruited 8 additional participants who required an additional recruitment visit (total of five visits) to measure the pure tone audiogram and take ear impressions. In total, four experimental visits, including two visits per masker type, were required for each participant. The visits were distributed across approximately four months during the fall of 2016 with intervals of at least three weeks between each visit to avoid learning effects of the sentence material as the material was repeatedly used (Simonsen et al., 2016). During the four experimental sessions, each participant sat on a fixed chair in front of the eye tracking system in a sound proof booth. The experimenter observed the real-time recording of the pupil response from the eye tracking system to evaluate the pupil recording quality. The height of the chair and the distance to the eye tracker (55 cm +/- 5 cm approximately) were adjusted individually until a stable, continuous pupil response was measured. The illumination in the measurement booth was fixed during the experiment to an average of 84.3 lux (SD=3.56 lux). The stationary noise and 4-talker masker were presented at eight identical SNRs between -12 dB and +16 dB distributed in steps of 4 dB. During each visit, only 1 of the 2 masker types was presented in two blocks of four randomized SNRs. In one block, the noise reduction scheme was turned on, and in the other block, the noise reduction scheme was turned off. During each visit, each noise reduction scheme setting (on or off) was tested at four SNR levels. We balanced the SNR levels for each visit, including two difficult and two easier SNRs (e.g., -12, -4, +4 and +12 dB SNR or -8, 0, +8 and +16 dB SNR). We balanced the setting of the noise reduction scheme and the presented masker types across visits and blocks. Each participant's visit started with a practice session in which the noise reduction scheme setting was the same as that in the starting block, and 20 sentences at an SNR of +4 dB were tested. The practice session ensured that the participants were confident with the experimental procedures as it may not be intuitive to inhibit movements and blinking during the sentence presentation. A sentence was scored as correct if all the words were correctly repeated.

## 2.6 Pupil data selection and cleaning

Pupil diameter values more than 2 standard deviations from the mean pupil diameter in a given trial were defined as blinks. Pupil traces with more than 25% of blinks between the start of baseline (final second pre-noise before the sentence onset) and the end of the post-masker were excluded from data analysis. For pupil traces with less than 25% of blinks, the blinks were interpolated linearly starting with 5 samples before and 7 samples after each blink (Siegle et al., 2008). The pupil response within each selected and de-blinked trace was smoothed by a 9-point moving average filter. The reference of the task evoked pupil dilation was the baseline, which corresponded to the average pupil diameter recorded during the final second of the three second presentation of the masker before the target speech onset. The PPD was calculated as the maximum pupil dilation between the onset of the sentence and the offset of the noise relative to the baseline pupil diameter for every trace (one pupil trace was recorded per sentence). For each participant and each condition, all the included de-blinked and smoothed traces (≤25) were time-aligned and averaged. For each SNR condition, at least 18 valid pupil traces (n=25 traces in total) with less than 25% of blinks were required per participant to consider the pupil data for the statistical analysis. Eighteen participants had the required number of valid pupil traces for each of the 32 testing conditions. Six participants had less than 18 valid pupil traces in at least one of the testing conditions, and two participants had missing data (<18 valid pupil traces) in at 3 test conditions. We calculated the average pupil trace across all the valid pupil traces per SNR condition and subject. The mean PPD was calculated based on the averaged pupil trace and thus provided the data for the statistical analysis per SNR and participant.

## 2.7 Statistical analyses

Pupil data selection and cleaning were applied to the pupil data from 24 participants (50% female). One participant was excluded due to unexpected attention problems. We measured 800 pupil traces during the experimental sessions (excluding the practice traces) per participant, and on average, 38 (SD=12.92) pupil traces were excluded per person. The corresponding sentence recognition scores for all 800 measured traces were included in the statistical analysis.

We applied linear mixed models (LMM) to analyze the data as LMMs tolerate missing values, while repeated measures ANOVA tests only use complete cases contrary to multilevel analyses. Moreover, mixed-effects models are more flexible in processing the multilevel structure of the data (i.e., the 8 different SNRs and 2 different hearing aid settings). We averaged over 25 sentences to obtain one ‘observation’ under each hearing aid setting and listening condition (SNR and masker type), which is commonly performed in pupillometry research (Koelewijn et al., 2012; Koelewijn et al., 2014; Ohlenforst et al., 2017a; Zekveld et al., 2011). A linear mixed-effects model was built in R-studio using the packages lme4 (Bates et al., 2014) and lmerTest (Kuznetsova et al., 2016). The function lmer was applied to fit the LMM to the data. First, we applied a 3-way LMM ANOVA to statistically compare the fixed effects of the masker types, SNR and noise reduction setting on the PPD and the sentence recognition performance separately to verify the hypothesis that the masker type and SNR range have an impact on speech recognition performance and the corresponding listening effort. The probability level of each LMM ANOVA was $p < 0.05$. We did not observe a significant 3-way interaction effect on the PPD, but we did observe a significant interaction between the SNR and noise reduction scheme setting. The model was collapsed across masker type, and an additional 2-way LMM ANOVA was applied to assess the effect of the SNR and noise reduction scheme setting and the corresponding interaction effect on the PPD.

The three-way interaction among the masker type, SNR and noise reduction scheme setting on sentence recognition performance was significant. We created two additional separate LMM ANOVAs to test the effect of the SNR of each masker type independently (stationary noise and 4-talker masker) on the percent-correct sentence recognition. In these models, the averaged percentage of correct sentence recognition scores for each SNR was treated as a dependent measure, and the participants were treated as a repeated measure, i.e., random effects. The fixed effects in each separate LMM ANOVA included the categorical variable SNR, the categorical variable noise reduction scheme setting and the interaction between the SNR and noise reduction scheme setting. We included the random effect of the SNR and noise reduction scheme as a random slope of SNR to allow each participant to have their own mean PPD size and effect of SNR or noise reduction scheme on PPD with both factors nested within participants. The *phia* package, including the *testInteractions* functions, was used to apply a post hoc interaction analysis. Pairwise comparisons of the noise reduction scheme setting (on or off) at each SNR level were conducted. The pairwise post-hocpost hoc analysis was separately applied to both outcome measures (PPD and sentence recognition performance), and a p-value correction using the Holm method was applied to correct for the multiple comparisons.

# 3. Results

## 3.1 Sentence recognition data

The results are displayed in Figures 3 and 4. Figure 3 shows the sentence recognition scores across the range of stationary noise masker SNRs with the noise reduction scheme on (solid, gray curve) or off (dashed, gray curve). The sentence recognition scores with the 4-talker masker are shown in Figure 4 with the noise reduction scheme on (solid, gray curve) or off (dashed, gray curve). The error bars represent the standard error of the mean.

The 3-way LMM ANOVA revealed significant main effects of SNR ($F_{[7,713]}$=1382.5, $p$<0.001), noise reduction scheme ($F_{[1,713]}$=524.4, $p$<0.001), and masker type ($F_{[1,713]}$=72.9, $p$<0.001), indicating that sentence recognition is affected by differences in the listening conditions (SNR and masker type) and the noise reduction processing algorithm. Furthermore, we found significant interactions between the SNR and noise reduction scheme ($F_{[7,713]}$=93.7, $p$<0.001), between the SNR and masker type ($F_{[7,713]}$=5.73, $p$<0.001) and among the SNR, noise reduction scheme and masker type ($F_{[7,713]}$=2.82, $p$<0.01). The interaction between the masker type and noise reduction scheme was not significant. The interaction effects of among the masker type, noise reduction scheme and SNR are larger in the mid-range of SNRs, while at relatively low and high SNRs, floor or ceiling effects of sentence recognition were observed.

Regarding the stationary noise masker, at relatively high SNRs between +16 dB and +8 dB, the participants achieved 100% sentence recognition independent of the setting of the noise reduction scheme. As the SNR decreased (+8 dB to -8 dB), sentence recognition rapidly decreased until the participants were unable to perform correct sentence recall at -12 dB SNR when the noise reduction scheme was turned off. At -12 dB SNR, the participants could correctly recognize approximately 12% when the noise reduction scheme was turned on. Overall, the sentence recognition curve at the level of 50% correct speech recognition was shifted by approximately 5.5 dB (see Figure 3) toward lower SNRs when the noise reduction scheme was turned on compared to that when it was turned off. The LMM ANOVA revealed significant main effects of SNR ($F_{[7,345]}$=846.2, $p$<0.001) and noise reduction scheme ($F_{[1,345]}$=332.5, $p$<0.001) and a significant interaction between the SNR and noise reduction scheme ($F_{[7,345]}$=68.8, $p$<0.001). We performed pairwise post hoc comparisons between the two noise reduction scheme settings (on or off) at each SNR level. Post hoc analysis revealed significant differences between the noise reduction scheme settings at -12 dB, -8 dB, -4 dB and 0 dB SNR (p<0.01, as indicated by gray diamonds in Figure 3).

Regarding the 4-talker masker, at SNRs between +16 dB and +8 dB, nearly 100% sentence recognition was achieved regardless of the noise reduction setting. The overall performance curve was shifted by approximately 5.1 dB toward the lower SNRs when the noise reduction scheme was turned on compared to that when it was turned off. By applying an LMM ANOVA, we found significant main effects of SNR ($F_{[7,345]}$=617.3, $p$<0.001) and noise reduction scheme ($F_{[1,345]}$=223.8, $p$<0.001) and a significant interaction between the SNR and noise reduction scheme ($F_{[7,345]}$=36.2, $p$<0.001). We performed pairwise post hoc comparisons between the two noise reduction scheme settings (on or off) at each SNR level. Significant differences were observed in the sentence recognition performance between the noise reduction scheme settings at -8 dB, -4 dB, 0 dB and +4 dB SNR (p<0.01, as indicated by gray diamonds in Figure 4).

Arcsine transformation prior to analyzing proportion data, such as the percent of correct responses, is known to stabilize the variance and normalize proportional data (Studebaker, 1985). We applied the arcsine transformation to the speech scores and performed the statistical analysis described in section 2.7 by using LMM ANOVAs of the speech data. The results revealed small differences in the F-and p-values compared to those obtained by analyzing the percentage scores. We chose to apply the statistical analysis of the speech data because the prior arcsine transformation did not change the results, and arcsine units are difficult to interpret as they fall into a numeric range that has little intuitive relationship to the proportionate performances.

## 3.2 Pupil data

Figure 3 shows the PPD data under the stationary noise masker conditions, and Figure 4 shows the PPD data under the 4-talker masker conditions across SNRs. The 3-way LMM ANOVA revealed the significant main effects of the SNR ($F_{[7,699.1]}$=26.82, $p$<0.001), noise reduction scheme ($F_{[1,699.1]}$=25.34, $p$<0.001), and masker type ($F_{[1,699.1]}$=21.37, $p$<0.01), and a significant interaction was observed between the SNR and noise reduction scheme ($F_{[7,699.1]}$=9.97, $p$<0.01). No significant interaction was observed between the masker type and SNR or masker type and noise reduction scheme. The interaction effect between the SNR and noise reduction scheme suggests that the SNR-dependency of the PPD differs when the noise reduction scheme is on from that when the scheme is off. We did not test two separate models for each masker type per sentence recognition performance. In an additional 2-way LMM ANOVA that collapsed across the level of masker type, the noise reduction scheme setting and masker type were not significant, which is similar to the interaction with the SNR. The 2-way LMM ANOVA revealed a significant main effect of noise reduction scheme setting ($F_{[1,715.05]}$=25.08, $p$<0.001), a significant main effect of SNR ($F_{[7,715.07]}$=25.94, $p$<0.001) and a significant interaction effect between the noise reduction scheme setting and SNR ($F_{[7,715.05]}$=9.72, $p$<0.001) on the PPD. Pairwise post hoc comparisons of the two noise reduction scheme settings (on or off) were applied at each SNR level. Significant differences were observed between the noise reduction scheme settings in the PPD measured at -8 dB, -4 dB, 0 dB and +4 dB SNR.

Figure 3 shows the averaged PPD data across SNRs for the stationary noise masker when the noise reduction scheme was active (black, solid line) and when the noise reduction scheme was inactive (black, dashed line). The PPD plateaued with relatively high SNRs between +16 and +8 dB where high performance was reached independently of the noise reduction scheme setting. When the noise reduction scheme was turned off, as the SNR further decreased, a steady increase in PPD was observed until a maximum PPD was reached at -4 dB SNR. The corresponding sentence recognition was approximately 38% correct. The maximum PPD was shifted by 4 dB toward lower SNRs when the noise reduction scheme was turned on, and this maximum corresponded to an approximately 52% correct sentence recognition. At the lowest SNR of -12 dB, relatively lower PPDs were observed under both noise reduction scheme settings.

Figure 4 shows the PPD data across SNRs with the noise reduction scheme on (black, solid curve) or off (black, dashed curve) under the 4-talker masker condition. The PPD measured with high SNRs between +16 dB and +8 dB was overall consistent but larger when the noise reduction scheme was off compared than that when it was on. Further decreases in the SNRs resulted in continuous increases in the PPD until the maximum PPD was reached between -4 dB and 0 dB SNR when the noise reduction scheme was off and between -8 and -4 dB SNR when the noise reduction scheme was on. The range of the maximum PPD was shifted by approximately 4 dB toward the lower SNRs when the noise reduction scheme was turned on compared to that when it was turned off.

## 3.3 Summary of the results

The preceding statistical analyses support the following summary of the results: The effect of the noise reduction scheme applied in this study on sentence recognition was to shift the performance function across SNRs by approximately 5.5 dB for the stationary masker and approximately 5.1 dB for the 4-talker masker toward the lower SNRs. For both masker types, the effect of the noise reduction scheme on listening effort (as measured by the PPD) was to shift the peak of the PPD function across SNRs by approximately 4 dB toward the lower SNR. In addition, in the case of the 4-talker masker, the noise reduction scheme lowered the average PPD by approximately 35% compared to the inactive noise reduction scheme.

## 4. Discussion

In the present study, the effect of a noise reduction scheme on sentence recognition and PPD was examined across a range of SNRs with two masker types. For both masker types, the noise reduction scheme had a large beneficial effect on sentence recognition, which was accompanied by a corresponding effect on listening effort, as indicated by the PPD.

### 4.1 Relationship among noise reduction scheme, SNR and speech recognition

For the stationary and 4-talker maskers, the sentence recognition performance was significantly improved when the noise reduction scheme was active compared to that when it was inactive. The results showed improved sentence recognition not only at performance levels of approximately 50% and higher but also at lower sentence recognition performances. Notably, sentence recognition was mainly improved across a large range of negative SNRs between 0 dB and -12 dB. The findings of the present study confirm and extend the previously shown benefits of a noise reduction scheme on sentence recognition with an approximately 50% successful performance rate (Wendt et al., 2017) at higher and lower performance levels. Additionally, the present study confirmed that the currently tested noise reduction scheme can significantly improve speech intelligibility in very challenging sound environments. Hence, this finding might allow hearing-impaired listeners to participate in communication situations that might otherwise be impossibly challenging.

## 4.2 Relationship among noise reduction scheme, SNR and PPD

In line with recent research (Ohlenforst et al., 2017a; Zekveld et al., 2014), the present results confirm that the changes in speech recognition are accompanied by changes in PPD. We found the maximum PPD with SNRs producing approximately 50% correct sentence recognition and relatively smaller PPDs at very low and very high SNRs. The indication that listening effort follows an inverted U-shape across a range of SNRs also supports the findings reported in a recent study (Wu et al., 2016) in which dual-task paradigms were applied to assess listening effort across a wide range of SNRs. Wu et al. (2016) found that second-task performance (reaction time) was the worst (i.e., longest) at SNRs for 30-50% speech recognition and better at both lower and higher SNRs. The change in the PPD function at positive SNRs when the percent-correct sentence recognition is saturated might be affected by the type of speech material used in the sentence recognition test. The transfer function of the speech intelligibility index is modifiable depending on the tested sentence material, and more difficult speech material can change the transfer function. Thus, the transfer function at positive SNRs might already be saturated for speech intelligibility index values that are not at the level of saturation. However, we designed this experiment to intentionally reach a ceiling in performance, although with very positive SNRs, a ceiling effect is achieved regardless of the presented speech material.

The statistical analysis revealed that the level of the SNR and the noise reduction scheme setting significantly affected the PPD. The impact of the masker type on the PPD was rather small, which might contrast with previous studies reporting that listening effort required for speech recognition is altered by the type of background masker (e.g., Koelewijn et al., 2012; Koelewijn et al., 2014).

Koelewijn and colleagues reported significantly larger pupil dilation responses for masker types containing speech information, and the increase in effort was mainly explained by the semantic inference with the target. However, Koelewijn and colleagues examined the impact of masker types on the PPD at similar intelligibility levels corresponding to 50% correct speech recognition. Therefore, comparisons between the PPDs of the different masker types were drawn at varying SNRs. Our data indicate that the PPDs are strongly affected by the SNRs, which is in line with the results of previous studies (Zekveld and Kramer, 2014; Ohlenforst et al., 2017). Hence, the differentiation between the effect of the SNR and masker type is not possible based on these aforementioned studies by Koelewijn and colleagues. Our results suggest that when examining the PPD across a range of intelligibility varying between 0 to 100%, a non-linear change in the PPD, with maximum PPDs occurring at approximately 50% recognition, could be observed independently of the masker type. Furthermore, the impact of the masker type might be less pronounced when testing fixed SNRs, which is in line with the results of previous work (see Wendt et al., 2018 in press).

One strength of the present study is the replication of previous findings, demonstrating the beneficial effect of a noise reduction scheme in hearing aids on sentence recognition and the PPD (Wendt et al., 2017). There were several factors that were kept constant between the setup of the recent study by Wendt and colleagues (2017) and the current study. In both studies, the same noise reduction scheme was tested during a sentence recognition task with identical stimulus material (HINT sentences in a 4-talker masker). Additionally, the large number of listeners (n=17) that participated in the study by Wendt et al., (2017) were used in the present study. Both studies contribute to the field of hearing research and listening effort by providing new valuable knowledge showing the possible benefits of a noise reduction scheme for hearing-impaired listeners wearing hearing aids.

## 5. Conclusion

The present study demonstrates that a noise reduction scheme in commercial hearing aids can reduce the effort required during speech recognition in stationary noise and a 4-talker masker. With both maskers, the noise reduction processing resulted in a shift in the performance (sentence recognition) function toward lower (more challenging) SNRs, and a corresponding shift in the PPD function was observed. For the 4-talker masker, in addition to the speech recognition-related reduction in the PPD, a main effect of noise reduction processing on the PPD was observed, indicating that the cognitive processing load and some aspects of listening effort may be reduced independent of the SNR. These results also confirm previous findings by showing that for hearing-impaired listeners using hearing aids during speech recognition, listening effort changes in a non-monotonic way as a function of the SNR. This knowledge is essential for future research in the field of listening effort and the hearing aid industry for improving the development of better hearing aid algorithms.

## 6. Acknowledgments

The authors would like to thank Per Bruun Brockhoff from the Institute for Mathematics and Computer Science at the Technical University of Denmark (DTU Compute) and Birgit Lissenberg-Witte from the department of Epidemiology and Biostatistics at the VU in Amsterdam for their support and advice with the statistical analyses. We would also like to thank Renskje Hietkamp for her support with the participant recruitment and data collection and Nicolas Le Goff and Jesper Jensen for the fruitful discussions about the tested hearing aid technology. We would like to thank Jacob Aderhold for technical support and advice with the hearing aids used in this study and Yang Wang for fruitful teamwork and discussions throughout the study. Finally, we wish to thank all the participants, the

European Commission (grant FP7-LISTEN607373) and the Oticon Foundation for supporting this study. Co-author GN was supported by the UK Medical Research Council (grant U135097131) and a grant from the Chief Scientist Office.

## Appendix

Table 1: Beta estimates of the sentence recognition performance scores and PPD at each SNR level show the mean differences between the inactive and active noise reduction scheme setting. The SNR levels are compared to the lowest SNR at -12 dB.

| SNRs [dB] compared to the reference SNR of -12 dB | -8 | -4 | 0 | +4 | +8 | +12 | +16 |
|---|---|---|---|---|---|---|---|
| Beta estimates of performance with the stationary noise masker | -33.68 | -40.57 | -5.68 | 7.92 | 11.03 | 11.67 | 11.40 |
| Beta estimates of performance with the 4-talker masker | -26.57 | -46.05 | -20.73 | -4.07 | 4.57 | 4.40 | 5.97 |
| Beta estimates of the PPD collapsed across stationary noise masker and 4-talker masker | -0.03 | 0.04 | 0.08 | 0.05 | 0.03 | 0.01 | 0.03 |

## References


Baer, T., Moore, B.C.J., Gatehouse, S. 1993. Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times. . Journal of Rehabilitation Research and Development 30, 49-72.

Bates, D., Mächler, M., Bolker, B., Walker, S. 2014. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Beatty, J. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological bulletin 91, 276.

Bentler, R., Wu, Y.-H., Kettel, J., Hurtig, R. 2008. Digital noise reduction: Outcomes from laboratory and field studies. . International journal of audiology 47, 447-460.

Borch Petersen, E., Wöstmann, M., Obleser, J., Lunner, T. 2017. Neural tracking of attended versus ignored speech is differentially affected by hearing loss. . Journal of Neuralphysiology 117, 18-27.

Brehm, J.W. 1999. The intensity of emotion. Personality and Social Psychology Review 3, 2-22.

Dillon, H., Lovegrove, R. 1993. Single microphone noise reduction systems for hearing aids: A review and an evaluation. Boston: Allyn and Bacon.

Engelhardt, P.E., Ferreira, F., Patsenko, E.G. 2010. Pupillometry reveals processing load during spoken language comprehension. The Quarterly Journal of Experimental Psychology 63, 639-645.

Festen, J.M., Plomp, R. 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. The Journal of the Acoustical Society of America 88, 1725-1736.

Francart, T., Van Wieringen, A., Wouters, J. 2011. Comparison of fluctuating maskers for speech recognition tests. International journal of audiology 50, 2-13.

Fraser, S., Gagné, J.-P., Alepins, M., Dubois, P. 2010. Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. journal of Speech, Language, and Hearing Research 53, 18-33.

Gatehouse, S., Gordon, J. 1990. Response times to speech stimuli as measures of benefit from amplification. . British Journal of Audiology. 24, 63-68.

Granholm, E., Asarnow, R.F., Sarkin, A.J., Dykes, K.L. 1996. Pupillary responses index cognitive resource limitations. Psychophysiology 33, 457-461.

Houben, R., v. Doorn-Bierman, M., Dreschler, W.A. 2013. Using response time to speech as a measure for listening effort. . International journal of audiology 52, 735-761.

Jensen, J., Pedersen, M.S. 2015. Analysis of beamformer directed single-channel noise reduction system for hearing aid applications, Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE. pp. 5728-5732.

Kahneman, D. 1973. Attention and effort Prentice-Hall Englewood Cliffs, NJ.

Kjems, U., Jensen, J. 2012. Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement, Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE. pp. 295-299.

Koelewijn, T., Zekveld, A.A., Festen, J.M., Kramer, S.E. 2012. Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. Ear and Hearing 33, 291-300.

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

Kramer, S.E., Kapteyn, T.S., Festen, J.M., Kuik, D.J. 1997. Assessing aspects of auditory handicap by means of pupil dilatation. Audiology 36, 155-164.

Kuchinsky, S.E., Ahlstrom, J.B., Vaden Jr., K.I., 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.

Kuznetsova, A., Bruun Brockhoff, P., H., B.C.R. 2016. lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-33., https://CRAN.R-project.org/package=lmerTest.

Larsby, B., Hällgren, M., Lyxell, B. 2008. The interference of different background noises on speech processing in elderly hearing impaired subjects. . International journal of audiology 47(sup2), S83-S90.

Le Goff, N. 2015. Amplifiying Soft Sounds - A Personal Matter. Oticon Whitepaper. Whitepaper, Oticon A/S.

Loizou, P.C., Kim, G. 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. . IEEE Transactions on Audio, Speech, and Language Processing 19, 47-56.

Lunner, T., Rudner, M., Rosenbom, T., Ågren, J., Ng, E.H.N. 2016. Using Speech Recall in Hearing Aid Fitting and Outcome Evaluation Under Ecological Test Conditions. . Ear and Hearing 37, 145S-154S.

McAuliffe, M.J., Wilding, P.J., Rickard, N.A., O'Beirne, G.A. 2012. Effect of Speaker Age and Speech Recognition and Perceived Listening Effort in Older Adults With Hearing Loss. . Journal of Speech, Language, and Hearing Research 55, 838-847.

McGarrigle, R., Munro, K.J., Stewart, A.J., Dawes, P. 2014. Listening effort and fatigue: are we talking about the same thing? International journal of audiology.

Naylor G & Johannesson RB (2009) Long-Term Signal-to-Noise Ratio at the Input and Output of Amplitude-Compression SystemsJ Am Acad Audiol 20:161–171 (2009) DOI: 10.3766/jaaa.20.3.2

Neher, T. 2014. Relating hearing loss and executive functions to hearing aid users'preference for, and speech recognition with, different combinations of binaural noise reduction and microphone directionality. . Frontiers in Neuroscience 8.

Neher, T., Grimm, G., Hohmann, V., Kollmeier, B. 2013. Do Hearing Loss and Cognitive Function Modulate Benefit From Different Binaural Noise-Reduction Settings? . Ear and Hearing 35, 52-62.

Ng, E.H.N., Rudner, M., Lunner, T., Rönnberg, J. 2015. Noise reduction improves memory for target language speech in competing native but not foreign language speech. Ear and hearing 36, 82-91.

Ng, E.H.N., Rudner, M., Lunner, T., Pedersen, M.S., Rönnberg, J. 2013. Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. International Journal of Audiology 52, 433-441.

Nielsen, J.B., Dau, T. 2009. Development of a Danish speech intelligibility test. International journal of audiology 48, 729-741.

Nordrum, S., Erler, S., Garstecki, D., D., S. 2006. Comparison of performance on the hearing in noise test using directional microphones and digital noise reduction algorithms. . American Journal of Audiology 15, 81-91.

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., Maess, B. 2012. Adverse Listening Conditions and Memory Load Drive a Common Alpha Oscillatory Network. . The Journal of Neuroscience 32, 12376-12383.

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

Ohlenforst, B., Zekveld, A.A., Jansma, E.P., Wang, Y., Naylor, G., Lorens, A., Lunner, T., Kramer, S.E. 2017b. Effects of Hearing Impairment and Hearing Aid Amplification on Listening Effort: A Systematic Review. Ear and hearing 38, 267.

Pals, C., Sarampalis, A., Başkent, D. 2013. Listening effort with cochlear implant simulations. Journal of Speech, Language, and Hearing Research 56, 1075-1084.

Panico, J., C., H.E. 2009. Influence of Text Type, Topic Familiarity, and Stuttering Frequency on Listener Recall, Comprehension, Pand Mental Effort. Journal of Speech, Language, and Hearing Research 52, 534-546.

Petersen, E.B., Wöstmann, M., Obleser, J., Stenfelt, S., Lunner, T. 2015. Hearing loss impacts neural alpha oscillations under adverse listening conditions. . Frontiers in Psychology 6.

Pichora-Fuller, M.K., Kramer, S.E., Eckert, M.A., Edwards, B., Hornsby, B.W., Humes, L.E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C.L. 2016. Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). Ear and hearing 37, 5S-27S.

Picou, E.M., Ricketts, T.A., Hornsby, B.W.Y. 2011. Visual Cues and Listening Effort: Individual Variability. Journal of Speech, Language, and Hearing Research 54, 1416-1430.

Picou, E.M., Ricketts, T.A., Hornsby, B.W. 2013. How hearing aids, background noise, and visual cues influence objective listening effort. . Ear and Hearing 34, e52-e64.

Ricketts, T., Lindley, G., Henry, P. 2001. Impact of compression and hearing aid style on directional hearing aid benefit and performance. . Ear and Hearing 22, 348-361.

Rosen, S., Souza, P., Ekelund, C., Majeed, A.A. 2013. Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. The Journal of the Acoustical Society of America 133, 2431-2443.

Sarampalis, A., Kalluri, S., Edwards, B., Hafter, E. 2009. Objective Measures of Listening Effort: Effects of Background Noise and Noise Reduction. Journal of Speech, Language, and Hearing Research 52, 1230-1240.

Scott, S.K., Rosen, S., Wickham, L., Wise, R.J. 2004. A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. The Journal of the Acoustical Society of America 115, 813-821.

Scott, S.K., Rosen, S., Beaman, C.P., Davis, J.P., Wise, R.J. 2009. The neural processing of masked speech: evidence for different mechanisms in the left and right temporal lobes. The Journal of the Acoustical Society of America 125, 1737-1743.

Shinn-Cunningham, B.G., Best, V. 2008. Selective attention in normal and impaired hearing. . Trends in amplification 12, 283:299.

Siegle, G.J., Ichikawa, N., Steinhauer, S. 2008. Blink before and after you think: blinks occur prior to and following cognitive load indexed by pupillary responses. Psychophysiology 45, 679-687.

Simonsen, L.B., Hietkamp, R.K., Bramsløw, L. 2016. Learning effects of repeated exposure to Hearing In Noise Test, Annual Conference of the British Society of Audiology, Coventry, UK.

Smeds, K., Wolters, F., Rung, M. 2015. Estimation of signal-to-noise ratios in realistic sound scenarios. Journal of the American Academy of Audiology 26, 183-196.

Studebaker, G. A. (1985). A rationalized arcsine transform. Journal of Speech, Language, and Hearing Research, 28(3), 455-462.

Tun, P.A., McCoy, S., Wingfield, A. 2009. Aging, Hearing Acuity, and the Attentional Costs of Effortful Listening. . Psychology and Aging 24, 761-766.

Van Der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., Kramer, J., Warmuth, E., Heekeren, H.R., Wartenburger, I. 2010. Resource allocation and fluid intelligence: Insights from pupillometry. Psychophysiology 47, 158-169.

Wendt, D., Hietkamp, R.K., Lunner, T. 2017. Impact of noise and noise reduction on processing effort: A pupillometry study. The Journal of the Acoustical Society of America 141, 4040-4040.

Wendt, D., Koelewijn, T., Książek, P., Kramer, S.E, Lunner, T. 2017. 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. Accpeted for Hearing Research

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.

Wu, Y.-H., Aksan, N., Rizzo, M., Stangl, E., Zhang, X., Bentler, R. 2014. Measuring listening effort: Driving simulator vs. simple dual-task paradigm. Ear and hearing 35, 623.

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. 2011. Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. Ear and hearing 32, 498-510.

**Figure legends**

Fig. 1: Averaged pure tone hearing thresholds of the left and right ears across frequencies (125 Hz to 8 kHz) among the twenty-four hearing-impaired participants. Error bars show the standard deviations of the mean.

Fig. 2: Spatial loudspeaker setup as used in Wendt et al., 2017. Target speech was presented from the front. Masker signals were presented at 90, 150, 210 and 270 degree azimuth. The stationary noise masker was presented as four individual point sources. For the four-talker masker, one single talker was presented from one loudspeaker each.

Fig. 3: Peak pupil dilation (PPD) (black color) and percentage-correct sentence recognition scores (gray color) are shown on the right y-axis across the signal-to-noise ratios (SNRs) with the stationary masker and the noise reduction scheme turned on or off. Error bars represent the standard error of the mean. Dark gray diamonds at -12, -8, -4, 0 and +4 dB SNR represent significant differences in sentence recognition performance between the active and inactive noise reduction in the pairwise comparison at each SNR level ($p<0.01$).

Fig. 4: Peak pupil dilation (PPD) (black color) and the percentage of correct sentence recognition scores (gray color) are shown on the right y-axis across the signal-to-noise ratios (SNRs) with the 4-talker masker and noise reduction scheme on or off. Error bars represent the standard error of the mean. Dark gray diamonds at -8, -4, 0 and +4 dB SNR represent significant differences in sentence recognition performance between the active and inactive noise reduction in the pairwise comparison at each SNR level ($p<0.01$).

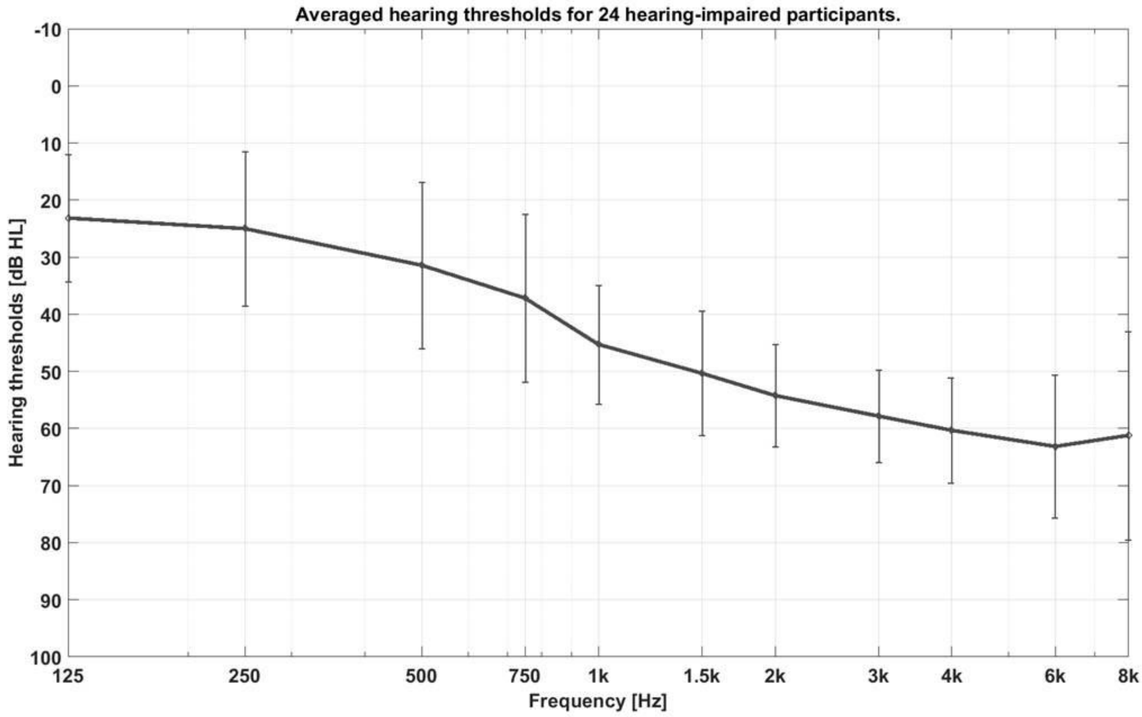


Fig. 1: Averaged pure tone hearing thresholds of the left and right ears across frequencies (125 Hz to 8 kHz) among the twenty-four hearing-impaired participants. Error bars show the standard deviations of the mean.

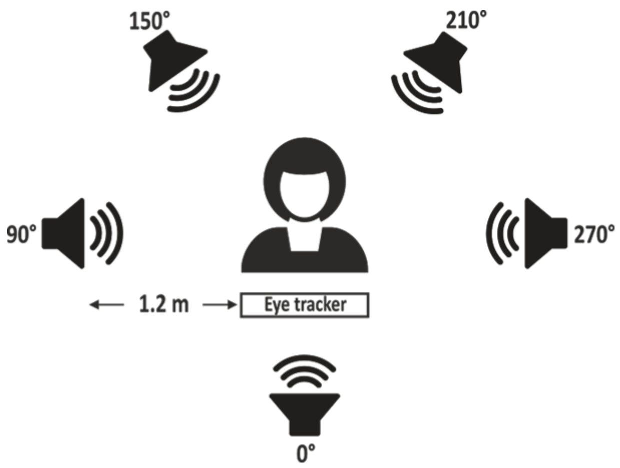


Fig. 2: Spatial loudspeaker setup as used in Wendt et al., 2017. Target speech was presented from the front. Masker signals were presented at 90, 150, 210 and 270 degree azimuth. The stationary noise masker was presented as four individual point sources. For the four-talker masker, one single talker was presented from one loudspeaker each.

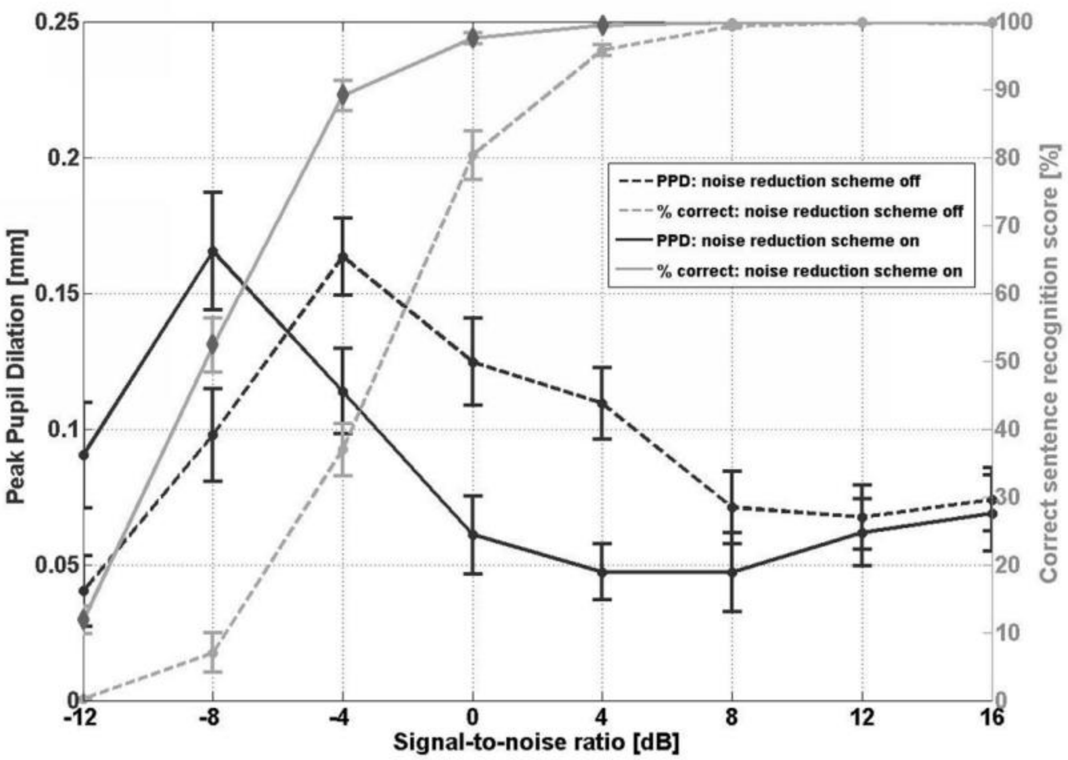


Fig. 3: Peak pupil dilation (PPD) (black color) and percentage-correct sentence recognition scores (gray color) are shown on the right y-axis across the signal-to-noise ratios (SNRs) with the stationary masker and the noise reduction scheme turned on or off. Error bars represent the standard error of the mean. Dark gray diamonds at -12, -8, -4, 0 and +4 dB SNR represent significant differences in sentence recognition performance between the active and inactive noise reduction in the pairwise comparison at each SNR level (p<0.01).

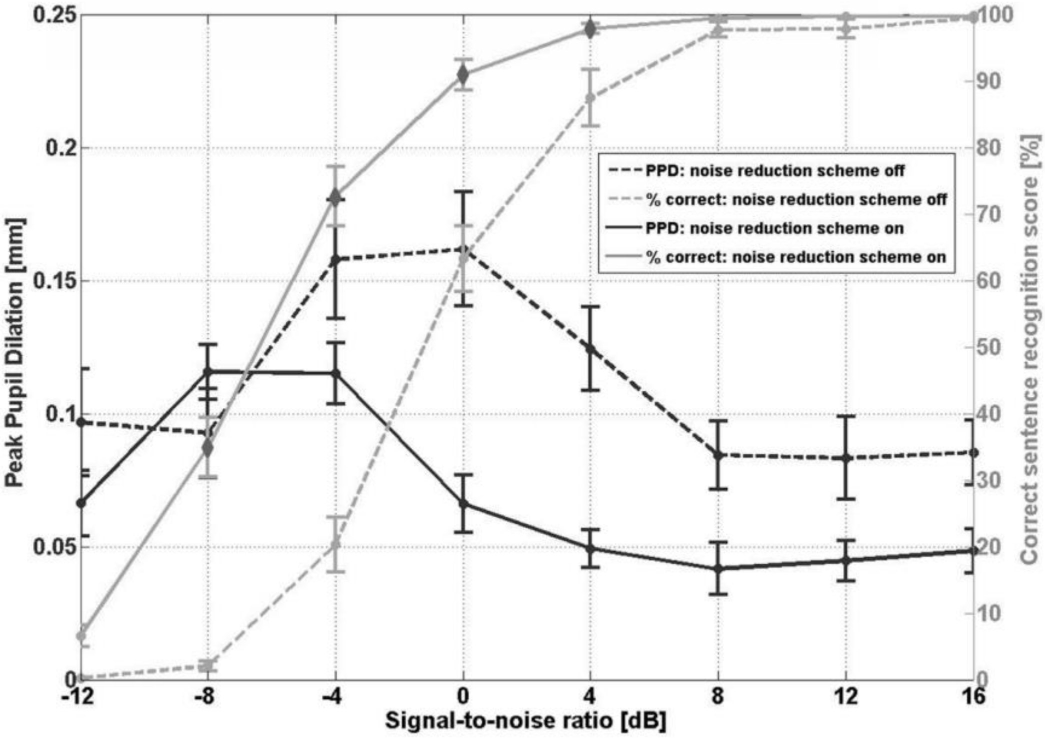


Fig. 4: Peak pupil dilation (PPD) (black color) and the percentage of correct sentence recognition scores (gray color) are shown on the right y-axis across the signal-to-noise ratios (SNRs) with the 4-talker masker and noise reduction scheme on or off. Error bars represent the standard error of the mean. Dark gray diamonds at -8, -4, 0 and +4 dB SNR represent significant differences in sentence recognition performance between the active and inactive noise reduction in the pairwise comparison at each SNR level ($p<0.01$).