Chapter 5

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

Ohlenforst, B.,
Wendt, D.,
Kramer, S. E,
Naylor, G.,
Zekveld, A. A.
Lunner, T.

Hearing Research (2018); Under Revision.
Abstract

Recent research showed that activating noise reduction scheme in hearing aids resulted in smaller peak pupil dilation (PPD), indicating reduced listening effort, at 50% and 95% correct sentence recognition in a 4-talker masker. The objective of this study was to measure the effect of noise reduction scheme (on vs. off) on PPD and sentence recognition across a wide range of SNRs from +16 dB to -12 dB, and for two masker types (4-talker and stationary noise). Relatively low PPDs were observed at very low (-12 dB) and at very high SNRs (+16 dB to +8 dB), presumably due to ‘giving up’ and ‘easy listening’, respectively. Maximum PPD was observed for SNRs at approximately 50% correct sentence recognition. Sentence recognition in both masker types was significantly improved by the noise reduction scheme, corresponding to shifting the performance vs. SNR function about 5 dB towards lower SNR. This intelligibility effect was accompanied by a corresponding effect on the PPD, shifting the peak about 4 dB towards lower SNR. In addition, for the 4-talker masker the PPD was overall smaller when noise reduction scheme was active versus inactive. We conclude that for the 4-talker masker, noise reduction scheme processing provides a listening effort benefit on top of any effect associated with improved intelligibility. Thus the effect of noise reduction scheme on listening effort incorporates more than we can explain by intelligibility alone, which emphasizes the potential importance of measuring listening effort in addition to traditional speech reception measures.
5.1 Introduction

Audiological evaluation and research on hearing aid signal processing has typically focused on changes or benefits in intelligibility, but has often failed to provide a complete picture of the processes involved during speech recognition (Dillon et al., 1993; Ricketts et al., 2001; Sarampalis et al., 2009). It has repeatedly been shown that traditional speech reception measures are insensitive to 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 from noise reduction processing in hearing aids may not be in enhanced speech intelligibility but rather in reduced listening effort.

According to the Framework for Understanding Effortful Listening (FUEL) (Pichora-Fuller et al., 2016), listening effort depends on a range of factors, including individual factors such as hearing ability and motivation to keep up listening and not to give up, 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 task demands become too high, or allocate less cognitive resources during very easy listening conditions (Ohlenforst et al., 2017a). In recent years, 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). Those methods include subjective assessments such as self-reports or 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 pupil response, fMRI or EEG measures (Kuchinsky et al., 2013; Obleser et al., 2012; Petersen et al., 2015). Crucially, listening conditions may affect listening effort even when speech intelligibility itself is not affected, for example when speech intelligibility is at 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 noise reduction scheme at ceiling performance reduced listening effort while speech in noise performance was unaffected. Therefore, assessing listening effort and speech performance simultaneously may uncover challenges or changes in processing speech that may not be evident with traditional measures.

Numerous studies in different research areas have shown that the pupil dilation increases with increasing processing load imposed by the task demands (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 to quantify effort required for speech recognition in 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). It has for instance been shown that the SNR (ranging from -20 dB to +16 dB) and the masker type (stationary and 1-talker masker) affect the pupil dilation during listening (Ohlenforst et al., 2017a). Recent studies indicate that effort is not necessarily monotonically related to task demands. Changes in effort rather follow an inverse U-shaped function, indicating that listeners may spend less effort due to ‘giving up’ at 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. They showed that the PPD, as an indication of cognitive processing load, was affected by an interaction of masker type and the hearing status of the individual. In the presence of a stationary noise masker, hearing-impaired listeners showed relatively large PPDs across a wide range of SNRs, while 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, maximum PPD was shown in the mid-range of SNRs, while relatively smaller PPDs were shown at low and high SNRs for 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 during listening may be different for normal-hearing and hearing-impaired listeners to reach similar speech understanding in daily life listening situations (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 quiet as well as in noisy environments. This may be accompanied by reduced listening effort. Advanced signal processing in hearing aids includes digital noise reduction scheme, which aims to reduce the level of interfering background noise by improving the SNR. Recent studies indicate that 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 and the multi-talker masker signal. This enhanced memory for the target words was interpreted as a sign of 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. They assessed the effect of a noise reduction scheme applying a combination of a digital noise reduction scheme and directional microphones. When the scheme was activated in the hearing aid, speech recognition performance at SNR50 was significantly improved, which was 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 a benefit of the noise reduction scheme on listening effort. The results demonstrated that measuring listening effort by assessing PPD provides a sensitive outcome measure of hearing aid benefit even at a high performance level, where 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 about the benefit of noise reduction processing on listening effort is still very limited as only a few listening conditions were tested in these studies. On the other hand, the effect of noise reduction processing on intelligibility has been studied by several groups of researchers. In that body of research, inconsistency in the diverse noise reduction processing schemes studied renders generalization problematic, especially as processing schemes increase
in sophistication over time. Some research has indicated that the application of noise reduction processing may not always be beneficial to speech intelligibility (Bentler et al., 2008; Nordrum et al., 2006). Such negative effects suggest that background noise may be removed but that the target speech might also be degraded. It has been suggested, that stronger or more aggressive signal processing may cause more signal enhancement, but may simultaneously introduce more degradation (Loizou et al., 2011). As a supporting example for this suggestion, in a recent study, the effect of noise reduction processing on sentence recognition was tested in the presence of 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 reverberant signal components and diffuse background-based noise reduction at mid to high frequencies, but did not include directionality. The results showed that sentence recognition was unaffected by moderate noise reduction processing, but that strong noise reduction processing reduced speech recognition by approximately 5%. The effect was replicated in a follow up study where the same acoustic test conditions were used on a group of habitual hearing aid users (Neher, 2014). It was found that strong noise reduction processing reduced speech recognition at -4 dB and 0 dB SNR compared to moderate or no noise reduction processing.

It still remains unclear how hearing-impaired listeners invest listening effort across a broader range of listening situations and how effortful listening relates to performance measures. The current study aimed to examine how a noise reduction scheme influenced sentence recognition and listening effort. The applied noise reduction scheme preserves speech and reduces noise in complex environments by means of 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 noise reduction processing on intelligibility will likely affect 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, besides this intelligibility effect, the noise reduction processing may have additional effects on the PPD, as suggested by recent literature on listening effort, demonstrating a benefit of hearing aid processing 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 an effect of noise reduction processing on listening effort, in combination with simultaneous knowledge about speech in noise performance, would further substantiate the value of measuring effort as an extra dimension in addition to traditional speech reception measures.

Recent research found better SRTs for speech recognition in the presence of a single-talker masker as compared with a stationary noise masker (Koelewijn et al., 2012). The additive interfering effect from the speech masker may cause a more difficult condition for speech recognition. We hypothesized, based on recent research, that better recognition performance for the 4-talker masker compared to the stationary noise masker, as the envelope modulations of the multi-talker masker should 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;
However, recent literature suggested that the intelligibility of speech, masked by additional interfering speech information, may require more mental effort compared to 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 better speech recognition is accompanied with larger PPDs for the 4-taker masker compared to the stationary noise masker. We hypothesized improved sentence recognition and reduced listening effort for SNRs corresponding to approximately 50% correct or better performance for the active noise reduction as compared to inactive noise reduction scheme. This hypothesis is motivated by two arguments. First, in previous research by Wendt and colleagues (2017), the SRT targeting 50% correct performance was significantly improved by the active noise reduction scheme compared to the inactive noise reduction scheme setting. Second, the segregation between 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.

5.2 Materials and Method

Participants
Twenty-five experienced hearing aid users were recruited at the Eriksholm Research Centre in Denmark. On average, participants had used hearing aids for 7.7 years (SD=3.1 years). The participants were between 46 and 77 years old (mean age 64.3 years, SD=9.4) and native Danish speakers. The audiometric inclusion criterion for the participants was symmetrical, mild to moderate sensorineural hearing thresholds. The average pure tone hearing thresholds had to lie between 35 dB and 60 dB HL (see Fig. 1) and air-bone gaps less than 10 dB between 500 Hz and 4000 Hz were required in both ears. All participants had normal or corrected-to-normal vision, and no history of neurological diseases, dyslexia or diabetes mellitus. All participants provided written informed consent and the study was approved by the local regional ethics committee (De Videnskabsetiske Komiteer for Region Hovedstaden).
Figure1: Averaged pure tone hearing thresholds for the left and the right ear across frequencies (125 Hz to 8kHz) for the twenty-four included hearing-impaired participants. Error bars show the standard deviations of the mean.

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 Fig. 2. The target sentences were spoken by a male talker and presented from the loudspeaker located at 0 degree azimuth. All sentences contain five words, 8-9 syllables are included in each sentence and single words do not contain more than four syllables (Nielsen et al., 2009). An example sentence is: “Filmen er rigtig godt lavet” (translation: “the movie was well made”). Sentence duration was on average 1.4 seconds. Listeners were presented with a training list of 20 sentences for each masker type, followed by eight lists of 25 sentences for every SNR. In order to cover the large amount of test conditions, the sentence material was re-used across four experimental visits. Recent research assessed possible learning effects due to repeated exposure to HINT sentences across three experimental visits with a gap of three weeks in between visits. The results showed that memory effects for the sentence material are not significant with limited exposure when the sentences were only presented once at each visit (Simonsen et al., 2016). The experimental visits in the current study were at least three weeks apart, and identical sentence material was not repeated within each visit to prevent learning effects of the speech material. Speech recognition performance was measured in the presence of a stationary noise or a 4-talker masker background. The 4-talker masker was created of four single-talker maskers, including two different male 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
We balanced the distribution of 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. For each trial, the masker started 3 seconds prior to the presentation of the sentence, and ended 3 seconds after the sentence offset. The participant 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 made 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 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, ecological sound environments for hearing-impaired listeners take place 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 the stationary masker and the 4-talker masker was presented at eight SNRs between -12 dB and +16 dB, distributed in steps of 4 dB. Per masker type, 25 sentences were presented for each SNR.

![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 one single talker was presented from one loudspeaker each.](image)

**Figure 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 one single talker was presented from one loudspeaker each.

### Noise reduction scheme

All participants were wearing an identical model of hearing aids during the sentence recognition test, which were 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 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, meaning that the microphone setting was close
to omnidirectional and no actual noise reduction was applied. The other setting resulted by turning the noise reduction scheme on. In that case 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 was applied to attenuate interfering sounds coming from the back of the listener.

**Pupillometry**

During the experiment, the pupil location and the pupil size were recorded by 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 the pupil size were recorded by the eye tracker and stored at a connected computer, including 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 case the participant moved the head or upper body or the real-time pupil recordings showed missing data in the pupil diameter, corrective actions were applied, such as the adjustment of the participants position, the distance to the eye tracker, or light adjustment.

**Procedure**

For 17 adults from the Eriksholm pool of participants, with recent pure tone audiogram data and recently made ear impressions (less than 6 month old), four experimental visits were required. We recruited 8 additional participants, for whom an extra recruitment visit (in total five visits) was required to measure the pure tone audiogram and to 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 2016 with gaps of at least three weeks in between each visit to avoid learning effects of the sentence material as it 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 a real-time recording of the pupil response from the eye tracking system to evaluate the quality of the pupil recording. The height of the chair and the distance to the eye tracker (55 cm +/- 5 cm approx.) 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 the 4-talker masker were presented at eight identical SNRs between -12 dB and +16 dB SNR, which were distributed in steps of 4 dB. During each visit only 1 out of 2 masker types was presented at two blocks of four randomized SNRs. For one block, the noise reduction scheme was turned on and for the other, the noise reduction scheme was turned off. During each visit, each noise reduction scheme setting (on versus 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 same noise reduction scheme setting as in the starting block
was tested for 20 sentences at an SNR of +4 dB. The practice session ensured that the participants were confident with the experimental procedure as it may not be intuitive to inhibit movements and blinking during the sentence presentation. A sentence was scored as correct if all words were correctly repeated.

**Pupil data selection and cleaning**

Pupil diameter values more than 2 standard deviations from the mean pupil diameter of a given trial were defined as blink. Pupil traces with more than 25% of blinks between the start of the baseline (last second of pre-noise before sentence onset) and the end of the post-masker were excluded from the data analysis. For pupil traces with less than 25% of blinks, blinks were interpolated linearly, starting 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 corresponds to the average pupil diameter recorded during the final second of the three second presentation of the masker, before 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 for 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 statistical analysis. Eighteen participants had the required number of valid pupil traces for each of the 32 test conditions. Six participants had less than 18 valid pupil traces for at least one test conditions and two participants had missing data (< 18 valid pupil traces) at 3 test conditions. We calculated an average pupil trace across all valid pupil traces per SNR condition and per subject. The mean PPD was calculated based on the averaged pupil trace and provided the data for the statistical analysis per SNR and participant.

**Statistical analyses**

Pupil data selection and cleaning was applied to pupil data from 24 participants (50% female). One participant had to be 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 LMM’s tolerate missing values, while repeated measures ANOVA only use complete cases contrary to multilevel analyses. Moreover, mixed-effects models are more flexible in dealing with the multilevel structure of the data (i.e. the 8 different SNRs and the 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 done 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
LMM to the data. First, we applied a 3-way LMM ANOVA for statistical comparison of the fixed effects of masker types, SNR and noise reduction setting separately on the PPD and the sentence recognition performance to verify the hypothesis that the masker type and the SNR range would have an impact on speech recognition performance and the corresponding listening effort. The probability level for each LMM ANOVA was $p<0.05$. We did not find a significant 3-way interaction on the PPD but a significant interaction between SNR and noise reduction scheme setting. The model was collapsed across masker types and an additional 2-way LMM ANOVA was applied to assess the effect of SNR and noise reduction scheme setting and the corresponding interaction on the PPD.

The three-way interaction between 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 SNR for each masker type independently (stationary noise and 4-talker masker) on percent-correct sentence recognition. In those models the averaged percentage correct sentence recognition scores for each SNR were treated as dependent measures with participants as the repeated measure and therefore the random effects. The fixed effects in each separate LMM ANOVA were the categorical variable SNR, the categorical variable noise reduction scheme setting and the interaction between SNR and noise reduction scheme setting. We included the random effect of SNR and noise reduction scheme as random slope of SNR, to allow each participant to have their own mean PPD size and their own 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 for the noise reduction scheme setting (on versus off) at each SNR level was carried out. The pairwise post-hoc analysis was separately applied for both outcome measures (PPD and sentence recognition performance) and a p-value correction according to the Holm method was applied to correct for multiple comparisons.
5.3 Results

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

We applied a 3-way linear mixed model (LMM) ANOVA, including noise reduction scheme, SNR and masker type as fixed factors to test the effect of all three factors on sentence recognition. We found 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 by the noise reduction processing algorithm. Furthermore, we found significant interactions between SNR and noise reduction scheme (F[7,713]=93.7, p<0.001), between SNR and masker type (F[7,713]=5.73, p<0.001) and between SNR, noise reduction scheme and masker type (F[7,713]=2.82, p<0.01). The interaction between masker type and noise reduction scheme was not significant. The interaction effects of masker type and noise reduction scheme with SNR are due to the fact that largest effects were observed in the mid-range of SNRs, while at relatively low and high SNRs, floor or ceiling effects of sentence recognition were observed.

For the stationary noise masker, at the relatively high SNRs between +16 dB and +8 dB SNR, participants achieved 100% sentence recognition independent of the setting of the noise reduction scheme. With decreasing SNR (+8 dB to -8 dB), sentence recognition dropped rapidly until the participants were not able to perform correct sentence recall at -12 dB SNR when the noise reduction scheme was turned off. At -12 dB SNR, participants were still able to recognize approximately 12% correct 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) towards lower SNRs when the noise reduction scheme was turned on compared to when it was turned off. A 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 SNR and noise reduction scheme (F[7,345]=68.8, p<0.001). We applied pairwise post-hoc comparisons between the two noise reduction scheme setting (on versus off) at each SNR level. The post-hoc analysis revealed significant differences between both noise reduction scheme settings at -12 dB, -8 dB, -4 dB and 0 dB SNR (indicated by gray diamonds for p<0.01 in Figure 3). Holm correction was applied to correct for multiple comparisons.
Figure 3: Peak pupil dilation (PPD) (black color) and percentage correct sentence recognition scores (gray color) on the right y-axis across signal-to-noise ratios (SNRs) for the stationary masker for noise reduction scheme on versus 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 active versus inactive noise reduction for pairwise comparison at each SNR level (p<0.01).

For the 4-talker masker, at SNRs between +16 dB and +8 dB, close to 100% sentence recognition was achieved regardless of noise reduction settings. The overall performance curve was shifted by approximately 5.1 dB towards lower SNRs when the noise reduction scheme was turned on compared to when it was turned off. Applying a 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 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 setting (on versus off) at each SNR level. Significant differences in sentence recognition performance between both noise reduction scheme settings revealed at -8 dB, -4 dB, 0 dB and +4 dB SNR (indicated by gray diamonds for p<0.01 in Figure 4). Multiple comparisons were accounted for by applying Holm correction.
Figure 4: Peak pupil dilation (PPD) (black color) and percentage correct sentence recognition scores (gray color) on the right y-axis across signal-to-noise ratios (SNRs) for the 4-talker masker for noise reduction scheme on versus 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 active versus inactive noise reduction for pairwise comparison at each SNR level (p<0.01).

Pupil data
Figure 3 shows the PPD for the stationary noise masker and Figure 4 shows the PPD for the 4-talker masker across SNRs. We tested a 3-way LMM ANOVA, including the noise reduction scheme, SNR and the masker type as fixed factors on the PPD. We found significant main effects of 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 between SNR and noise reduction scheme ($F[7,699.1]=9.97$, $p<0.01$). There was no significant interaction between masker type and SNR or masker type and noise reduction scheme. The interaction effect between SNR and noise reduction scheme suggests that the SNR-dependency of the PPD is different when noise reduction scheme is on vs. off. One additional LMM ANOVA model, in which we collapsed across the level of masker type, was built. We did not test two separate models for each masker type as done for the sentence recognition performance. The reason for applying a 2-way LMM ANOVA that does not differentiate between masker types is that the interaction between SNR, noise reduction scheme setting and masker type was not significant. 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 between noise reduction scheme setting and SNR \((F[7,715.05]=9.72, p<0.001)\) on the PPD. Pairwise post-hoc comparisons between the two noise reduction scheme setting (on versus off) was applied at each SNR level. Significant differences between both noise reduction scheme settings were found for the PPD measured at -8 dB, -4 dB, 0 dB and +4 dB SNR. Holm correction was applied to correct for multiple comparisons.

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

Figure 4 shows the PPD across SNR for the noise reduction scheme on (black, solid curve) versus off (black, dashed curve) for the 4-talker masker condition. The PPD measured for the high SNRs between +16 dB and +8 dB was overall consistent, but larger for the noise reduction scheme off versus when it was turned on. Further decreasing SNRs resulted in continuously increasing PPD until a 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 turned on. The range of maximum PPD was shifted by approximately 4 dB towards lower SNRs when the noise reduction scheme was turned on versus when it was turned off.

**Results summary**

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 by approximately 5.1 dB for the 4-talker masker, towards lower SNRs. For both masker types, the effect of the noise reduction scheme on listening effort (as measured by PPD) was to shift the peak of the PPD function across SNRs by approximately 4 dB towards lower SNR. In addition, in the case of the 4-talker masker, noise reduction scheme lowered the average PPD by approximately 35% compared to inactive noise reduction scheme. The beta estimates on the sentence recognition performance scores and the PPD for each SNR level can be found in Appendix C.
5.4 Discussion

In the present study, the effect of a noise reduction scheme on sentence recognition and PPD was examined across a range of SNRs for 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.

Relation between noise reduction, SNR and speech recognition

For the stationary and the 4-talker masker, sentence recognition performance was significantly improved when the noise reduction scheme was active versus when it was inactive. The results showed improved sentence recognition for performance levels around 50% and higher, but also for 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 from the present study confirm and extend previously shown benefits of a noise reduction scheme on sentence recognition around 50% correct performance (Wendt et al., 2017) as well as 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 might allow hearing-impaired listeners to participate in communication situations that might otherwise be impossibly challenging.

Relation between noise reduction, SNR and PPD

In line with recent research (Ohlenforst et al., 2017a; Zekveld et al., 2014), the present results confirm that changes in speech recognition are accompanied by changes in PPD. We found a maximum PPD for 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 findings from a recent study (Wu et al., 2016), where dual-task paradigms were applied to assess listening effort across a wide range of SNRs. Wu et al. found that second-task performance (reaction time) was worst (i.e. longest) around SNRs for 30-50% speech recognition, and better at both lower and higher SNRs. The change of the PPD function at positive SNRs, when percent correct sentence recognition is saturated, might be affected by the type of speech material used for the sentence recognition test. That is as the difficulty of the speech material can change the transfer function. 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. In other words, 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 on purpose to end up at ceiling performance and at very positive SNRs ceiling performance will be reached anyway, regardless of the presented speech material.

The statistical analysis on the PPD revealed that the level of SNR and the noise reduction scheme setting significantly affect the PPD. The results did not suggest that the PPD, as indication of listening effort during speech recognition, is altered by the type of background masker. The null-effect of masker type on the PPD is in contrast to previous pupillometry studies, showing that speech recognition and the PPD changed depending on the masker
type (Koelewijn et al., 2012; Koelewijn et al., 2014). The application of neural speech tracking in different sound environments underlined that multi-talker sound scenarios are especially for hearing-impaired listeners problematic (Borch Petersen et al., 2017). The researchers suggest that hearing-impaired listeners seem to track the entire auditory scene, including target and masking speech in multi-talker environments. Hearing-impaired listeners seem to have difficulties to neurally inhibit the masking speech, which could relate to the often reported difficulties with speech intelligibility in multi-talker sound scenarios (Shinn-Cunningham et al., 2008). The noise reduction scheme activation in the present study grants better speech recognition performance by improving the SNR conditions for the hearing-impaired listeners. Better SNR conditions might allow better inhibition of interfering information which is may be reflected by the main effect of noise reduction processing on the PPD.

One strength of the present study is the replication of previous findings, showing a beneficial effect of a noise reduction scheme in hearing aids on sentence recognition and the PPD (Wendt et al., 2017). Several factors that were kept constant between both studies. 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). Also, a large number of listeners (n=17) that participated in the study by Wendt et al., (2017) were repeatedly tested in the present study. Both studies accommodate the field of hearing research and listening effort with new valuable knowledge by showing possible benefits of a noise reduction scheme for hearing-impaired listeners wearing hearing aids.

5.5 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. For both maskers, noise reduction processing resulted in a shift of the performance (sentence recognition) function towards lower (more challenging) SNRs, and a corresponding shift of the PPD function. For the 4-talker masker, on top of the speech recognition-related reduction in PPD, a main effect of noise reduction processing on the PPD was found. This may indicate that cognitive processing load and some aspects of listening effort may be reduced, independent of the SNR. The 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 SNR. This knowledge is essential for future research in the field of listening effort and for the hearing aid industry to improve the development of better hearing aid algorithms.

5.6 Acknowledgements

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 from the Eriksholm Research Center in Denmark for her support with the participant recruitment and the data collection and Nicolas Le Goff and Jesper Jensen from the Oticon Headquarters in Denmark for 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.