



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

Wendt, Dorothea; Koelewijn, Thomas; Książek, Patrycja; Kramer, Sophia E.; Lunner, Thomas

Published in:
Hearing Research

Link to article, DOI:
[10.1016/j.heares.2018.05.006](https://doi.org/10.1016/j.heares.2018.05.006)

Publication date:
2018

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):
Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research*, 369, 67-78. <https://doi.org/10.1016/j.heares.2018.05.006>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Accepted Manuscript



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

Dorothea Wendt, Thomas Koelewijn, Patrycja Książek, Sophia E. Kramer, Thomas Lunner

PII: S0378-5955(17)30529-4

DOI: [10.1016/j.heares.2018.05.006](https://doi.org/10.1016/j.heares.2018.05.006)

Reference: HEARES 7554

To appear in: *Hearing Research*

Received Date: 3 November 2017

Revised Date: 7 April 2018

Accepted Date: 9 May 2018

Please cite this article as: Wendt, D., Koelewijn, T., Książek, P., Kramer, S.E., Lunner, T., Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test, *Hearing Research* (2018), doi: 10.1016/j.heares.2018.05.006.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1 **Toward a more comprehensive understanding of the impact of masker type and**
2 **signal-to-noise ratio on the pupillary response while performing a speech-in-**
3 **noise test.**

4 Dorothea Wendt^{1,2}, Thomas Koelewijn³, Patrycja Książek¹, Sophia E. Kramer³,
5 Thomas Lunner^{1, 2, 4, 5}.

6
7 ¹*Eriksholm Research Centre, Snekkersten, Denmark*

8 ²*Hearing Systems, Hearing Systems group, Department of Electrical Engineering, Technical University of*
9 *Denmark, Kgs. Lyngby, Denmark*

10 ³*Section Ear & Hearing, Dept. of Otolaryngology-Head and Neck Surgery, and Amsterdam Public Health*
11 *Research Institute VU University Medical Center, Amsterdam, The Netherlands*

12 ⁴*Division of Technical Audiology, Department of Clinical and Experimental Medicine, Linköping University,*
13 *Linköping, Sweden,*

14 ⁵*Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden,*

15
16
17 **Keywords:**

18 Listening Effort, Pupillometry, Pupil dilation, Speech-in-Noise Test, Speech Recognition, Signal-to-Noise
19 Ratio, Growth Curve Analysis,

20
21 **Financial Disclosures/Conflicts of Interest:**

22 This research was funded by the Oticon Foundation (grant 14-0845).

23
24 **Correspondence:**

25 Dorothea Wendt
26 Eriksholm Research Centre, Snekkersten, Denmark
27 Email: dowe@eriksholm.com

28

29

30 Abstract

31 Difficulties arising in everyday speech communication often result from the acoustical environment,
32 which may contain interfering background noise or competing speakers. Thus, listening and
33 understanding speech in noise can be exhausting. Two experiments are presented in the current
34 study that further explored the impact of masker type and Signal-to-Noise Ratio (SNR) on listening
35 effort by means of pupillometry. In both studies, pupillary responses of participants were measured
36 while performing the Danish Hearing in Noise Test (HINT; Nielson and Dau, 2010). The first
37 experiment aimed to replicate and extend earlier observed effects of noise type and semantic
38 interference on listening effort (Koelewijn et al., 2012). The impact of three different masker types,
39 i.e. a fluctuating noise, a 1-talker masker and a 4-talker masker on listening effort was examined at
40 a fixed speech intelligibility. In a second experiment, effects of SNR on listening effort were
41 examined while presenting the HINT sentences across a broad range of fixed SNRs corresponding
42 to intelligibility scores ranging from 100 % to 0 % correct performance. A peak pupil dilation
43 (PPD) was calculated and a Growth Curve Analysis (GCA) was performed to examine listening
44 effort involved in speech recognition as a function of SNR. The results of two experiments showed
45 that the pupil dilation response is highly affected by both masker type and SNR when performing
46 the HINT. The PPD was highest, suggesting the highest level of effort, for speech recognition in the
47 presence of the 1-talker masker in comparison to the 4-talker babble and the fluctuating noise
48 masker. However, the disrupting effect of one competing talker disappeared for intelligibly levels
49 around 50 %. Furthermore, it was demonstrated that the pupillary response strongly varied as a
50 function of SNRs. Listening effort was highest for intermediate SNRs with performance accuracies
51 ranging between 30 % -70 % correct. GCA revealed time-dependent effects of the SNR on the
52 pupillary response that were not reflected in the PPD.

53

54 **1. INTRODUCTION**

55 Everyday communication requires recognizing and understanding speech in adverse listening
56 situations. External sound sources such as background noise or competing speech can degrade the
57 target speech, which makes following a conversation effortful. Presence of the masking noise
58 demands extra cognitive resources to process, comprehend, and remember speech (Rönnberg et al.,
59 2013). Allocation of additional resources can furthermore lead to higher listening effort, which has
60 recently been defined as *the deliberate allocation of mental resources to overcome obstacles to goal*
61 *pursuit when carrying out a listening task* (Pichora-Fuller et al., 2016). Consequences of increased
62 effort can be, for example, higher levels of mental distress and fatigue leading to stress, greater need
63 for recovery after work, or increased incidence of stress-related sick leave (Gatehouse and Gordon,
64 1990; Kramer et al., 2006; Edwards, 2007; Hornsby, 2013).

65 Within recent years, there has been a growing interest in identifying factors that cause
66 difficulties occurring during speech perception in noise. In audiological research, speech perception
67 is commonly explored using speech audiometry. Traditionally, speech-in-noise tests (e.g. Plomp
68 and Mimpen, 1979; Hagerman, 1982) are applied to measure either the proportion of correctly
69 repeated speech items (intelligibility) - usually single words or single sentences - or the speech
70 reception threshold (SRT) when the speech intelligibility is fixed. Recent studies have demonstrated
71 that measuring speech recognition performance or SRTs within a speech-in-noise test do not
72 capture the whole picture of speech perception. In fact, it has been demonstrated that while
73 maintaining similar intelligibility, listening effort varies depending on acoustical and linguistic
74 aspects of the speech and the masker signal (Koelewijn et al., 2012b; Wendt et al., 2016, 2017). For
75 instance, masker types containing speech/linguistic information might lead to increased effort even
76 with a constant speech intelligibility (Koelewijn et al., 2012a).

77 Physiological measures, such as pupillometry, have been applied recently to examine the
78 effort accompanying understanding speech in adverse listening environments. It has been shown
79 that listening effort increases together with increasing task demands, which is reflected in an
80 increased pupil size until processing resources are exceeded (Janisse, 1977; Beatty, 1982,
81 Kahnemann and Beatty, 1966; Zekveld et al., 2010, 2011). By applying pupillometry within a
82 speech-in-noise test, several studies explored the impact of hearing status, masker type, or speech
83 intelligibility on listening effort (e.g., Koelewijn et al., 2014; Zekveld et al., 2010; Wendt et al.,
84 2017). Those studies indicated that the impact of masking noise on speech perception and listening
85 effort can be manifold and intricate depending on characteristics of the masker or the SNR, which
86 will be discussed in more detail below. Moreover, the relationship between the performance of the
87 participants and their pupil response during speech processing is not fully captured yet.

88 Toward the goal of a more comprehensive understanding of listening effort measured by
89 means of pupillary response within a speech-in-noise test, the present work includes two studies
90 that examined the effect of masker type and SNR on effort. Two different experiments will be
91 presented that both applied pupillometry within the Danish Hearing In Noise Test (HINT, Nielsen
92 and Dau 2011): Experiment 1 is exploring the impact of different noise masker types that contain
93 linguistic information on effort when speech intelligibility is kept constant at 50 % and 84 %.
94 Experiment 2 investigates changes in listening effort across a broad range of fixed SNRs
95 corresponding to speech intelligibility between 0 % and 100 % correct recognition. Both
96 experiments were designed to provide insights into the relationship between listening effort
97 (reflected in the pupillary response) and recognition performance (as indicated by the SRT or %-
98 correct performance) using the HINT test across a wide range of acoustic scenarios.

99 *About the impact of the masker type on listening effort:*

100 Generally, two different ways have been distinguished in how the masker signal interferes with the
101 speech signal. If the masker signal coincides in spectrum and time with the speech signal, it is
102 referred to as *energetic masking* (Pollack, 1975; Brungart et al., 2001). Energetic masking is
103 supposed to take place at more peripheral stages of processing. All masking that is not considered
104 as energetic masking and occurs at a more central processing stage is often designated as
105 *informational masking* (Pollack, 1975). Many attempts have been made to disentangle the impact of
106 energetic and informational masking on auditory processing by using different types of maskers
107 (Festen and Plomp, 1990; Hygge et al., 1992; Brungart et al., 2001). In particular, speech-on-speech
108 masking has been of interest when studying speech perception. The interfering speech signal leads
109 to an informational masking effect due to its lexical and semantic content that creates contextual
110 overlap with the target speech (Kidd et al., 2008). However, the interfering content seems to have
111 only little influence on the intelligibility of the target speech. Festen and Plomp (1990) compared
112 speech recognition for three different types of masker, namely a stationary noise, a fluctuating noise
113 and a 1-talker masker. It was reported that the temporal dips of the interfering masker led to a better
114 recognition performance in the presence of a fluctuating noise compared to a stationary noise
115 masker. In addition, recognition performance for speech masked by a 1-talker masker was
116 comparable to the performance for the fluctuating noise masker. Interestingly, at low fixed
117 intelligibility (e.g., 50 %) participants have even shown lower (better) SRT scores for the 1-talker
118 masker compared to the fluctuating noise masker (Festen and Plomp, 1990; Koelewijn et al.,
119 2012b). While the effect of competing linguistic information on the speech recognition performance
120 seems to be rather small and sometimes beneficial, it has been demonstrated that its impact on
121 listening effort is major. Koelewijn et al. (2012a, b) examined the pupillary response within a Dutch
122 speech-in-noise test (Versfeld et al., 2000) for three different masker types. The authors
123 demonstrated significantly larger pupil dilation responses for speech masked by a 1-talker masker

124 compared to fluctuating noise and/or stationary noise. As concluded by the authors, this increase in
125 effort was mainly explained by the semantic inference with the target speech.

126 Unfortunately, the design used by Koelewijn et al. (2012) did not allow to distinguish
127 between the masking based on voice characteristics (e.g., timbre, fundamental frequencies) and the
128 actual semantic content. This was due to the fact that the fluctuating noise used in this study only
129 contained the speech envelopes while voice characteristics, which enable us to recognize human
130 speech (even for unknown languages) and let us to tell different voices apart, were not preserved.
131 Hence, these voice characteristics available in the 1-talker masker could explain the observed effect
132 rather than the linguistic information. Some studies have indicated that for babble maskers
133 comprising up to three competing talkers, masking is still highly affected by individual voice
134 characteristics of the talker, whereas for four or even more competing talkers the characteristics of
135 the individual voices become less prominent (Simpson and Cooke, 2005). At the same time, several
136 studies indicated that the intelligibility of the target sentence decreases with increasing number of
137 competing talkers (Simpson and Cooke, 2005; Rosen et al., 2013). This has to be related to the
138 decreasing possibility of listening into the dips since the envelope of the summed signal starts to
139 smooth out with an increasing number of talkers. However, not so much is known about how
140 listening effort changes with more than one interfering talker in the background.

141 The goal of Experiment 1 was to disentangle the impact of linguistic information (semantic
142 interference) and voice characteristics on listening effort and performance (i.e. SRTs) measured
143 within the Danish HINT. This by examining the impact of different masker types including a 1-
144 talker vs a 4-talker masker on listening effort. For that purpose, an experiment similar to Koelewijn
145 et al. (2012) was conducted. Listening effort was measured by means of pupil dilation and three
146 different masker types were used (a fluctuating noise, a 1-talker and a 4-talker masker). By using a
147 4-talker masker in contrast to a 1-talker masker, the SRT was expected to be higher (worse) due

148 reduced opportunities of listening in the dips. At the same time, linguistic information provided by
149 the 4-talker masker was considered to be less audible than for the 1-talker masker, which should
150 lead to less semantic interference and, hence, a smaller pupil dilation response. Consequently, it was
151 hypothesized that listening effort is highest for speech recognition within the 1-talker masker
152 compared to the 4-talker masker (H1), while at the same time, the 4-talker babble was expected to
153 result in higher (worse) SRT due to a smaller beneficial effect of dip listening (H2).

154 *About the impact of speech intelligibility and SNR on listening effort:*

155 Recent studies indicated that listening effort changed in a non-linear way with decreasing SNRs.
156 Ohlenforst et al. (2017) explored the impact of hearing-impairment on listening effort as a function
157 of SNRs for speech recognition in noise. Peak pupil dilations (PPD) were measured for participants
158 performing a speech-in-noise test at eight different SNRs. The authors reported increased PPDs,
159 suggesting higher listening effort, with decreasing SNRs. Interestingly, the pupil dilation reached a
160 maximum value until speech recognition performance was around 40-50 %. When recognition
161 decreased even further (i.e. to < 40 % correct recognition) the pupil dilation dropped again. This
162 decline in pupil dilation was interpreted as a sign of giving up because listening might become too
163 difficult in those conditions, which is supported by the Framework for Understanding Effortful
164 Listening (FUEL, see Pichora-Fuller et al., 2016).

165 The FUEL assumes that listening effort is not only affected by the task demands, but further by the
166 individual's motivation to complete the task. When task demands increase, due to decreasing SNRs,
167 more cognitive resources are allocated presumably leading to elevated levels of effort. However,
168 resources are limited and when task demands become too high and benefits are no longer outweighs
169 these costs, signs of "quitting" might be observed (Pichora-Fuller et al., 2016). A non-linear change
170 of the effort in form of an inverted U-shape has further been reported by Wu et al. (2016). The
171 authors measured reaction times within a dual-task paradigm with a primary sentence recognition

172 task and a secondary (visual) tasks across a range of SNRs. Again, reaction times became shorter
173 and subjective effort ratings lower at lowest SNRs with recognition performance below 50 %
174 indicating reduced effort. Similar findings of a neural activity breakdown with increasing memory
175 loads have been reported in other studies examining alpha power of the electroencephalogram (see
176 e.g. Wisniewski et al., 2017; Sander et al., 2012 for a visual task or Petersen et al., 2015 for an
177 auditory task; McMahon et al., 2016 for combining EEG and pupil dilation) and further by fMRI
178 studies (Reuter-Lorenz and Cappell, 2008; Grady, 2012). Taken together, those studies
179 demonstrated that when testing effort at a few constituent SNRs only, one might not cover the
180 whole pattern of listening effort and its changes across a broader range of listening situations. In
181 particular, the breaking point of listening effort, as indicated by the highest pupil dilation or reaction
182 time was reported at around 40-50 % speech recognition performance, remains undetected.

183 Moreover, recent studies indicated that changes in effort can be found in listening situations
184 with constant performance levels, which led to the assumption that changes in performance (as
185 indicated by % correct performance or SRT) and listening effort (as indicated by the pupil dilation)
186 are not necessarily related to each other (see Koelewijn et al., 2012a; McGarrigle et al., 2014;
187 Wendt et al., 2017). Those studies indicated that listening effort could point towards problems
188 occurring during speech recognition that are not addressed by performance data and vice versa.

189 The motivation of Experiment 2 was to expand the finding of Experiment 1 by including a
190 number of important differences. Instead of fixed intelligibility, changes in listening effort were
191 explored within a speech-in-noise test across a wide range of eight different SNRs. Thereby, a wide
192 range of acoustic scenarios can be covered including ecological listening situations with high
193 speech intelligibility (Smeds et al., 2015). Based on previous studies (Ohlenforst et al., 2017; Wu et
194 al., 2016), it was hypothesized that the pupil dilation would change as a function of SNR with
195 having a maximum dilation around 40-50 % speech intelligibility (H3). In order to investigate

196 listening effort in a more realistic acoustic environment, pupillary response was measured within a
197 spatial setup of loudspeaker presenting either a 4-talker babble or a stationary noise while
198 participant performing the Danish HINT test. It was expected that the maximum pupil dilation
199 would occur at lower (negative) SNRs for the 4-talker babble compared to the stationary noise (H4).
200 A further motivation for combining those two studies (Experiment 1 and 2) was towards a better
201 understanding of the pupillary response indicating the listening effort involved in a speech-in-noise
202 test (namely the Danish HINT test; Nielsen and Dau, 2011). For that purpose, the findings of both
203 experiments will be discussed with regard to potential applications of pupillometry within a speech-
204 in-noise test as an assessment tool for clinical populations.

205

206 **2. EXPERIMENT 1**

207 **2.1. MATERIALS AND METHODS**

208 **2.1.1 Participants**

209 Nineteen participants (aged from 18 to 63 years, mean 32.7 years, 9 male) with normal hearing
210 participated. They were native Danish speakers and had pure tone hearing thresholds for both ears
211 of 20 dB hearing level (HL) or better for octave frequencies between 125 Hz - 4 kHz and 30 dB HL
212 or better for octave frequencies between 6 - 8 kHz. The participants had no history of eye diseases
213 or eye operations. The experiment was carried out without the use of glasses or contact lenses.
214 Ethical approval for the study was obtained from the Research Ethics Committees of the Capital
215 Region of Denmark.

216 **2.1.2. Stimuli and procedure**

217 Danish sentences, spoken by a male speaker, from the HINT (Nielsen and Dau, 2011) were
218 presented via headphones with three different maskers, i.e. a 1-talker masker, a 4-talker masker and
219 a temporary fluctuating noise masker. The 1-talker masker consisted of a single female talker

220 reading text from the newspaper. The masker was created by concatenating two speech streams
221 uttered by two different female speakers reading from a newspaper. All breathing pauses (i.e.
222 speech pauses longer than 50ms) were removed. The masker was furthermore spectrally shaped to
223 obtain the same long-term average frequency spectrum as the target sentences. All speaker specific
224 short-term fluctuations of the masker were maintained. The 4-talker masker was created by
225 overlapping two male and two female talkers (all reading text from a newspaper), of which the
226 audio files had the same long-term average frequency spectrum as the HINT sentences. Finally, the
227 fluctuating masker consisted of a noise with the same average frequency spectrum and similar
228 intensity fluctuations of the HINT sentences. To mimic similar temporal intensity fluctuations, the
229 noise signal was multiplied by the envelope of the HINT sentences for two separate frequency
230 bands below and above 1 kHz (Festen and Plomp, 1990).

231 HINT sentences were presented with one out of the three different masker audio files. In
232 each trial the masker started 3 s before the onset of each HINT sentence and continued for 3 s after
233 sentence offset. The length of each trial varied depending on the length of the presented HINT
234 sentence, which had a mean duration of about 1.5 s. After masker offset, the participants were asked
235 to repeat back the HINT sentence. The total experiment consisted of six different conditions
236 including three different masker types (fluctuating, 1-talker, and 4-talker) and two different SRTs
237 that were performed in a blocked fashion. To ensure comparable speech intelligibility, every
238 participant performed the test at his or her individual SRTs corresponding to either 84 % or 50 %
239 sentence intelligibility respectively by using a staircase procedure based on full sentence correct
240 scoring. To obtain the SRT at 50 % intelligibility, a 1-up-1-down procedure was applied (Plomp
241 and Mimpen, 1979). After a correct response, the SNR increased by 2 dB and after an incorrect
242 response the SNR decreased by 2 dB. In order to measure the SRT at 84 % intelligibility, a 4-up-1-
243 down procedure was used. For each block, the SNR of the first trial started below threshold (i.e. -

244 15 dB SNR). The first sentence of each block was repeated until the participant correctly repeated
245 the entire sentence. The sound level of the mixed signal (speech and noise) was constant at 70 dB
246 SPL, regardless of SNR. Each block consisted of 33 trials and took about 15 min. After the second
247 and fourth block, participants had a break of 10 min.

248 In total, 6 blocks, i.e. one for each condition, with 33 trials each were presented in a
249 randomized order. In addition, participants performed one training block consisting of 30 sentences
250 (10 sentences for each noise masker type) at the beginning of the session. The complete
251 measurement took about 2.5 hours per participant.

252 **2.1.3 Apparatus**

253 During the speech perception task the pupil diameter of both eyes were recorded by an eye-tracker
254 system (iView X RED System, SensoMotoric Instruments, Teltow, Germany) with a sampling rate
255 of 60 Hz. An infrared camera that tracked the eye and head position automatically was placed in
256 front of the listener to record both eyes. The presentation of stimuli was controlled by a PC using
257 MATLAB (MathWorks, Natick, MA) based programming. Auditory signals were routed through a
258 sound card (RME Hammerfall DSB multiface II, Audio AG, Haimhausen, Germany) and presented
259 via closed headphones (Sennheiser HDA 200, Wedemark, Germany) in a double-walled and
260 acoustics-treated room (IAC Acoustics, Hvidovre, Denmark). The participants were seated 60 cm
261 from the eye-tracker and the luminance in the booth was adapted such that the pupil diameter was
262 around the middle of its dynamic range. The pupil size and pupil x- and y-traces of both eyes were
263 recorded to detect horizontal and vertical eye movements, respectively.

264 **2.1.4. Pupil Data Processing**

265 Pupil data were processed using MATLAB (MathWorks, Natick, MA) in line with a previous study
266 (see Wendt et al., 2017). Pupil traces of the first 3 trials were removed from further analysis. For all
267 remaining traces the mean pupil dilation and standard deviation was calculated from 3 s prior to the

268 sentence onset until the noise offset. Pupil diameter values more than 3 standard deviations smaller
269 than the mean were coded as eye-blinks. Eye-blinks were removed by a linear interpolation that
270 started about 80 ms before and ended 150 ms after the blinks. Trials that consisted for more than 20
271 % of their duration of eye-blinks, gross artefacts or missing data were excluded from further
272 analysis. A moving average filter with a symmetric rectangular window of 117 ms length was used
273 to smooth the de-blinked trials and to remove any high-frequency artefacts. All remaining traces
274 were baseline corrected by subtracting the mean pupil size as measured within the 1 s preceding to
275 sentence onset from each individual trace. After baseline correction traces were averaged for each
276 condition. Consistent with previous studies the peak pupil dilation (PPD) was calculated between 3
277 s and 7.5 s of stimulus presentation (Zekveld et al., 2010, 2011; Koelewijn et al., 2012; Wendt et al.,
278 2017). This time segment was chosen since a local peak of the pupillary response is usually
279 observed within that segment. Furthermore, it is assumed that the listener would process the
280 sentence and prepare the task (repeating back) during that interval. The PPD was calculated for each
281 participant and each condition.

282

283 **2.2. RESULTS EXPERIMENT 1**

284 **2.2.1. Behavioural data**

285 The average SRTs were calculated for all three masker types and both intelligibility scores for each
286 participant. Results for each condition averaged over participants are shown in Figure 1. An
287 ANOVA on the SRTs revealed a main effect of intelligibility ($F[1,18] = 358, p < 0.001$) indicated
288 by a significantly higher SRT at 84 % compared to 50 % intelligibility. In addition, a main effect of
289 masker type was shown ($F[1,18] = 285, p < 0.001$). No significant interaction between intelligibility
290 and masker type was observed. Post-hoc analysis was performed by t-tests (two-tailed paired
291 samples) and revealed higher thresholds for the 4-talker babble compared to fluctuating noise ($p <$

292 0.001) and the 1-talker masker ($p < 0.001$). Furthermore, higher SRTs were measured for the
293 fluctuating noise compared to the 1-talker masker ($p < 0.001$).

294

295 *[FIGURE 1 about here]*

296

297 **2.2.2. Pupillometry**

298 Figure 2 depicts the PPDs averaged across all participants for 50 % and 84 % speech intelligibility
299 and all three masker types (fluctuating noise, 1-talker, and 4-talker).

300

301 *[FIGURE 2 about here]*

302

303 An ANOVA on the PPDs revealed an effect of intelligibility ($F[1,18] = 8.85, p = 0.008$) indicated
304 by significant higher PPDs at 50 % compared to 84 % intelligibility. In addition, a main effect for
305 masker type ($F[1,18] = 3.90, p = 0.029$) and an interaction between masker type and intelligibility
306 were found ($F[1,18] = 3.6, p < 0.046$). Post-doc analysis was performed by t-tests (two-tailed paired
307 samples) and revealed higher PPDs at 50 % compared to 84 % intelligibility for the fluctuating
308 noise masker ($p = 0.003$) and the 4-talker masker ($p = 0.005$), but not for the 1-talker masker.
309 Moreover, t-tests performed between masker types at 84 % intelligibility showed larger PPDs for
310 the 1-talker compared to the 4-talker ($p = 0.009$) and the fluctuating noise masker ($p = 0.006$). No
311 differences in the PPDs between the masker types were found at the 50 % speech intelligibility.

312

313 **2.3. DISCUSSION EXPERIMENT 1**

314 Experiment 1 examined the impact of masker type and intelligibility on the SRT and on the
315 listening effort while performing aHINT. Data indicated a main effect of masker type on the SRTs.

316 A lower (better) SRT was measured for the 1-talker masker compared with the two other masker

317 types. Moreover, a lower SRT was measured for the fluctuating noise compared to the 4-talker
318 masker. In other words, the lowest reception thresholds were found for the 1-talker masker, slightly
319 higher thresholds for fluctuating noise, and the highest thresholds for the 4-talker masker
320 independent of the intelligibility. These results are in line with previous studies (e.g., Festen and
321 Plomp, 1990, Koelewijn et al., 2012a; Holube et al., 2011).

322 On basis of the behavioural data, one could argue that those relatively low SRTs observed
323 for the fluctuating noise and 1-talker masker stemmed from an easier differentiation between target
324 and masker signal compared to the 4-talker masker. Both masker types fluctuate in level of which
325 listeners might take advantage by using the temporal minima within the masker signal to detect the
326 relevant speech cues, which is often referred to as *listening-in-the-dips* or *dip-listening* (Miller and
327 Licklider, 1950; Howard-Jones and Rosen, 1993). Amount of the overlapping energy of target and
328 masker increases with increasing number of interfering talkers, which will furthermore reduce the
329 spectro-temporal gaps and, therefore, the possibilities for dip-listening are reduced. Hence,
330 energetic masking is supposed to increase with an increasing number of speakers (e.g. Rosen et al.,
331 2013). This is reflected in the current study by the higher (worse) SRTs for the 4-talker masker
332 compared to the other maskers, which is in line with our hypothesis (H2). Differences in the SRT
333 were further observed between the fluctuating and the 1-talker masker, indicated by lower (better)
334 SRTs for the fluctuating noise. Those differences have been reported before and might result due to
335 semantic interference during the recognition of speech in the presence of an interfering talker (e.g.
336 Koelewijn et al., 2012a).

337 In contrast to what might be expected based on a relatively low SRT in the 1-talker masker
338 condition, the pupil data showed the largest PPD for the 1-talker masker compared with the two
339 other masker types when the presentation level of the masker was relatively low (84 %
340 intelligibility). On the assumption that a higher pupil dilation is indicating higher effort, those

341 results suggest that the participants allocated more resources when the speech was masked by one
342 interfering talker. Increased effort in the presence of one competing talker independent of the SRT
343 has been demonstrated before (Koelewijn et al., 2012a, b, 2014; Ohlenforst et al., 2017) and
344 supports our hypothesis that effort is highest for the 1-talker condition due to highest amount of
345 intelligible interfering linguistic and semantic information (H1). Even though energetic masking is
346 supposed to increase with increasing number of competing talkers, it is assumed that the distinction
347 between target (speech) and noise (4-talker masker) might be facilitated as the background noise
348 becomes less similar to the target signal. At the same time, individual words are less intelligible
349 within the 4-talker babble and, thus, lexical interference might be reduced compared to the 1-talker
350 condition (see e.g. Rosen et al., 2013; Hoen et al., 2007). Our findings support this assumption
351 insofar as the relatively high SRT for the 4-talker babble may indicate the increased energetic
352 masking and reduced opportunity of dip-listening compared to the other maskers. At the same time
353 PPDs were significantly reduced in the 4-talker masker compared to the 1-talker masker which
354 might stem from reduced linguistic interference of the auditory masker. Furthermore, there was no
355 difference found between the PPD measured in the fluctuating noise and the 4-talker babble at 84 %
356 intelligibility, which is also in line with H1.

357 When the presentation level of the maskers was relatively high (50 % intelligibility), all
358 maskers showed larger PPD compared to the 84 % intelligibility and no differences in the PPDs
359 between the masker types were observed anymore. That is, PPDs were similar for all masker types
360 at the 50 % intelligibility, i.e. in a situation where behavioural data (SRTs) differed dramatically
361 and are not predicting the PPD at all. Note that this is in contrast to what has been reported by
362 Koelewijn et al. (2012a). In their study, the authors observed a pronounced effect of the 1-talker
363 masker on the PPD also at 50 % intelligibility. In general, the results of Experiment 1, in particular

364 the behavioural data and the PPD for the 4-talker masker, emphasize the dissociation between
365 performance and listening effort.

366 *Motivation for Experiment 2*

367 The pupillary response has been commonly measured using speech-in-noise tests by adapting the
368 SNR to examine the listening effort at a controlled speech intelligibility (e.g. 50 % or 84 % correct
369 sentence recognition; Zekveld et al., 2010; Koelewijn et al., 2012a). This adaptive procedure has
370 been applied in Experiment 1 as well. As a consequence, comparisons between the PPDs of the
371 different masker types were drawn at varying SNRs. For instance, the SNRs between the 1-talker
372 and the 4-talker masker differed up to 13 dB at the 50 % intelligibility (approximate -13 dB SNR
373 for the 1-talker vs 0 dB SNR for the 4-talker masker). Recent literature, however, reported that
374 listening effort strongly depends on the SNR (e.g. Ohlenforst et al., 2017; Wu et al., 2016). Those
375 studies indicated that effort is changing across SNRs with highest effort at approximately 50 %
376 correct speech intelligibility. Differences in the PPDs observed in Experiment 1 between the 1-
377 talker and the 4-talker masker might have been occurred due to different masker types, but those
378 effects could also partly stem from differences in the SNR. Hence, distinguishing between the effect
379 of SNR and masker type is not feasible when examining the PPD within an adaptive procedure of
380 varying SNRs to achieve a fixed speech intelligibility as realised in Experiment 1.

381 Experiment 2 aimed to gain insides into the effect of SNR on listening effort under more ecological
382 test conditions. With that goal in mind, two changes were made in the paradigm in Experiment 2.
383 First, pupillary responses were measured at fixed SNRs ranging from -20 dB to 8 dB SNR to cover
384 a broad range of listening situations (and intelligibility scores between 0 – 100 %). Second, stimuli
385 were presented over spatially arranged loudspeakers instead of headphones. This was realized to
386 examine listening effort within a more realistic acoustical setting where spatial cues can be utilized
387 to distinguish between different sources such as interaural time and level differences. Even though

388 aided listening were not tested in the current study, measuring the pupillary response within a
389 spatial arrangement of loudspeakers is an important step towards testing listeners using hearing-
390 aids. The main focus of Experiment 2 was to investigate the impact of SNR on listening effort, thus
391 pupillary response was measured across eight different SNRs ranging from 0% to 100% correct
392 recognition. The pupillary response was measured with two different masker types, i.e. a 4-talker
393 masker (same as in Experiment 1, but spatially separated) and a stationary noise masker without
394 temporal fluctuations to maximize the masking.

395 **3. EXPERIMENT 2**

396 **3.1. MATERIALS AND METHODS**

397 **3.1.1. Participants**

398 Twenty-nine listeners (aged from 50 to 77 years, mean 65.7 years, 9 males) with normal hearing
399 participated in Experiment 2. The listeners were native Danish speakers and had average pure tone
400 hearing thresholds of 25 dB hearing level (HL) or better for octave frequencies between 125 Hz - 4
401 kHz for both ears. Furthermore, the accepted thresholds at 6 kHz were 25 to 55 dB (HL) or better
402 and 25 to 60 dB (HL) or better at 8 kHz, depending on the age of the participants (ISO standard
403 7029:2017). The participants had no history of eye diseases or eye operations. Ethical approval for
404 the study was obtained from the Research Ethics Committees of the Capital Region of Denmark.

405 **3.1.2. Stimuli and procedure**

406 Danish male sentences from the HINT corpus were presented with two different masker types, i.e.
407 either with a 4-talker masker or a stationary noise masker within a spatial setup of five
408 loudspeakers. The HINT sentences were presented from a loudspeaker positioned in front of the
409 listener at 0°. The other four peripheral loudspeakers, positioned at $\pm 90^\circ$ and $\pm 150^\circ$ with a
410 distance of 1.2 m to the listener's side or back, were presenting the maskers (see Figure 3). The 4-
411 talker masker was realized by presenting four single talkers, including two male and two female

412 voices, reading a text passage from a newspaper (same as in Experiment 1). Each single-talker was
413 spatially presented via one of the four peripheral loudspeakers in a randomized order, whereby the
414 position of a single-talker with the same gender was balanced across all conditions. Uncorrelated
415 stationary noise was presented through all 4 peripheral loudspeakers as well. Both the 4-talker
416 masker and the stationary noise masker had the same long-term-average spectrum as the HINT
417 sentences. Per masker type, sentences were presented at eight different SNRs ranging between -20
418 dB and +8 dB, distributed in steps of 4 dB SNR. Note that the goal of this experiment was to cover
419 the whole psychometric function including 0% and 100% correct speech recognition and, thus,
420 pupillary response was measured across a broad range of SNRs including extreme listening
421 situations corresponding to 0% correct recognition. The sound level of the masker was kept
422 constant at 70 dB SPL and the level of the speech was changed according to the SNR condition.
423 The masker levels were kept constant to ensure that the noise would not become too loud at the low
424 SNRs. In addition, changing the noise levels might allow the participants to make assumptions
425 about the upcoming task difficulty. All 16 conditions (eight SNRs vs two masker types) were
426 presented in a block design. Each block contained 25 trials leading to 400 trials per participants.
427 Within each trial, the noise started always 3 s before the sentences onset and ended 3 s after
428 sentence offset. Participants were instructed to repeat back the sentence when the noise stopped.
429 Participants performed two training blocks for each condition consisting of 20 trials to get
430 familiarized with the testing setup and the procedure. The complete measurement took about 5
431 hours and was divided into two testing sessions. Within one session, participants performed all
432 eight SNR conditions of one masker type. Half of the participants started with the 4-talker masker,
433 the other half with the stationary noise masker.

434

435

[FIGURE 3 about here]

436

437

3.1.3. Apparatus

438 The same eye-tracker was used as in Experiment 1 and the presentation of the stimuli was
439 controlled by a PC using MATLAB (MathWorks, Natick, MA) based programming. Auditory
440 signals were routed through the same soundcard as in Experiment 1 but this time were played back
441 via loudspeakers Genelec 8040A (Genelec Oy, Iisalmi, Finland). The experiment was conducted in
442 a double-walled, acoustics-treated IAC-NORDIC (IAC Acoustics, Hvidovre, Denmark) sound
443 booth. Luminance in the booth was kept constant at around 135 lux throughout both testing
444 sessions. For participant with relatively large pupils, the luminance was adapted. The participants
445 were seated 60 cm from the eye-tracker.

446

447

3.1.4. Pupil Data Processing

448 For each participant 400 traces were recorded (25 per condition). Pupil data of 29 participants were
449 analysed. Trials that consisted for more than 25 % of their duration of eye-blinks, gross artefacts or
450 missing data were excluded from the further analysis. Due to missing data, two participants were
451 excluded from the further data analysis. Data smoothing, baseline-correction and time-alignment of
452 the data were then performed according to the procedure described in Experiment 1. The PPD was
453 calculated for each participant and each condition.

454

455

3.2 DATA ANALYSIS

456

3.2.1. Linear Mixed Model

457 Linear mixed models (LMM) were chosen to analyze the performance data and the PPDs. A linear
458 mixed-effects model was built in R-studio (version 1.0.153 with programming language R for
459 Windows version 3.3.3) by using the package lme4 (Bates et al., 2014). The function lmer was

460 applied to fit LMM to the data. Two different 2-way LMM ANOVAs were performed for statistical
461 comparison of the effect of SNR and masker type, one for the behavioral performance data (%
462 correct performance) and the other for the pupil data. In both models, SNRs were treated as
463 dependent measures, thus as fixed factors, with participants as the repeated measure and, therefore,
464 as a random factor.

465 **3.2.2. Growth Curve Analysis**

466 In experiments on listening effort using the pupillary response, the PPD and/or mean pupil dilation
467 within pre-defined time segments are commonly analysed. However, some limitations have been
468 pointed out with an approach that does explore potential effects by analysing the pupillary response
469 at a particular point in time or for a time-averaged response (Mirman, 2014). As a consequence,
470 recent studies used statistical models to examine the *morphology of the pupillary response* by
471 modelling pupil dilation as a function of time (Winn et al., 2015; Kuchinsky et al., 2013). To
472 account for effects reflected in the time-course of the pupillary response, aforementioned studies
473 applied a Growth Curve Analysis (GCA) as proposed by Mirman (2008). GCA is a multi-level
474 regression technique that fits orthogonal polynomials to time course data in order to analyse time-
475 depended differences between conditions and between individual participants.

476 In the current study, a third-order (cubic) orthogonal polynomial was applied with fixed effects of
477 SNR. Additionally, all the polynomial terms were included in the model as a random term in order
478 to represent the distributed variance at the individual level. The model was applied to the overall
479 time course of the pupil dilation within a time window starting at 2 s until 7 s of stimulus
480 presentation. The model used a linear combination of three orthogonal polynomials including linear,
481 quadratic and cubic components. The intercept term is supposed to reflect the average height of the
482 curve, linear term refers the overall angle or slope of the curve, and the quadratic term reflects the
483 symmetric rise and fall rate around a primary inflection point (shape of the primary inflection). The

484 cubic term reflects (asymmetric) differences in the rise and fall and, thus, in the steepness of the
485 curve around inflection points (see Mirman, 2008). Higher-order components were not included in
486 the analysis due to ambiguity in their interpretation as well as due to the fact that they led to an
487 overfitting of the pupil curve (see Książek, 2017). The lme4 package (Bates et al., 2015) was used
488 in R for the GCA computations. The model was applied twice to investigate the effect of SNR on
489 the pupillary response, i.e. once for the stationary noise and another time for the 4-talker masker.
490 The model formula and output can be seen in Table 1 for the stationary masker and Table 3 for the
491 4-talker masker. Model output was fitted to the data with four different conditions as a reference for
492 a direct statistical comparison. It means that the model fit (AIC, BIC, LogLik) was kept at the same
493 level, yet the fixed effects were printed in a different order with respect to the condition chosen to
494 be a reference.

495

496 **3.3 RESULTS**

497

498 *[FIGURE 4 about here]*

499

500 **3.3.1. Speech Recognition Performance**

501 Figure 4 shows the performance data, i.e. the averaged recognition scores for the HINT sentences,
502 averaged across all participants for both masker types as a function of SNR. Participants achieved
503 high recognition performance (100 % correct) at the SNRs between +4 and +8 dB SNR. With
504 decreasing SNR (0 dB to -4 dB), recognition dropped rapidly until the participants were able to
505 perform approximately around 5-7 % correct at -12 dB SNR. At -16 and -20 dB SNR, participants'
506 sentence recognition was impossible and performance dropped to 0 % for both masker types. The
507 LMM ANOVA revealed a significant main effect of SNR ($F = 892.0, p < 0.001$) and a small but

508 significant interaction of SNR and masker type ($F = 2.3$, $p = 0.021$). No effect of masker type was
509 found ($p = 0.091$). Post-hoc pairwise t-tests were performed to examine the effect of masker type on
510 the recognition performance between 8 different SNRs (Bonferroni corrected $p = 0.006$ for pairwise
511 t-tests). Significant differences between the two masker types were only revealed at 4 dB SNR ($p =$
512 0.004), indicating a lower recognition performance for the 4-talker masker.

513

514 3.3.2 Pupil Data

515 *Linear Mixed Model*

516 Figure 5 depicts the PPD for the stationary noise masker and Figure 6 shows the PPD for the 4-
517 talker masker as a function of SNRs. For both masker types, the PPD converged to small values at
518 SNRs between 4 and 8 dB corresponding to high performance that almost reach 100 % speech
519 intelligibility. With decreasing SNR, PPD gradually increased and reached maximum PPDs
520 between -4 dB and -8 dB SNR. The corresponding sentence recognition was at approximately
521 between 30 % (at -8 dB SNR) and 70 % (at -4 dB SNR) correct performance. With SNR
522 decreasing below -8 dB SNR, the PPDs dropped again successively and reached a minimum at -20
523 dB SNR corresponding to 0 % speech recognition.

524 [FIGURE 5 about here]

525

526 [FIGURE 6 about here]

527

528 A 2-way LMM ANOVA was tested including the SNR and the masker type as fixed factors on the
529 PPD. A significant main effect of the SNR ($F = 25.9$, $p < 0.001$) and a significant main effect of the
530 masker type ($F = 6.7$, $p < 0.01$) were found. However, no interaction between SNR and masker type
531 was revealed ($p = 0.9$). Pairwise t-tests were performed on the PPD between adjoining SNRs

532 (Bonferroni corrected $p = 0.006$ for pairwise t-tests). For the stationary noise masker, significant
533 differences in the PPDs were revealed between -16 and -12 dB SNR ($p = 0.004$), between -12 and -
534 8 dB SNR ($p = 0.001$), between -4 and 0 dB SNR ($p = 0.001$), and between 0 and 4 dB SNR ($p =$
535 0.003). Note that there were no differences between -20 and -16 dB SNR ($p = 0.154$), between -8
536 and -4 dB SNR ($p = 0.570$), and between 4 and 8 dB SNR ($p = 0.797$).

537 For the 4-talker masker, significant differences were found between -12 and -8 dB SNR ($p < 0.001$)
538 and between -4 and 0 dB SNR ($p = 0.001$). Note that no differences between -20 and -16 dB SNR
539 ($p = 0.141$), between -16 and -12 dB SNR ($p = 0.223$), between -8 and -4 dB SNR ($p = 0.664$), and
540 between 4 and 8 dB SNR ($p = 0.750$) were found.

541

542 *[FIGURE 7 about here]*

543

544 *[FIGURE 8 about here]*

545

546

547 *Growth Curve Analysis*

548 The pupil curves for both masker types are depicted in Figure 7 (stationary noise) and Figure 8 (4-
549 talker). Two different analysis were carried out, one for each masker type. A first analysis was
550 carried for the stationary masker. The model formula, the model fit, and the output for the GCA are
551 presented in Table 1. GCA demonstrated a main effect of SNR on all terms depending on the
552 reference condition (see Table 1), which will be discussed more detailed in the following. A
553 significant effect was found for the intercept ($p < 0.05$) and the linear term ($p < 0.001$) for all
554 reference conditions. Furthermore, a significant effect of SNR on the pupillary response was

555 revealed on the quadratic term ($p < 0.001$) for three reference conditions (-12, -4 and 4dB SNR as
556 reference) and on the cubic term ($p < 0.001$) for two reference conditions (-12 and -4 dB SNR as
557 reference). Summing up, a GCA demonstrated that the overall height (intercept) and slope of the
558 pupil dilation (linear term) changed with SNRs for the stationary masker. There was also a
559 significant effect on the quadratic and cubic term for three reference conditions, indicating changes
560 in the symmetric rise and fall rate around a central inflection point (quadratic) and a more delayed
561 peak of the response (cubic term) for more unfavorable SNRs.

562 Planned comparisons were made between some SNRs where the PPDs were similar or did not differ
563 significantly (see previous section about the analysis of the PPD data) by using the GCA model (see
564 Table 2). For the stationary noise masker, the GCA revealed significant differences in the overall
565 time course of pupil dilation between -12 and 4 dB SNR in the intercept term ($p < 0.001$), the
566 quadratic term ($p < 0.001$), and the cubic term ($p = 0.013$), indicating a higher overall pupil dilation,
567 a higher acceleration/deceleration around the central inflection point, and a more delayed peak of
568 the response for the -12 SNR condition. Furthermore, effects on the intercept and the linear term
569 were identified between 4 and 8dB SNR, pointing towards a higher average pupil response and a
570 higher slope of the entire pupil response at 4 dB SNR (see Table 2).

571

572

573 *[Table 1 about here]*

574

575 *[Table 2 about here]*

576 A separate GCA analysis was carried out for the 4-talker masker (Table 3). A significant effect of
577 SNR was found for the quadratic term independent of the reference conditions ($p < 0.05$) suggesting
578 that –similar to the stationary noise- the rise and fall rate around a central inflection point changed
579 with decreasing SNRs. In addition, a significant effect of SNR on the cubic term ($p < 0.001$)

580 indicated that the delay of the peak response changed across SNRs (for conditions with -12, -4 and
581 4dB SNR as reference). Moreover, a significant effect of SNR on the overall slope (linear term) and
582 on the average height (intercept) of the pupil dilation was revealed for -4 and -12 dB SNR as
583 reference conditions ($p < 0.001$). Similar as for the stationary masker, planned comparisons were
584 performed between some SNR conditions for the 4-talker masker with similar PPDs (cf. Table 4).
585 For the 4-talker masker, the GCA revealed significant differences between -12 and 4dB SNR in the
586 overall height of the pupil dilation (intercept; $p < 0.001$) and the rise and fall rate around the
587 inflection point (quadratic term; $p < 0.001$). Furthermore, the cubic term significantly differed for
588 the comparison between 4 vs 8dB SNR ($p = 0.049$).

589 *[Table 3 about here]*

590 *[Table 4 about here]*

591

592 **3.4. DISCUSSION EXPERIMENT 2**

593 The results of Experiment 2 indicated a strong impact of the SNR on the PPDs. Highest PPDs were
594 measured for intermediate SNRs corresponding to 30 %-70 % correct recognition. Lowest PPDs
595 were revealed at higher SNRs due to a more favourable listening condition and also at lower SNRs
596 where listening became impossible (as indicated of the recognition performance). Interestingly, the
597 impact of the masker type was rather small and only small differences were found between the
598 PPDs of different maskers at the corresponding SNRs. Note that this is not in line with our
599 hypothesis (H4), which predicted a maximum pupil dilation at lower (negative) SNRs for the 4-
600 talker babble compared to the stationary noise.

601 A GCA was applied on the pupillary response independently for both masker types that revealed
602 further (time-dependent) characteristics of the pupil curve are affected by the SNR. For both masker
603 types, differences in the intercept, linear, cubic, and quadratic term were identified depending on the

604 reference condition. Independent of the reference condition, an impact of SNR on the overall height
605 and the overall slope of the pupillary response occurred for the stationary masker. An effect of SNR
606 on the rise and fall rate around the primary inflection was identified for the 4-talker masker
607 independent of the reference condition. Moreover, selected comparisons between some SNR
608 conditions with similar PPDs identified differences in the overall time course of the pupil dilation,
609 which were not necessarily covered by the PPD analysis. For instance, differences were detected at
610 favourable SNRs, i.e. between 4 and 8 dB SNR, in the overall height and slope of the pupil curve
611 for the stationary noise, and in delay in peak of the response for the 4-talker babble. At both SNRs
612 speech intelligibility was very high (with recognition performance at around 100 %) and no
613 significant differences in the PPDs occurred. Further, differences between -12 and 4 dB SNR were
614 found in the overall size of the pupillary response (for both masker types), the overall slope (for the
615 stationary masker) as well as in the steepness of the primary inflection (for the 4-talker masker).
616 Note that in those two conditions, the PPDs were very similar (in particular for the 4-talker masker),
617 however time-depending changes in the pupil dilation were still detectable between the two SNR
618 conditions where the recognition performances differed dramatically (i.e. below 10% at -12dB and
619 above 90% at 4dB SNR). The results encourage the analysis of the overall time course of the
620 pupillary response.

621

622 **4. GENERAL DISCUSSION**

623 The two experiments of the present study explored the impact of masker type and SNR on the pupil
624 dilation response using the Danish HINT test. Experiment 1 focused on the impact of semantic and
625 linguistic interference on the pupil dilation at fixed speech intelligibility. Experiment 2 examined
626 the pupil dilation as a function of SNR. Both the effect of masker type as well as changes in pupil

627 dilation as a function of SNRs will be discussed in the following. Finally, challenges and some
628 general considerations when combining pupillometry and a speech-in-noise test will be discussed.

629 The results from Experiment 1 showed lowest (best) SRTs for the 1-talker masker, followed by
630 slightly higher SRTs for the fluctuating noise, and highest (worst) SRTs for the 4-talker masker.

631 This was found at 50 % and at 84 % speech intelligibility. These findings are supported by a
632 previous study from Koelewijn et al. (2012a). One could argue that the relatively low SRTs for the
633 1-talker masker originated from a better and easier differentiation between target speech and masker
634 even at similar intelligibility. However, the pupil data showed a larger pupillary response for the 1-
635 talker masker compared with both other masker types, which suggested that more cognitive
636 resources were invested when the noise masker contained intelligible speech information leading to
637 increased listening effort. These results are in line with H1 that predicted an effect of semantic
638 interference of the masker on the pupil response. Interestingly, this impact of maskers containing
639 speech information was only found at 84 % speech intelligibility, which is not in line with
640 Koelewijn et al. (2012a). Furthermore, it was hypothesized that the effect of the energetic masking
641 should be most pronounced with the 4-talker babble (H2). Again, that was in line with the findings
642 of Experiment 1. Higher (worse) SRTs were found within the 4-talker masker condition due to an
643 increased energetic masking. Our results indicate a distinction between the impacts of informational
644 vs energetic masking. Whereas the effect of semantic or linguistic content of the masker was
645 highest on the PPD and thus on listening effort, the effect of energetic masking was most
646 pronounced for the SRT data and less for the pupil data. Furthermore, the distinction between SRT
647 and PPD data further supports the assumption that performance and listening effort are not always
648 related to each other, which is supported by previous studies (Koelewijn et al., 2012a; Mc Garrigle
649 et al., 2014; Wendt et al., 2017). Note that comparisons of PPDs were drawn between listening
650 situations that highly differ in the SNRs, especially when comparing the 4-talker masker to the other

651 masker conditions. SRTs measured for different masker types differed by almost 13 dB SNR.
652 Literature indicate that the pupillary response can be further affected by the SNRs (Zekveld and
653 Kramer, 2014; Ohlenforst et al., 2017). Hence, a differentiation between the effect of SNR and
654 masker type would not be possible based on the findings in Experiment 1.

655 Experiment 2 was conducted with the primary goal of exploring the impact of SNR on the pupillary
656 response. The results of Experiment 2 suggested that the PPD changed non-linearly across SNRs in
657 the form of an inverted U-shape: PPD were highest at intermediate SNRs between -8 and -4 dB
658 SNR. With increasing SNRs, the PPDs decreased due to gradually decreasing task demands and
659 listening became easier due to a more favourable SNR. In addition, PPDs demonstrated that the
660 highest effort was reached when speech intelligibility was between 30 % and 70 % correct
661 recognition, which is in line with our hypothesis (H3) and previous studies. Zekveld and Kramer
662 (2014) assessed pupil dilation within a speech-in-noise test across a wide range of intelligibility
663 between 0 % to 99 % correct recognition. The authors reported that pupil dilation was largest at
664 intermediate intelligibility. Recently, Ohlenforst et al. (2017) investigated changes in the pupillary
665 response across a range of SNRs for people with normally hearing and with hearing impairment. It
666 was demonstrated that, again, the PPDs showed a peak at around 40 % -50 % correct speech
667 recognition in both stationary noise and in a 1-talker masker. This non-linear trend of listening
668 effort across a broad range of SNRs had been reported by applying other methods and techniques.
669 For instance, Wu et al. (2016) investigated listening effort employing a dual-task paradigm using
670 primary speech recognition task simultaneously with a secondary visual task. Reaction times were
671 measured within the secondary task as an indicator of the listening effort involved in speech
672 recognition. Results indicated that the reaction times changed in form of an inverted U-shape across
673 SNRs, with a maximum reaction time at intermediate SNRs corresponding to intelligibility between
674 30 % - 50 % correct recognition.

675 Those findings of a non-linear trend of the listening effort as a function of increasing task demands
676 are supporting the FUEL (Pichora-Fuller et al., 2016). The framework is assuming that listening
677 effort involved in speech understanding in noise is mainly modulated by two dimensions: the task-
678 demand dimension and the motivation dimension. Both can be integrative or independently
679 affecting the cognitive resources that are allocated and, thus, the listening effort within a listening
680 task. The demands mainly depend on external factors that are entailed with the input (such as a
681 degraded signal due to noise, but also due to a hearing loss) or the task (e.g. instructions or
682 complexity of the task). The motivation dimension is more internally controlled and depends on the
683 individual's criterion for the importance of success (Pichora-Fuller et al., 2016). In line with FUEL,
684 effort would increase due to changes in the task demands such as with decreasing SNRs from
685 favourable (e.g. between 4 and 0 dB SNR) to intermediate SNRs (between -4 and -8 dB SNR). Our
686 data indicated that participants are willing to spend increased effort with decreasing and more
687 unfavourable SNRs covering an intelligibility range between 30 % - 70 %. However, with further
688 decreasing SNRs, task demands escalated causing a drop in effort, which is probably due to a drop
689 in the motivation. In extreme listening situations, high task demands, as they would be imposed by
690 low (unfavourable) SNRs, could lead to signs of "quitting" or "giving up" (Pichora-Fuller et al.,
691 2016). In other words, understanding and recognizing speech becomes impossible at very low
692 SNRs. This can lead to disengagement which is probably demonstrated in Experiment 2 at -20 or -
693 16 dB SNR. Note that this breaking point would be undetectable by examining the performance
694 data only. Even though the performance was gradually dropping with decreasing SNRs, the point
695 where the effort was peaking – and would drop with further increasing task demands - would not be
696 reflected in the behavioural data. It was furthermore speculated that the fact that the largest
697 pupillary response was observed in ranges around 50% performance levels, might actually suggest

698 that effort peaks in difficulty ranges where listeners could actively change their own performance
699 level within the speech-in-noise test by exerting more effort.

700

701 *Considerations when testing listening effort for speech recognition in noise*

702 Two experiments were presented in the current study, both applying pupillometry together with a
703 speech-in-noise test that is commonly used for speech audiometry. The results suggested that
704 listening effort changed depending on the masker type as well as on the SNR. While the PPD was
705 larger for maskers containing speech, the PPD was further affected by the SNR and seemed to peak
706 between 30-70 % speech intelligibility. This maximum in PPD was referred to as the breaking point
707 since it might indicate the point where –with further increasing demands- effort would drop due to
708 too high task demands and/or dropping motivation. This can be a problem when examining the PPD
709 at a fixed speech intelligibility, since small changes in the SNRs might have a huge impact on the
710 PPDs when testing at or around this breaking point. Furthermore, it would be difficult to evaluate
711 whether smaller PPDs would indicate reduced effort due to lower task demands or whether smaller
712 PPDs point towards reduced effort as a consequence of giving-up. Hence, when investigating the
713 PPDs at fixed intelligibility, changes in PPDs around the 50 % performance level should be
714 interpreted with caution when the SNRs differ between conditions. Furthermore, when investigating
715 an impact of the masker type on effort by means of PPDs, it can be advisable not only to analyse the
716 PPD or the mean pupil dilation, but also to characterize the whole pupil curve. Several approaches
717 have been realized to model changes in pupil dilation over time (Kuchinsky et al., 2013; Mirman,
718 2014; Winn et al., 2015). Our results indicated that by applying the GCA, time-dependent
719 differences in the pupillary response can be detected that were not necessarily reflected in the PPD
720 alone. Furthermore, time-dependent differences in the overall slope of the pupil curve were
721 identified in situations where speech recognition performances were at ceiling and around 100 %

722 performance. This is in line with previous studies that reported of applying GCA to gain a more
723 sensitive pupillary analysis (Kuchinsky et al., 2013, Winn et al., 2015). Even though a GCA has
724 been successfully applied in recent studies, further work is required to gain a better understanding
725 about how terms of the GCA are related to different aspects of listening effort while performing a
726 speech-in-noise test. In other words, there is still a lack of knowledge in interpreting time-
727 dependent differences in a shape of the pupillary curve in terms of effort involved in speech
728 recognition. Other methods have been proposed to investigate time-dependent changes of the pupil
729 dilation. For instance, Winn (2016) analysed both the growth of the pupillary response (dilation)
730 and the reduction of this response (constriction), and pointed to the later as an indicator of the
731 release from effort. He suggested that analysing the constriction of the pupillary response after the
732 presentation of the stimulus can reveal difficulties due to hearing accuracy (cochlear implant user vs
733 normally hearing listener) and spectral degradation (unprocessed speech vs. vocoded speech).

734

735 **5. CONCLUSIONS**

736 Two main observations were replicated in the current study using the HINT test: First, listening
737 effort is highly affected by the masker type. This was reflected by the largest pupil dilation for
738 speech recognition in the presence of a 1-talker masker in comparison to a 4-talker masker and a
739 fluctuating noise masker. Increased effort in the presence of one interfering speaker, however, was
740 not reflected by the SRTs at fixed intelligibility. Second, listening effort changes with SNR.
741 Pupillary response changed non-linearly across a range of fixed SNRs that corresponded to a wide
742 range of speech recognition between 0 % to 100 % correct performances. Pupil dilation was largest
743 for intermediate SNRs. This point indicated maximum effort and was interpreted as a breaking
744 point since effort would drop with decreasing SNRs. Hence, it was suggested that by means of pupil

745 dilation, effects of motivation or disengagement on listening effort, that were not necessarily
746 reflected by performance measures, were captured during the speech-in-noise test.

747 Our findings led to two main conclusions that should be considered in light of the experimental
748 methods and data analysis when testing pupillary response within the speech-in-noise test: First,
749 listening effort changes non-monotonically as a function of SNR with highest effort around 50%
750 performance. Thus, changes in the pupillary response at SRTs around 50 % level of performance
751 should be interpreted with caution since they might indicate either lower effort due to reduced task
752 demands or lower effort due to disengagement or giving up. Second, when assessing changes in the
753 pupillary response over time, difficulties arising during speech recognition in noise can be detected
754 that are not necessarily covered by the PPD alone. In general, our data support the assumption that
755 by collecting pupil responses using a (traditional) speech-in-noise test, a more complete picture of
756 difficulties arising from speech recognition in different types of background sound can be obtained.

757

758 **6. Acknowledgments**

759 This research was funded by the Oticon Foundation (grant 14-0845). The authors would like to
760 thank Hans van Beek, Sanne Mehrfeld Møller and Josefine Juul Jensen for their support with the
761 preparation and data collection of the experiments. Furthermore, we thank Renskje K. Hietkamp for
762 her support with the participant recruitment and planning of both experiments.

763 **References**

- 764 Aston-Jones, G., Cohen, J. D., 2005. An integrative theory of locus coeruleus-norepinephrine
765 function: Adaptive gain and optimal performance. *Annual Review of Neuroscience* 28, 403–450.
766
- 767 Bates, B., Mächler, M., Bolker, B. M., Walker, S. C., 2015. Fitting Linear Mixed-Effects Models
768 using lme4. *Journal of Statistical Software* 67, 1-48.
769
- 770 Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing
771 resources. *Psychological Bulletin* 91, 276–292.
772
- 773 Brungart, D. S., Simpson, B. D., Ericson, M. A., et al. 2001. Informational and energetic masking
774 effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of*
775 *America* 110, 2527–2538.
776
- 777 Festen, J. M., Plomp, R., 1990. Effects of fluctuating noise and interfering speech on the speech-
778 reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America*
779 88, 1725–1736.
780
- 781 Grady, C., 2012. The cognitive neuroscience of ageing. *Nature Reviews Neuroscience* 13, 491–505.
782
- 783 Hagerman, B., 1982. Sentences for testing speech intelligibility in noise. *Scandinavian Audiology*
784 11(2): 79–87.
785
- 786 Hoen, M., Meunier, F., Grataloup, C. L., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., and
787 Collet, L., 2007. Phonetic and lexical interferences in informational masking during speech-in-
788 speech comprehension. *Speech Communication*. 49, 905–916.
789
- 790 Holube, I. (2011) Speech intelligibility in fluctuating maskers. *Proceedings of the International*
791 *Symposium on Auditory and Audiological Research*, [S.I.], v. 3, p. 57-64.
792
- 793 Howard-Jones, P. A., Rosen, S., 1993. The perception of speech in fluctuating noise. *Acustica* 78,
794 258–272.
795
- 796 Hygge, S., Ronnberg, J., Larsby, B., et al., 1992. Normal-hearing and hearing- impaired subjects’
797 ability to just follow conversation in competing speech, reversed speech, and noise backgrounds.
798 *Journal of Speech and Hearing Research* 35, 208–215.
799
- 800 ISO 7029:2017. Acoustics – Statistical Distribution of Hearing Thresholds related to Age and
801 Gender.
802

- 803 Janisse, M. P., 1977. *Pupillometry: The Psychology of the Pupillary Response*. Washington, DC:
804 Hemisphere Publishing Corporation.
- 805
- 806 Kahneman, D., Beatty, J., 1966. Pupil diameter and load on memory. *Science* 154, 1583–1585.
- 807
- 808 Kidd, Jr G., Mason, C. R., Richards, V. M., Gallun, F. J., Durlach, N. I., 2008. Informational
809 masking. *Auditory perception of sound sources*, (Springer), pp 143-189.
- 810
- 811 Koelewijn, T., Zekveld, A.A., Festen, J.M., Kramer, S.E., 2012a. Pupil dilation uncovers extra
812 listening effort in the presence of a single-talker masker. *Ear and Hearing* 33, 291.
- 813
- 814 Koelewijn, T., Zekveld, A.A., Festen, J.M., Rönnerberg, J., Kramer, S.E., 2012b. Processing Load
815 Induced by Informational Masking Is Related to Linguistic Abilities. *International Journal of*
816 *Otolaryngology* 2012, 1–11. doi:10.1016/j.specom.2005.09.004
- 817
- 818 Koelewijn, T., Zekveld, A. A., Festen, J. M. & Kramer, S. E., 2014. The influence of informational
819 masking on speech perception and pupil response in adults with hearing impairment. *Journal of the*
820 *Acoustical Society of America*, 135 (3), 1596–1606.
- 821
- 822 Kramer, S. E., Kapteyn, T. S., Houtgast, T., 2006. Occupational performance: Comparing normally-
823 hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work.
824 *International Journal of Audiology*, 45, 503–512.
- 825
- 826 Książek, P., (2017). *Statistical Modelling of Pupil Curves to Quantify Differences in Processing*
827 *Effort on Group- and Individual- Level*. Master Thesis. Technical University of Denmark (DTU),
828 Hearing Systems, Department of Electrical Engineering,
- 829
- 830 Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., Eckert, M.
831 A., 2013. Pupil size varies with word listening and response selection difficulty in older adults with
832 hearing loss. *Psychophysiology*, 50, 23–34.
- 833
- 834 Lunner, T., Rudner, M., Rosenbom, T., et al. 2016. Using speech recall in hearing aid fitting and
835 outcome evaluation under ecological test conditions. *Ear and hearing*, 37(Suppl 1), 145S–154S.
- 836
- 837 McGarrigle, R., Munro, K., Dawes, P., Stewart, A., Moore, Barry, J. G., Amitay, S., 2014.
838 Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology
839 Cognition in Hearing Special Interest Group “white paper”. *International Journal of Audiology*,
840 53(7), 433–440. doi:0.3109/14992027.2014.890296.
- 841
- 842 McMahon CM, Boisvert I, de Lissa P, Granger L, Ibrahim R, Lo CY, Miles K and Graham PL (2016)
843 Monitoring Alpha Oscillations and Pupil Dilation across a Performance-Intensity Function. *Front.*
844 *Psychol.* 7:745. doi: 10.3389/fpsyg.2016.00745

- 845
846 Miller, G. A., Licklider, J. C. R., 1950. The intelligibility of interrupted speech. *Journal of the*
847 *Acoustical Society of America* 22, 167–173.
848
- 849 Mirman, D., 2014. *Growth curve analysis and visualization using R*. New York: CRC Press.
850
- 851 Mirman, D., Dixon, J. A., Magnuson, J. S., 2008. Statistical and computational models of the visual
852 world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59,
853 475–494.
854
- 855 Nielsen, J. B., Dau, T., 2011. The Danish hearing in noise test. *International Journal of Audiology*
856 50, 202–208.
857
- 858 Ohlenforst, B., Zekveld, A.A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N.J.,
859 Kramer, S.E., 2017. Impact of stimulus-related factors and hearing impairment on listening effort as
860 indicated by pupil dilation. *Hearing Research* 351, 68–79. doi:10.1016/j.heares.2017.05.012
861
- 862 Petersen, E. B., Wöstmann, M., Obleser, J., Stenfelt, S., Lunner T., 2015. Hearing loss impacts
863 neural alpha oscillations under adverse listening conditions. *Frontiers in Psychology*. 6:177.
864 doi:10.3389/fpsyg.2015.00177
865
- 866 Pichora-Fuller, M.K., Kramer, S.E., Eckert, M.A., Edwards, B., Hornsby, B.W., Humes, L.E.,
867 Lemke, U., Lunner, T., Matthen, M., Mackersie, C.L., 2016. Hearing impairment and cognitive
868 energy: The framework for understanding effortful listening (FUEL). *Ear and hearing* 37, 5S-27S.
869
- 870 Plomp, R., Mimpen, A. M., 1979. Improving the reliability of testing the speech reception threshold
871 for sentences. *Audiology*, 18, 43–52.
872
- 873 Pollack, I., 1975. Auditory informational masking. *Journal of the Acoustical Society of America*, 57
874 (Suppl. 1), 5.
875
- 876 Reuter-Lorenz, P. A., Cappell, K. A., 2008. Neuro cognitive aging and the compensation
877 hypothesis. *Current Directions in Psychological Science*. 17, 177–182. doi: 10.1111/j.1467-
878 8721.2008.00570.x
879
- 880 Rosen, S., Souza, P., Ekelund, C., Majeed, A.A., 2013. Listening to speech in a background of other
881 talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of*
882 *America* 133, 2431 (2013); doi: 10.1121/1.4794379
883
- 884 Rönnerberg, J., Lunner, T., Zekveld, A.A., Sörqvist, P., Danielsson, H., Lyxell, B., Hahlström, Ö.,
885 Signoret, C., Stenfelt, S., Pichora-Fuller, K.M., Rudner, M., Rudner, M., 2013. The Ease of

- 886 Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in*
887 *Systems Neuroscience* 1–17. doi:10.3389/fnsys.2013.00031/abstract
- 888
- 889 Sander, M.C., Werkle-Bergner, M., and Lindenberger, U., 2012. Amplitude modulations and inter-
890 trial phase stability of alpha-oscillations differentially reflect working memory constraints across
891 the life span. *Neuroimage* 59, 646–654. doi: 10.1016/j.neuroimage.2011.06.092
- 892
- 893 Simpson, S.A., Cooke, M., 2005. Consonant identification in N-talker babble is a nonmonotonic
894 function of N. *The Journal of the Acoustical Society of America* 118, 2775–2778.
895 doi:10.1121/1.2062650
- 896
- 897 Smeds, K., Wolters, F., Rung, M., (2015). Estimation of signal-to-noise ratios in realistic sound
898 scenarios. *Journal of the American Academy of Audiology* 26, 183–196.
- 899
- 900 Wendt, D., Dau, T., Hjortkjær, J. (2016). Impact of Background Noise and Sentence Complexity on
901 Processing Demands during Sentence Comprehension. *Frontiers in Psychology* 1429 7(345). doi:
902 10.3389/fpsyg.2016.00345
- 903
- 904 Wendt, D., Hietkamp, R.K., Lunner, T. 2017. Impact of noise and noise reduction on processing
905 effort: A pupillometry study. *Ear and Hearing* 141, 4040-4040.
- 906
- 907 Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution
908 on listening effort revealed by pupil dilation. *Ear and Hearing*, 36, 153–165.
- 909
- 910 Winn, M.B., (2016). Rapid Release From Listening Effort Resulting From Semantic Context, and
911 Effects of Spectral Degradation and Cochlear Implants *Trends in Hearing*, 20: 1–17
- 912
- 913 Wisniewski, M.G., Thompson, E.R., Iyer, N., 2017. Theta- and alpha-power enhancements in the
914 electroencephalogram as an auditory delayed match-to-sample task becomes impossibly difficult.
915 *Psychophysiology*. 00:000–000. [https://doi.org/ 10.1111/psyp.12968](https://doi.org/10.1111/psyp.12968)
- 916
- 917 Wu, Y.-H., Stangl, E., Zhang, X., Perkins, J., Eilers, E. 2016. Psychometric functions of dual-task
918 paradigms for measuring listening effort. *Ear and hearing* 37, 660-670.
- 919
- 920 Zekveld, A., Kramer, S., Festen, J., 2010. Pupil Response as an Indication of Effortful Listening:
921 The Influence of Sentence Intelligibility. *Ear and hearing*.
- 922
- 923 Zekveld, A.A., Kramer, S.E., 2014. Cognitive processing load across a wide range of listening
924 conditions: Insights from pupillometry. *Psychophysiology* 51, 277–284. doi:10.1111/psyp.12151
- 925

926 Zekveld, A. A., Kramer, S. E., Festen, J. M., 2011. Cognitive load during speech perception in
927 noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and hearing* 32,
928 498–510.

ACCEPTED MANUSCRIPT

929 **FIGURE LEGENDS**

930

931

932 **Figure 1:** Mean SRTs averaged across participants measured for three different masker types at
933 two different intelligibility scores (50 % and 84 %). Error bars indicate one standard error from the
934 mean.

935

936 **Figure 2:** PPD in mm per masker type and intelligibility scores averaged across all participants.
937 Error bars indicate one standard error from the mean.

938

939 **Figure 3:** Experimental setup of the loudspeakers. HINT sentences were presented from the
940 loudspeaker in the front (0°). The masker was presented from the loudspeaker at the side and the
941 back of the listener ($\pm 90^\circ$ and $\pm 150^\circ$).

942

943 **Figure 4:** Mean recognition scores (in %) for the stationary noise and the 4-talker masker
944 averaged across all participants. Error bars indicate one standard error from the mean.

945

946 **Figure 5:** Averaged PPDs and word recognition scores across all participants for the stationary
947 noise masker for eight different SNRs. Error bars indicate one standard error from the mean.

948

949 **Figure 6:** Averaged PPDs and word recognition scores across all participants for the 4-talker
950 babble masker for eight different SNRs. Error bars indicate one standard error from the mean.

951

952

953 **Figure 7:** Mean pupil response averaged across participants per SNR in the stationary masker
954 condition. Sentence onset was at 3 s. The baseline value was calculated as the mean pupil value one
955 second preceding the sentence onset (i.e. between 2 s and 3 s).

956

957 **Figure 8:** Mean pupil response averaged across participants per SNR in the 4-talker babble masker.
958 Sentence onset was at 3 s. The baseline value was calculated as the mean pupil value one second
959 preceding the sentence onset (i.e. between 2 s and 3 s).

960

961 **Table 1:** Linear Mixed-Effects Model formula and output of the GCA for the pupil dilation recorded
 962 in conditions with the stationary noise. The effect of SNR on the all terms was tested against 4
 963 different references, i.e. against -20 dB, -12 dB, -4 dB, 4 dB.

Formula code: PupilDilation ~ (1 + Linear + Quadratic + Cubic) * SNR + (1 + Linear + Quadratic + Cubic Subject)												
Model fit: AIC: -34546.3; BIC: -34225.3; LogLik: 17316.2; Deviance: -34632.3; Df. resid: 12857												
Terms	Reference: -20dB			Reference: -12dB			Reference: -4dB			Reference: 4dB		
	β	t	p	β	t	p	B	t	p	β	t	p
Intercept	<u>0.0186</u>	<u>2.743</u>	<u>0.009**</u>	<u>0.038</u>	<u>5.575</u>	<u><0.001**</u>	<u>0.063</u>	<u>9.215</u>	<u><0.001**</u>	<u>0.014</u>	<u>2.064</u>	<u>0.046*</u>
Linear	<u>0.216</u>	<u>3.192</u>	<u>0.002**</u>	<u>0.359</u>	<u>5.325</u>	<u><0.001**</u>	<u>0.633</u>	<u>9.379</u>	<u><0.001**</u>	<u>0.255</u>	<u>3.754</u>	<u><0.001**</u>
Quadratic	-0.118	-1.909	0.062	<u>-0.228</u>	<u>-3.681</u>	<u><0.001**</u>	<u>-0.380</u>	<u>-6.14</u>	<u><0.001**</u>	<u>-0.44</u>	<u>-7.066</u>	<u><0.001**</u>
Cubic	-0.039	-0.905	0.369	<u>-0.142</u>	<u>-3.288</u>	<u>0.002**</u>	<u>-0.204</u>	<u>-4.71</u>	<u><0.001**</u>	-0.0456	-1.045	0.3

964 * p<0.05; ** p<0.01.

965 AIC – Akaike Information Criterion,

966 BIC – Bayesian Information Criterion,

967 LogLik – Logarithmic Likelihood,

968 Deviance- a measure of the goodness of the model fit,

969 Df. Resid – Degree of Freedom for Residual,

970 **Table 2:** Planned contrasts: The effect on the pupil dilation between chosen SNRs in the stationary
 971 noise condition. Contrasts adjusted with the multivariate *t* adjustment.

Contrast	-12dB vs. 4dB			-4dB vs. -8dB			4dB vs. 8dB		
	Term: 4 dB, Reference: -12dB			Term:-8dB, Reference: -4dB			Term:8dB, Reference: 4dB		
Terms	β	t	p	β	t	p	β	t	p
Intercept	<u>-0.024</u>	<u>-6.277</u>	<u><0.001**</u>	0.004	0.947	0.343	<u>-0.010</u>	<u>-2.539</u>	<u>0.011*</u>
Linear	-0.104	-1.982	0.047	0.004	-0.078	0.938	<u>-0.132</u>	<u>-2.506</u>	<u>0.012*</u>
Quadratic	<u>-0.213</u>	<u>-4.212</u>	<u><0.001**</u>	0.081	1.611	0.107	0.051	0.953	0.341
Cubic	<u>0.097</u>	<u>2.474</u>	<u>0.013*</u>	-0.0001	-0.012	0.991	0.004	0.113	0.91

972

973 * $p < 0.05$; ** $p < 0.01$;

974

975 **Table 3:** Linear Mixed-Effects Model formula and output of the GCA for the pupil dilation recorded
 976 in conditions with the 4-talker masker. The effect of SNR on the all terms was tested against 4 differ
 977 ent references, i.e. against -20 dB, -12 dB, -4 dB, 4 dB.
 978

Formula code: PupilDilation ~ (1 + Linear + Quadratic + Cubic) * SNR + (1 + Linear + Quadratic + Cubic Subject)												
Model fit: AIC: -35375.5; BIC: -35054.5; logLik: 17730.8; Deviance: -35461.5; Df. resid: 12857												
Terms	Reference: -20dB			Reference: -12dB			Reference: -4dB			Reference: 4dB		
	β	t	p	β	t	p	β	t	P	β	t	p
Intercept	0.0057	0.776	0.443	<u>0.019</u>	<u>2.534</u>	<u>0.016**</u>	<u>0.069</u>	<u>9.489</u>	<u><0.001**</u>	0.006	0.771	0.446
Linear	0.0462	0.575	0.569	<u>0.163</u>	<u>2.032</u>	<u>0.049*</u>	<u>0.48</u>	<u>6.008</u>	<u><0.001**</u>	0.086	1.068	0.292
Quadratic	<u>-0.143</u>	<u>-2.522</u>	<u>0.014**</u>	<u>-0.242</u>	<u>-4.291</u>	<u><0.001**</u>	<u>-0.152</u>	<u>-9.440</u>	<u>0.009**</u>	<u>-0.377</u>	<u>-6.635</u>	<u><0.001**</u>
Cubic	-0.0594	-1.445	0.153	<u>-0.141</u>	<u>-3.433</u>	<u>0.001**</u>	<u>-0.343</u>	<u>-8.347</u>	<u><0.001**</u>	<u>-0.097</u>	<u>-2.342</u>	<u>0.022*</u>

979 * $p < 0.05$; ** $p < 0.01$.

980 AIC – Akaike Information Criterion,

981 BIC – Bayesian Information Criterion,

982 LogLik – Logarithmic Likelihood,

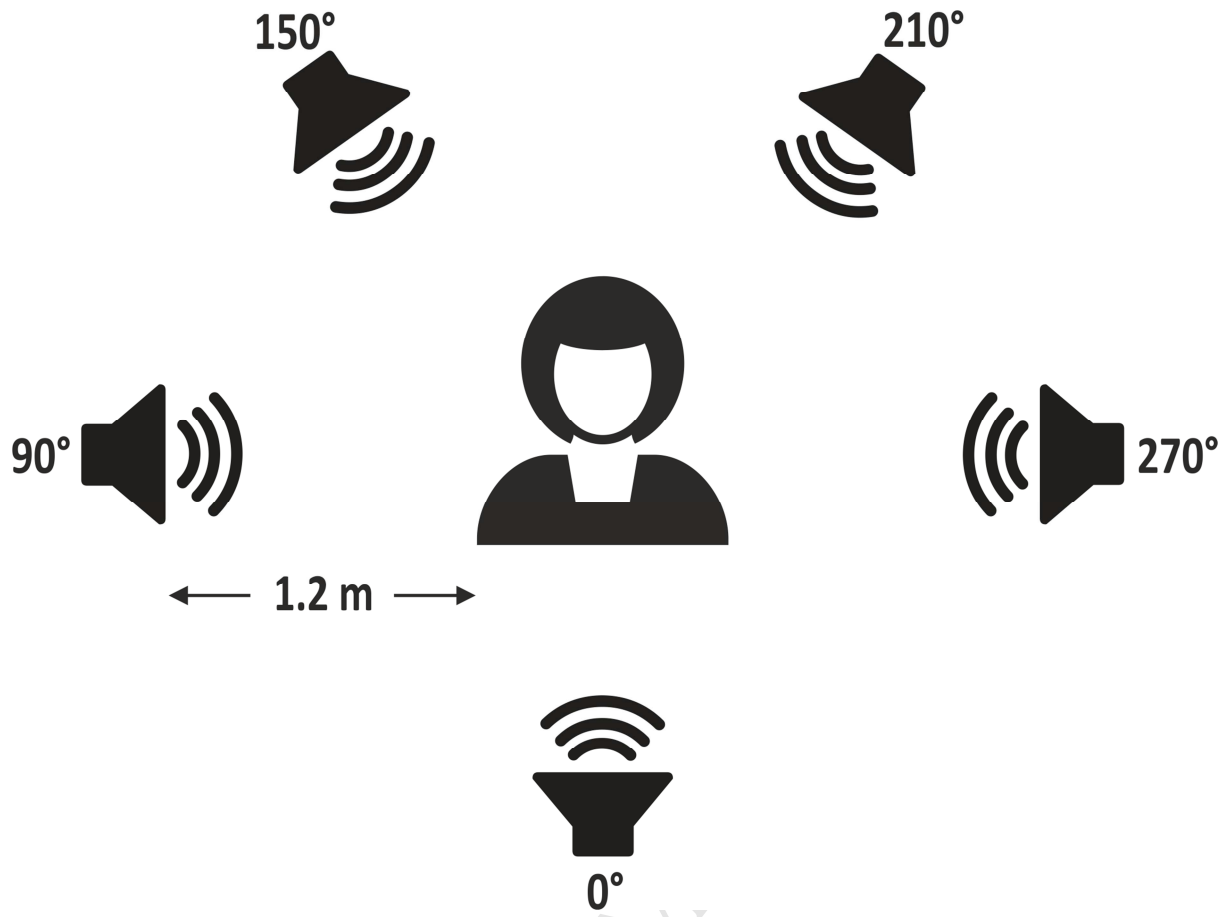
983 Deviance- a measure of the goodness of the model fit,

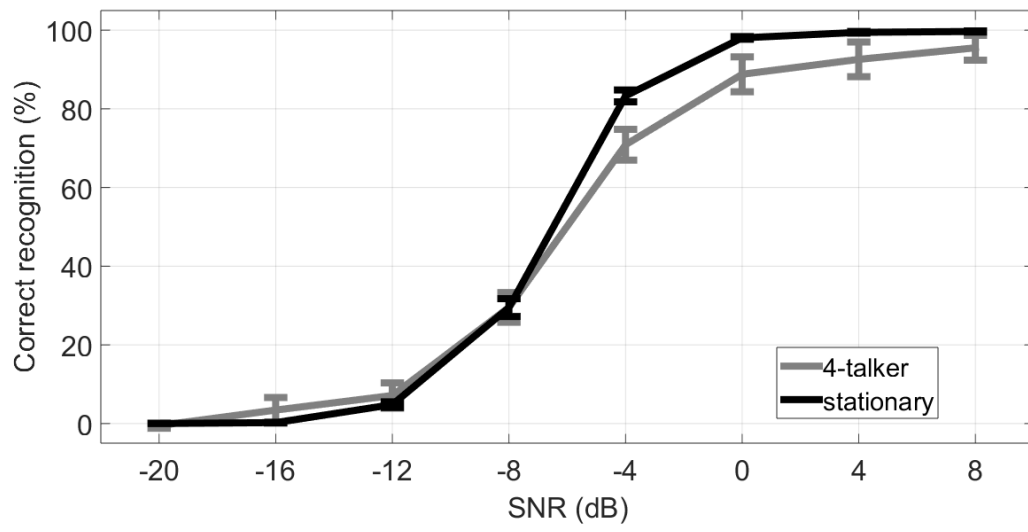
984 Df. Resid – Degree of Freedom for Residuals

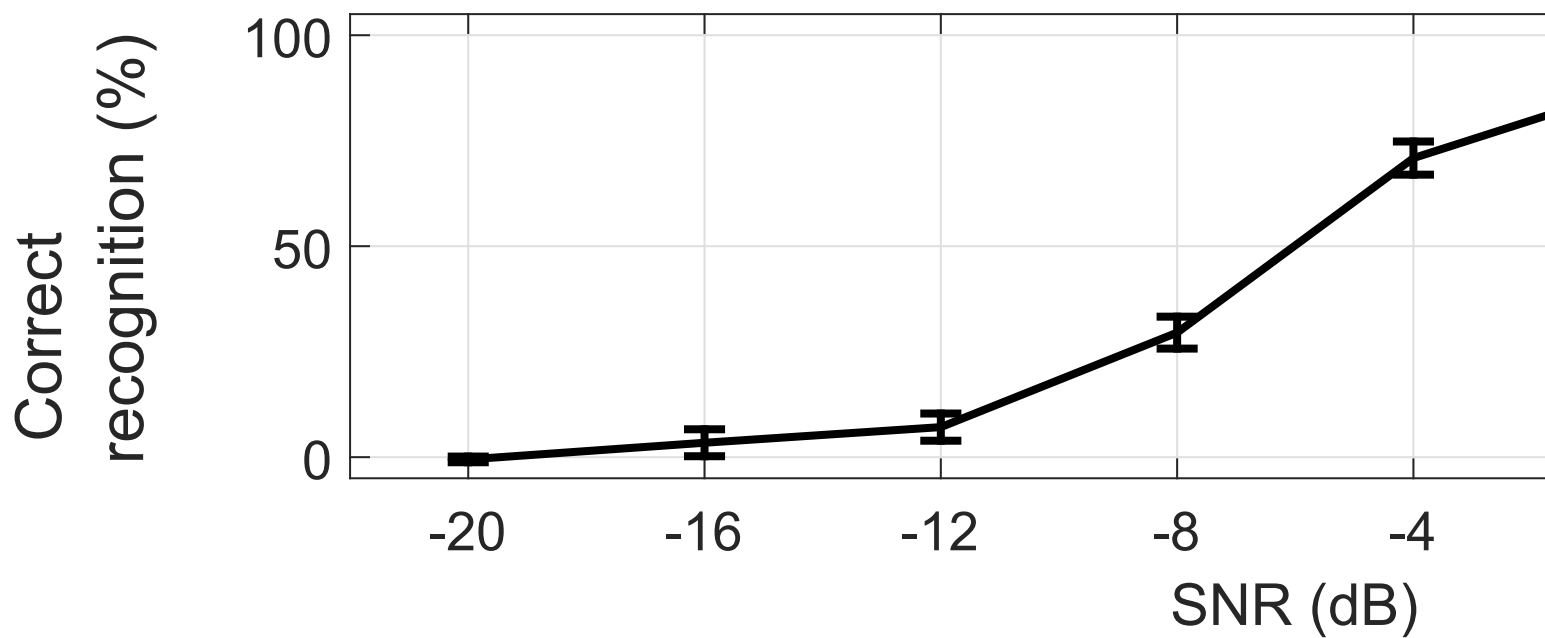
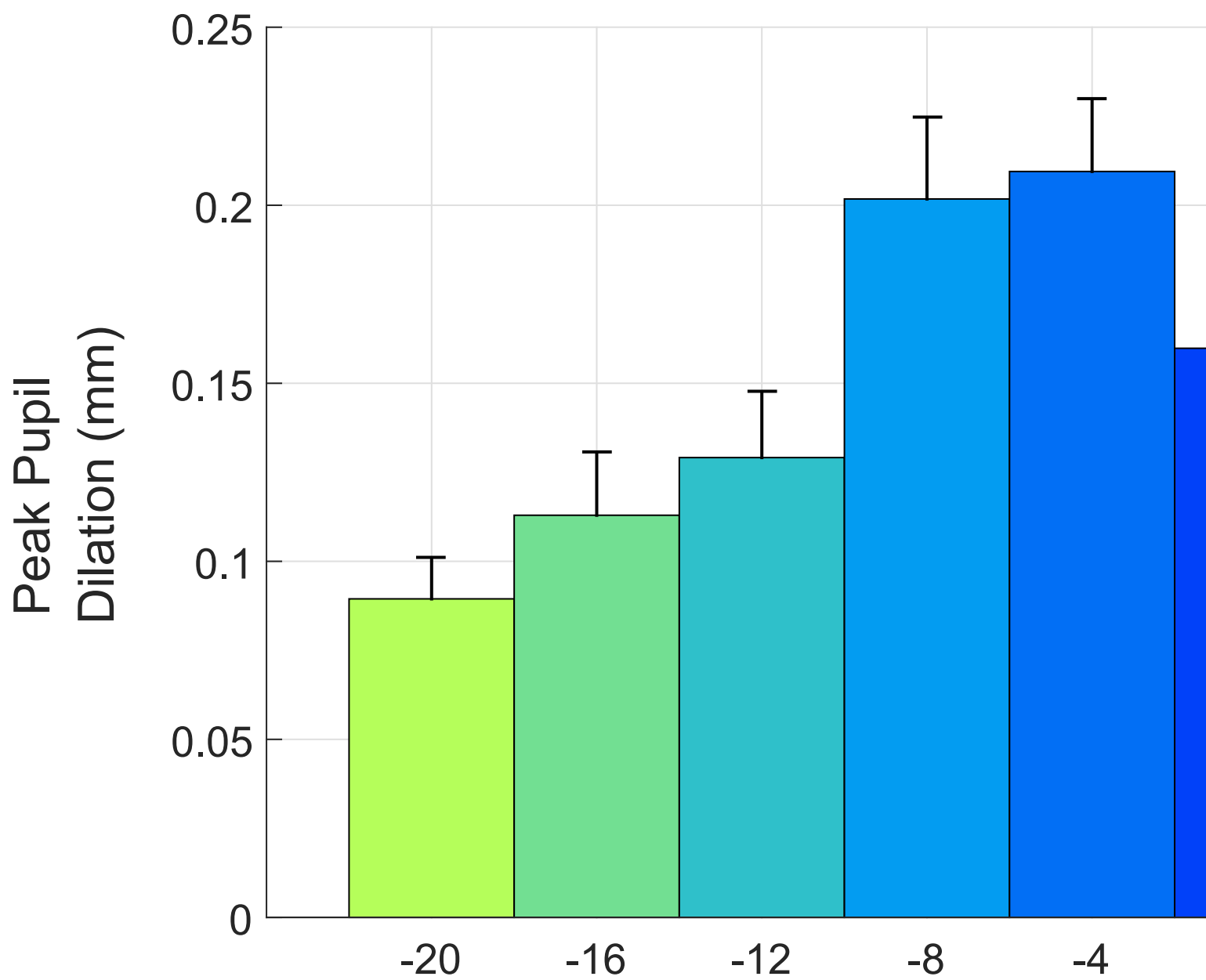
985 **Table 4:** Planned contrasts: The effect on the pupil dilation between the chosen SNRs in the 4-talker
 986 masker condition. Contrasts adjusted with the multivariate *t* adjustment.
 987

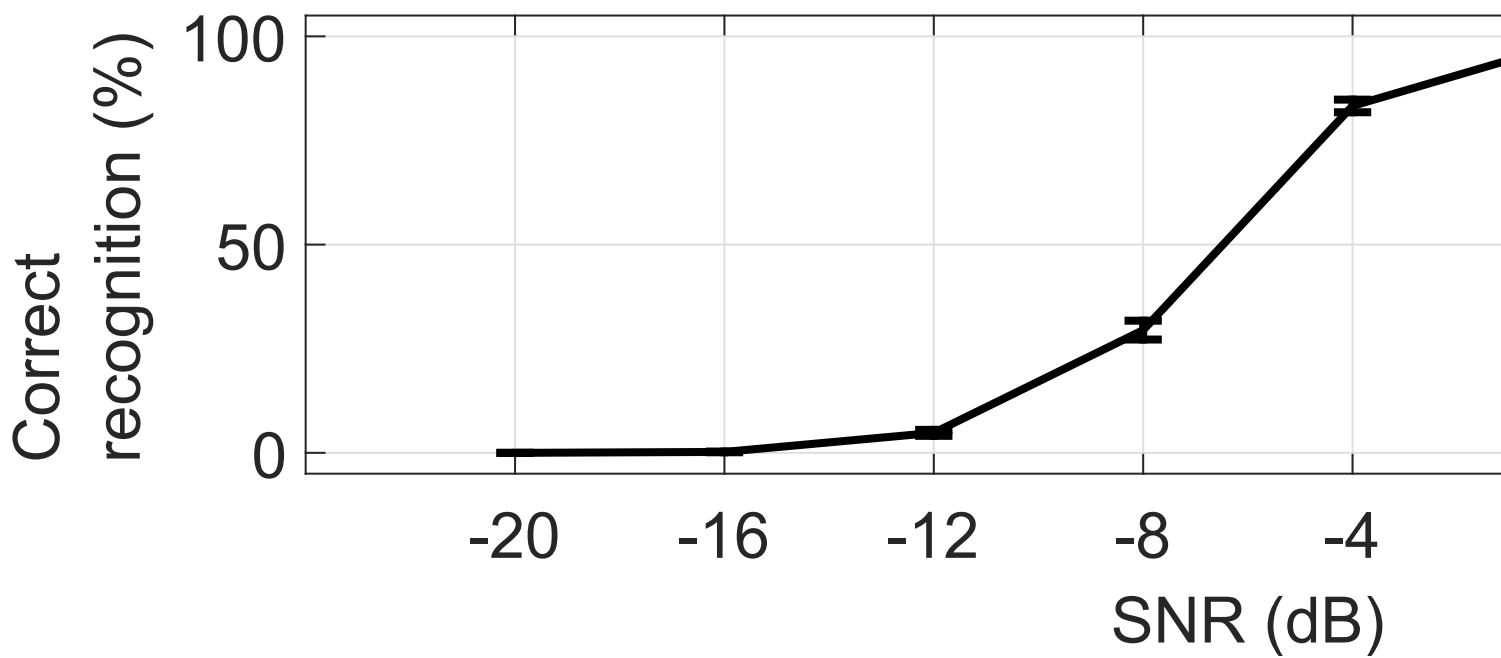
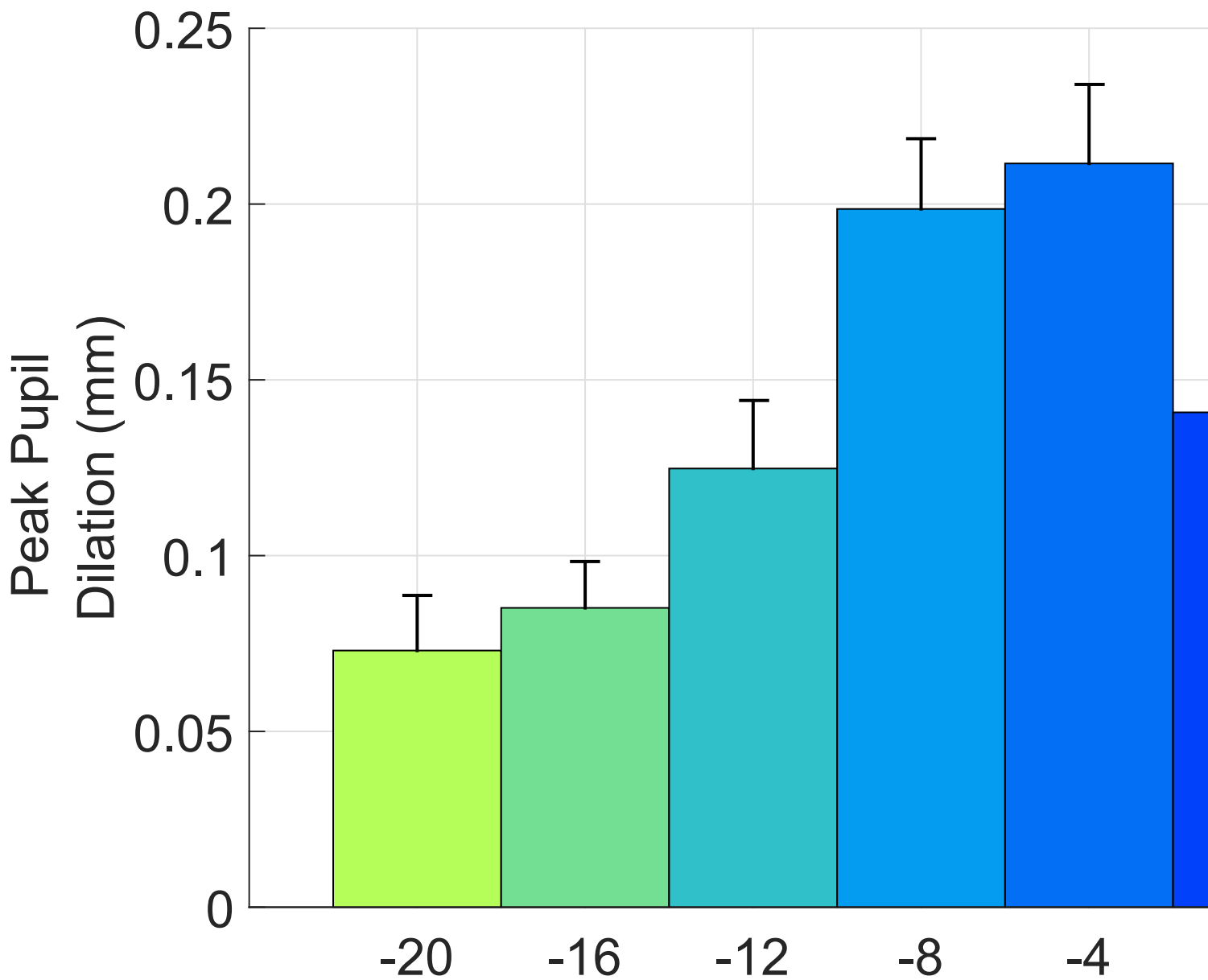
Contrast	-12dB vs. 4dB			-4dB vs. -8dB			4dB vs. 8dB		
	Term: 4 dB, Reference: -12dB			Term: -8dB, Reference: -4dB			Term: 8dB, Reference: 4dB		
Terms	β	t	p	β	t	p	β	t	p
Intercept	<u>-0.013</u>	<u>-3.532</u>	<u><0.001**</u>	0.003	0.901	0.367	-0.004	-0.963	0.335
Linear	-0.077	-1.512	0.131	0.035	0.700	0.484	0.026	0.508	0.611
Quadratic	<u>-0.136</u>	<u>-2.771</u>	<u>0.005**</u>	0.085	1.756	0.079	0.047	0.955	0.34
Cubic	0.044	1.164	0.245	0.012	0.328	0.743	<u>0.074</u>	<u>1.963</u>	<u>0.049*</u>

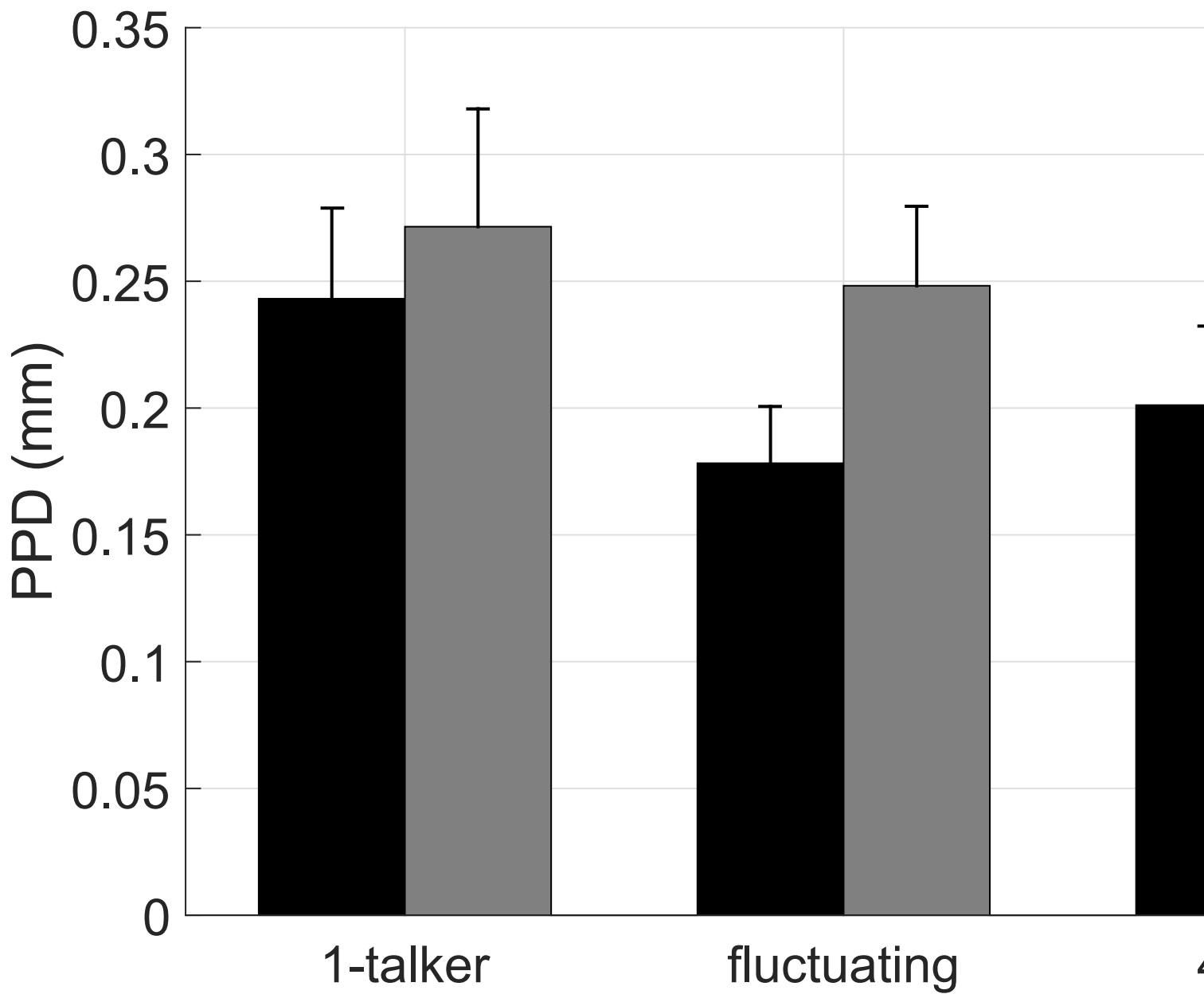
988
 989 * $p < 0.05$; ** $p < 0.01$

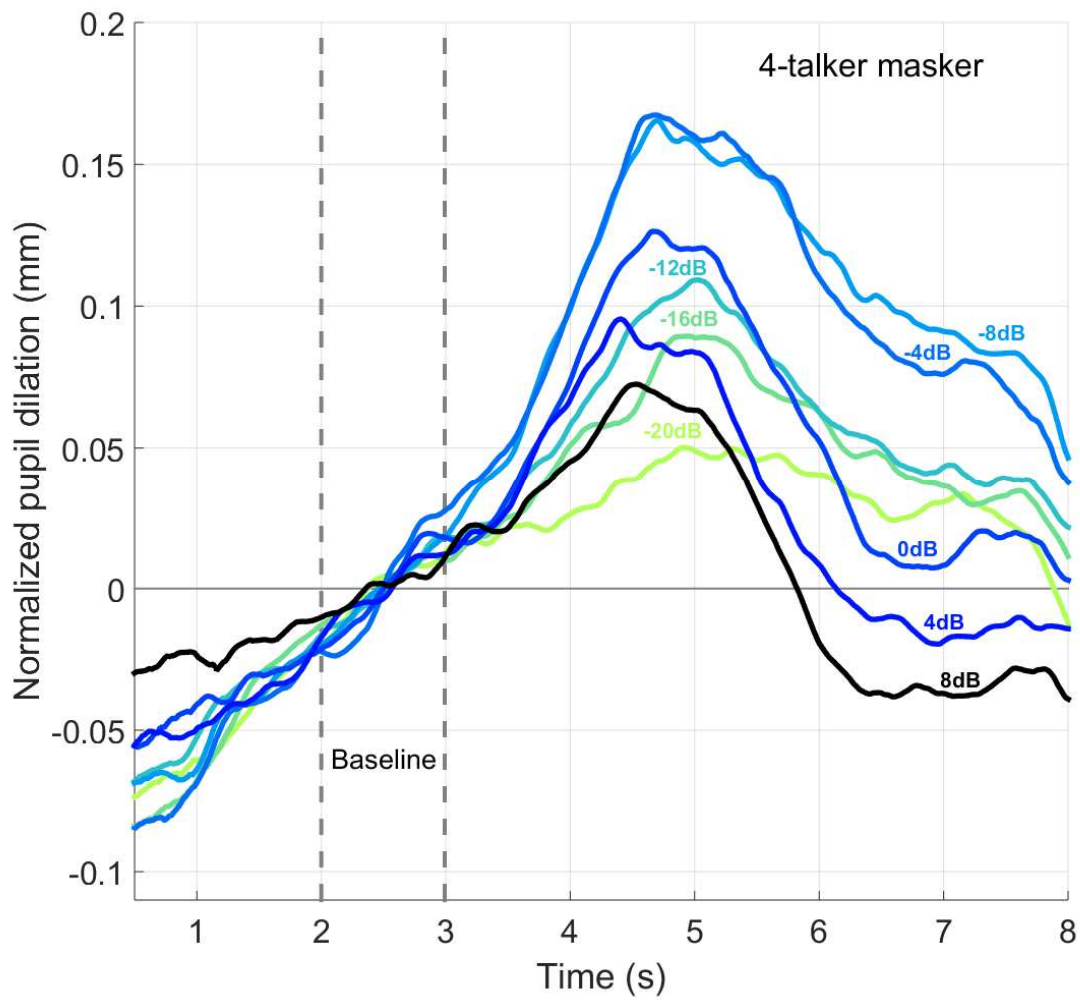


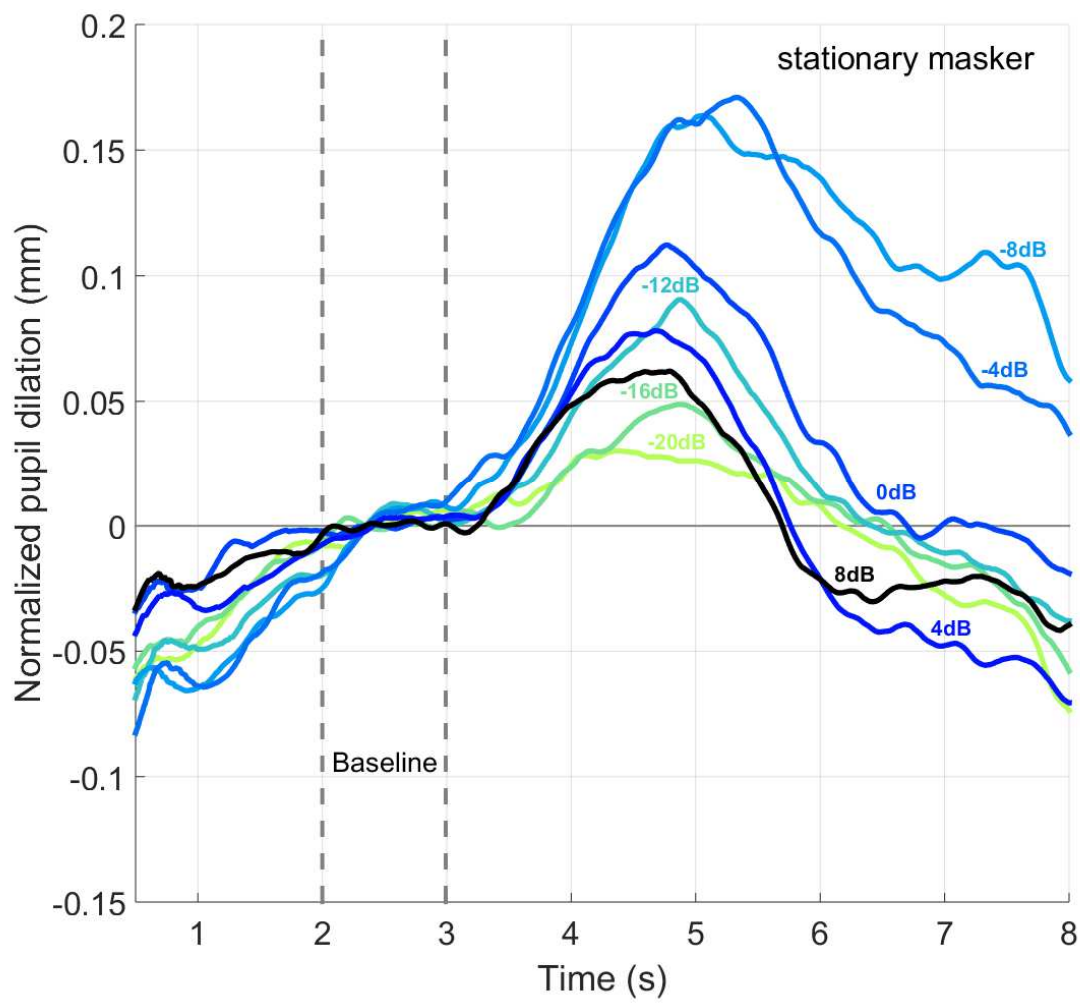


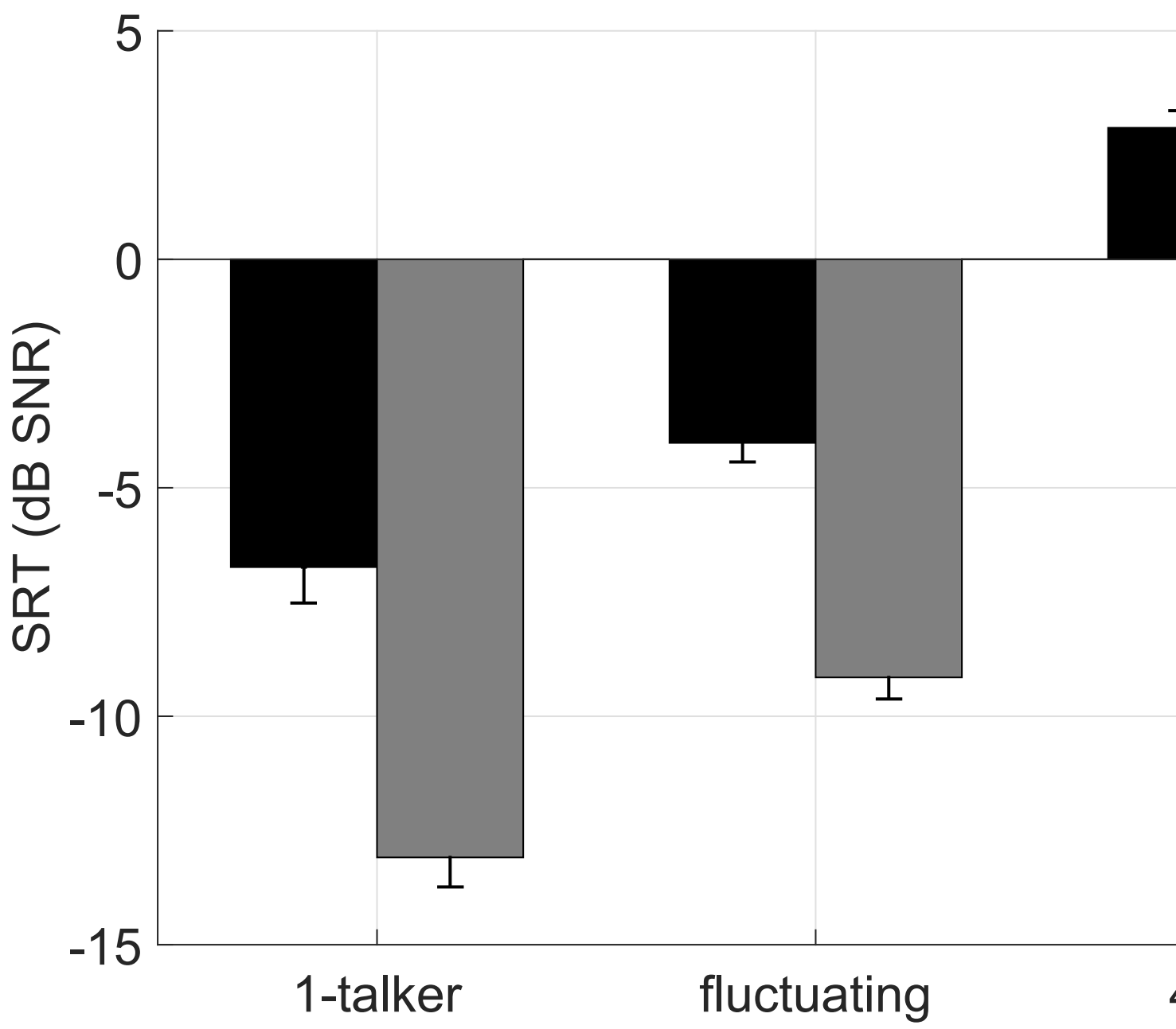












Highlights:

- Two experiments explored the impact of masker type and Signal-to-Noise Ratio on listening effort by means of pupillometry using a speech-in-noise test.
- Listening effort is highly affected by the masker type and the semantic interference of the masker.
- Pupillary response changed non-linearly across a range of fixed SNRs that corresponded to a wide range of recognition performance.
- The pupillary response demonstrated that listening effort is highest at intermediate SNRs corresponding to 30-70% speech intelligibility. Reduced pupillary response was measured at higher (favourable) SNRs corresponding to high intelligibility close to 100 %, reflecting lower listening effort likely due to a favourable listening situation and low task demands. Pupillary response was furthermore reduced at low (unfavourable) SNRs corresponding to intelligibility between 0-30%, which suggested that listeners spent less resources probably due to disengagement and giving up in those adverse listening situations.