Perceptual consequences of different signal changes due to binaural noise reduction
Do hearing loss and working memory capacity play a role?
Neher, Tobias; Grimm, Giso; Hohmann, Volker

Published in:
Ear and Hearing

DOI:
10.1097/AUD.0000000000000054

Publication date:
2014

Document version
Accepted manuscript

Citation for published version (APA):

Terms of use
This work is brought to you by the University of Southern Denmark through the SDU Research Portal. Unless otherwise specified it has been shared according to the terms for self-archiving.
If no other license is stated, these terms apply:
• You may download this work for personal use only.
• 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 this open access version
If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 02. Sep. 2020
Perceptual consequences of different signal changes due to binaural noise reduction: Do hearing loss and working memory capacity play a role?

Dr. Tobias Neher, Dr. Giso Grimm, and Prof. Dr. Volker Hohmann

Medical Physics and Cluster of Excellence Hearing4all

Oldenburg University
D-26111 Oldenburg
Germany

Address for correspondence:
Dr. Tobias Neher
Dept. of Medical Physics and Acoustics
Oldenburg University
D-26111 Oldenburg
Germany

Phone: +49 (0)441 2172-321
Fax: +49 (0)441 2172-350
Email: tobias.neher@uni-oldenburg.de

Conflicts of Interest and Source of Funding:

There are no conflicts of interest. This research was funded by the DFG Cluster of Excellence EXC 1077/1 “Hearing4all”.
STRUCTURED ABSTRACT

Objectives

In a previous study, Neher et al. (2013) investigated whether pure-tone average hearing loss (PTA) and working memory capacity (WMC) modulate benefit from different binaural noise reduction (NR) settings. Results showed that listeners with smaller WMC preferred strong over moderate NR even at the expense of poorer speech recognition due to greater speech distortion, whereas listeners with larger WMC did not. To enable a better understanding of these findings, the main aims of the current study were (1) to explore the perceptual consequences of changes to the signal mixture, target speech, and background noise caused by binaural NR, and (2) to determine whether response to these changes varies with WMC and PTA.

Design

As in the previous study, four age-matched groups of elderly listeners (with $N = 10$ per group) characterized by either mild or moderate PTAs and either better or worse performance on a visual measure of WMC participated. Five processing conditions were tested, which were based on the previously used (binaural coherence-based) NR scheme designed to attenuate diffuse signal components at mid to high frequencies. The five conditions differed in terms of the type of processing that was applied (no NR, strong NR, or strong NR with restoration of the long-term stimulus spectrum) and in terms of whether the target speech and background noise were processed in the same manner or whether one signal was left unprocessed while the other signal was processed with the gains computed for the signal mixture. Comparison across these conditions allowed assessing the effects of changes in high-frequency audibility (HFA), speech distortion (SD), and noise attenuation and distortion (NAD). Outcome measures included a dual-task paradigm combining speech recognition...
with a visual reaction time (VRT) task as well as ratings of perceived effort and overall preference. All measurements were carried out using headphone simulations of a frontal target speaker in a busy cafeteria.

Results
Relative to no NR, strong NR was found to impair speech recognition and VRT performance slightly, and to improve perceived effort and overall preference markedly. Relative to strong NR, strong NR with restoration of the long-term stimulus spectrum and thus HFA did not affect speech recognition, restored VRT performance to that achievable with no NR, and increased perceived effort and reduced overall preference markedly. SD had negative effects on speech recognition and perceived effort, particularly when both speech and noise were processed with the gains computed for the signal mixture. NAD had positive effects on speech recognition, perceived effort, and overall preference, particularly when the target speech was left unprocessed. VRT performance was unaffected by SD and NAD. None of the datasets exhibited any clear signs that response to the different signal changes varies with PTA or WMC.

Conclusions
For the outcome measures and stimuli applied here, the current study provides little evidence that PTA or WMC affect response to changes in HFA, SD, and NAD caused by binaural NR. However, statistical power limitations suggest further research is needed. This research should also investigate whether partial HFA restoration combined with some pre-processing that reduces co-modulation distortion results in a more favorable balance of the effects of binaural NR across outcome dimensions, and whether NR strength has any influence on these results.
1 Keywords

2 Hearing loss, cognition, hearing aids, noise reduction, individualization
INTRODUCTION

Modern digital hearing aids (HAs) are equipped with a range of signal processing algorithms including dynamic range compression, noise reduction (NR), and directional microphones (e.g., Kates 2008; Schaub 2008). The performance of these algorithms is controlled by various parameters, which are typically set based on technical and perceptual evaluations. Eventually, certain algorithm settings are obtained that – according to the evaluation criteria applied – are considered to be optimal. However, while on average the obtained settings may confer some benefit to the intended target group, it is very common to observe large inter-individual variability in outcome. Presumably, this is due to the fact that HA users can differ in terms of a multitude of, for example, peripheral, central-auditory, or cognitive characteristics, even if they have similar audiograms and ages (cf., CHABA 1988). Thus, it is of interest to identify any user characteristics that can explain a noteworthy proportion of this variability, as this would enable the development of fitting rules that could take such dependencies into account.

Nevertheless, probably because of the large number of available HA algorithms and potentially salient user characteristics, relatively little progress has been made with respect to tailoring HA signal processing better to individual users so far. As a matter of fact, the only user data that routinely are taken into account when fitting HAs are pure-tone audiometric thresholds, which provide the basis for prescribing amplification (e.g., Dillon 2012). As far as other types of HA technology such as NR or microphone directionality are concerned, the influence of hearing thresholds on outcome is rather unexplored and thus unclear (cf., Houben et al. 2012; Luts et al. 2010; Ricketts & Mueller 2000). Instead, audiological researchers have become increasingly interested in the role of cognitive factors for benefit from HA signal processing. Previous research has shown that cognitive factors such as working memory capacity (WMC), selective attention, and processing speed are necessary
for effective speech communication in noise (e.g., Akeroyd 2008; Humes et al. 2013; Neher et al. 2009). Based on a review of 20 studies on speech recognition and cognitive abilities, Akeroyd (2008) also found that while pure-tone average hearing loss (PTA) was the primary predictor of speech-in-noise performance, WMC as quantified using a measure of reading span (RS) was the second most important predictor.

Simultaneous processing and short-term storage of information in the form of mental representations are the key components of working memory (e.g., Daneman & Carpenter 1980). They are thought to draw on a number of executive functions, including the ability to update information, shift attention, and inhibit such behavior if necessary (e.g., Baddeley 1986). Given that the importance of these functions for speech understanding under challenging conditions is well documented (e.g., Rönnberg 2003; Shinn-Cunningham 2008) and that these functions typically, albeit not necessarily, decline with increasing age (e.g., Nyberg et al. 2012; Salthouse 1982), it is not surprising that researchers have concentrated on how WMC affects HA outcome. More precisely, researchers have looked at the influence of WMC on performance with dynamic range compression (Foo et al. 2007; Gatehouse et al. 2006a, 2006b; Lunner & Sundewall-Thorén 2007), NR processing (Lunner 2003; Ng et al. 2013), and frequency compression (Arehart et al. 2013). Taken together, the results from these studies – which stem almost exclusively from (objective) measures such as speech recognition or recall of the final words of a set of sentence-in-noise stimuli (Ng et al. 2013) – suggest that listeners with larger WMC perform better with stronger forms of HA signal processing, whereas listeners with smaller WMC perform better with more moderate forms of HA signal processing.

In a previous experiment, we extended these studies in a number of ways (Neher et al. 2013). Firstly, we controlled both PTA and WMC to be able to assess their effects on outcome not only individually but also in combination. Secondly, we included both objective
and subjective outcome measures, as these two types of measures can produce divergent data patterns (e.g., Brons et al. 2013). We controlled PTA and WMC by testing four age-matched groups of elderly hearing-impaired (HI) listeners with either smaller (H+) or larger (H-) PTAs and either larger (C+) or smaller (C-) WMC as assessed using a RS measure after Daneman and Carpenter (1980). Specifically, our first group (‘GH+C+’) was characterized by mild hearing loss and larger WMC, our second group (‘GH+C−’) was characterized by moderate hearing loss and larger WMC, our third group (‘GH−C−’) was characterized by mild hearing loss and smaller WMC, and our fourth group (‘GH−C−’) was characterized by moderate hearing loss and smaller WMC. In terms of HA signal processing, we used a binaural coherence-based NR algorithm designed to suppress diffuse signal components at mid to high frequencies. We varied the strength of this algorithm from inactive through moderate to strong, and performed measurements at a number of fixed signal-to-noise ratios (SNRs) between −4 and 8 dB. More specifically, we used an (objective) dual-task paradigm combining a speech recognition task with a visual reaction time (VRT) task assumed to index listening effort (cf., Sarampalis et al. 2009). In addition, we collected (subjective, binary) preference judgments. All measurements were made using headphone simulations of a frontal speech target in a busy cafeteria. Results showed that, for all groups, speech recognition was unaffected by moderate NR processing, whereas strong NR processing reduced recognition by about 5%. Similarly, VRT performance was unaffected by moderate NR but slightly impaired by strong NR for all groups. Preference scores (collapsed across SNR) showed that all groups preferred some over no NR processing. Furthermore, the two C- groups preferred strong over moderate NR processing, while for the two C+ groups preference did not differ significantly between the moderate and strong settings.

The fact that we only found group differences in our preference data and that smaller WMC was associated with a preference for stronger NR seems at odds with the findings
Neher et al.  Noise reduction, hearing loss, and working memory

summarized above. To recapitulate, these suggested that HA users with smaller WMC fare better with less aggressive signal processing whereas HA users with larger WMC fare better with more aggressive signal processing. Thus, the main purpose of the current study was to gain a better understanding of our previous findings. Specifically, we wanted to investigate whether the previously observed group differences could be explained based on differences in response to different types of signal changes caused by the binaural NR algorithm. Typically, NR processing results in both desirable (e.g., noise attenuation) and undesirable (e.g., speech distortion) signal changes (e.g., Loizou 2007). In general, both types of effects can be expected to increase with stronger NR processing (e.g., Loizou & Kim 2011). In the current study, we made use of a simulated HA that allowed us to tease apart the dominant signal changes caused by the binaural NR algorithm under investigation. More precisely, we investigated the separate effects of restoring the long-term stimulus spectrum or high-frequency audibility (HFA), distortion of the target speech, and attenuation and distortion of the background noise. We then employed a number of objective and subjective outcome measures to explore the perceptual consequences of these changes for our four groups of listeners.

A number of earlier studies indirectly addressed the effects of HFA by investigating how high-frequency bandwidth influences speech understanding, cognitive effort, and perceived sound quality. With respect to speech understanding, these studies showed that extended (>3-3.5 kHz) bandwidth generally results in small improvements for listeners with gently sloping mild to moderate hearing losses and that this benefit decreases as high-frequency PTAs increase (e.g., Amos & Humes 2007; Hornsby et al. 2011; Hornsby & Ricketts 2006; Skinner & Miller 1983). With respect to cognitive effort, one study investigated the effect of low-pass filtering speech-in-noise stimuli on the time it took normal-hearing (NH) children to respond to them (McCreery & Stelmachowicz 2013). It was
found that, even when speech recognition remained unaffected, reduced bandwidth caused slower responses, which was taken as an indicator for greater cognitive effort. With respect to perceived sound quality, one study found listeners with gently sloping, mild to moderate hearing losses to prefer music and movie stimuli with extended (>5.5 kHz) bandwidths, while for listeners with steeply sloping hearing losses the opposite was true (Ricketts et al. 2008). In another study, preferred bandwidth of music stimuli was investigated with NH and HI listeners (who apparently exhibited different degrees and configurations of hearing loss), and in this case only the NH listeners were found to prefer the wideband (>4 kHz) condition (Franks 1982).

The effects of signal distortion on listener response have also been addressed previously. For example, Arehart et al. (2010) investigated the relative impact of noise, linear filtering, and non-linear processing on the perceived sound quality of speech signals for NH and HI listeners. They found that the quality ratings of their HI listeners were affected at least as much if not more than the quality ratings of their NH participants by the different types of signal degradation applied. Even though this suggests that hearing loss increases the susceptibility to such effects, other studies have basically come to the opposite conclusion (e.g., Stelmachowicz et al. 1999; Tan & Moore 2008). In another study, Arehart et al. (2013) systematically investigated the effects of noise, distortion due to frequency compression, and WMC on the speech recognition performance of 26 HI listeners aged 62-92 yr. WMC was assessed using a measure of RS after Daneman and Carpenter (1980). As the amount of overall signal degradation due to both noise and distortion increased, listeners with larger WMC achieved better speech recognition than listeners with smaller WMC. Arehart et al. therefore concluded that smaller WMC is associated with increased susceptibility to signal degradation caused by the cumulative effects of noise and at least some types of HA signal processing.
In summary, although previous research has dealt with the perceptual consequences of HFA and signal distortion, results are sometimes inconsistent or are lacking altogether for speech-in-noise stimuli and elderly HI populations. In addition, the effects of PTA and WMC have not been explored in detail so far – neither in isolation nor in combination. Furthermore, although Arehart et al. (2013) pointed out that the effects of WMC they observed under conditions of noise combined with distortion due to frequency compression appear to also hold for dynamic range compression, it is uncertain whether this is also the case for algorithms whose specific purpose it is to reduce background noise. In our previous study, the C- listeners had apparently experienced our fixed-SNR stimuli as more strenuous and had favored stronger NR processing, even at the expense of poorer speech recognition due to greater speech distortion. In the current study, we therefore hypothesized that they would be more affected by noise and less sensitive to any signal distortions due to binaural NR. In contrast, because the C+ listeners had appeared less affected by the noise, we hypothesized that they would be more sensitive to any signal distortions.

In terms of objective measures, we employed the (combined) speech recognition and VRT tasks from our earlier study. Previously, we had been unable to correct the VRT scores for any inter-individual differences in baseline performance (e.g., due to differences in dexterity) as is common practice in dual-task studies involving elderly participants (e.g., Desjardins & Doherty 2013; Hornsby 2013; Picou et al. 2013; Somberg & Salthouse 1982). It is therefore possible that these scores had been insensitive to any existing differences among our groups of listeners. Thus, an additional aim of the current study was to re-examine potential group differences in VRT performance by modifying our procedure such that baseline correction was possible. In addition, we included two subjective outcome measures. Firstly, we assessed overall preference again. This time, however, we collected paired-comparison ratings instead of binary judgments to enable our listeners to indicate the extent
to which a given processing condition was preferred over another. Secondly, we included listening effort scaling (e.g., Luts et al. 2010) as a direct measure of perceived effort that we could benchmark our VRT measure against.

In view of the shortage of comparable research, the current study was rather exploratory in nature. Nevertheless, in accordance with the above-mentioned hypotheses we expected to find group differences (or effects of WMC) in our preference data, and perhaps also in the baseline-corrected VRT and listening effort scaling data. Furthermore, in accordance with the reviewed literature, we hypothesized that restored HFA would lead to improved VRT performance and perhaps also to a minor increase in speech recognition, at least for the two H+ groups. Finally, we expected that if target speech distortion (SD) was to have any effect on outcome it would be negative, while for noise attenuation and distortion (NAD) the opposite would be the case.

**MATERIALS AND METHODS**

Ethical approval for all experimental procedures was obtained from the ethics committee of the University of Oldenburg.

**Participants**

As in our previous study, 40 listeners participated whom we stratified into four groups (with $N = 10$ per group) based on the medians of our sample’s PTA and RS data. Out of these, 37 participants had taken part in our previous study; the remaining three listeners (one from each of the $G_{H-C+}$, $G_{H+C-}$, and $G_{H-C-}$ groups) were unavailable for the current study. Based on an extra recruitment round we were able to find three participants who closely resembled the three unavailable listeners in terms of age, PTA, and RS performance. In line with the inclusion criteria applied previously, these listeners also had bilateral, gradually sloping, sensorineural hearing losses, asymmetry in air-conduction thresholds of no more...
than 15 dB HL at any audiometric frequency between 125 Hz and 8 kHz, air-bone gaps of no
larger than 15 dB HL at any audiometric frequency between 500 Hz and 4 kHz, and self-
reported normal or corrected-to-normal vision. Thirty-three out of the final 40 participants
were habitual HA users with at least nine months of HA experience; the other seven
participants were nonusers. Furthermore, all participants had experience with similar research
studies. They were paid on an hourly basis for their participation.

From Table 1 it is apparent that the four groups were closely matched in terms of age,
PTA, and RS. Average age ranged from 73.0 to 76.6 yr. Average PTA, which was calculated
across ears for the standard audiometric frequencies from 500 Hz to 4 kHz, was 36.8 and 38.8
dB HL for the two H+ groups and 55.4 and 58.7 dB HL for the two H- groups. A one-way
analysis of variance (ANOVA) confirmed the lack of significant group differences in terms
of age (F[3,36] = 0.9; p > 0.05; partial-eta squared, $\eta^2_P = 0.07$). Another one-way ANOVA
(F [3,36] = 34.6; $p < 0.0001$, $\eta^2_P = 0.74$) followed by a post hoc analysis with Bonferroni
correction confirmed significant differences in terms of PTA between all pairs of groups with
different hearing status (all $p < 0.0001$), and no significant difference in terms of PTA
between any two groups with the same hearing status (all $p = 1.0$).

**Reading Span Test**

Cognitive function was characterized using the RS test after Daneman and Carpenter
(1980), which is a visual measure of WMC. This test is increasingly being used in
audiological research (e.g., Arehart et al. 2013; Desjardins & Doherty 2013; Neher et al.
2011; Ng et al. 2013), and our implementation closely mimics that of other researchers (see
Neher et al. 2013 for details). For the 37 participants who had taken part in the previous study
the RS data collected previously were used. For the other three participants, RS performance
was measured in accordance with the procedure described in (Neher et al. 2013). As before,
the performance measure used was the percentage of correctly recalled first and final words presented across a total of 54 sentences. Average RS performance was 44.6% and 44.1% (standard deviations: 6.0% and 6.0%) of correctly recalled first and final words for the two C+ groups and 28.1% and 27.4% (standard deviations: 4.3% and 5.2%) for the two C- groups (see Table 1). A one-way ANOVA (\(F[3,36] = 31.1; p < 0.0001, \eta^2_p = 0.72\)) followed by a post hoc analysis with Bonferroni correction confirmed significant differences in terms of RS between all pairs of groups with different cognitive status (all \(p < 0.0001\)) and no significant difference in terms of RS between any two groups with the same cognitive status (all \(p = 1.0\)).

**Physical Test Setup**

The test setup employed was very similar to the one used previously. Inside a soundproof booth, two computer screens and a keyboard were located. One screen was used for displaying information to the participants. The other screen, which the participants were unable to see during the tests, was used by the experimenter for scoring the participants’ responses to the speech stimuli (see below). All test software was implemented in MatLab (MathWorks, Natick, USA). Audio playback was via an Auritec (Hamburg, Germany) Earbox Highpower soundcard and a pair of Sennheiser (Wennebostel, Germany) HDA200 headphones. Calibration was carried out using a Brüel & Kjær (B&amp;K; Nærum, Denmark) 4153 artificial ear, a B&amp;K 4134 1/2” microphone, a B&amp;K 2669 preamplifier, and a B&amp;K 2610 measurement amplifier.

**Speech Stimuli**

The speech stimuli closely resembled those from our previous study. They were based on recordings from the Oldenburg sentence material (Wagener et al. 1999), which consists of 120 sentences that are low in semantic context and that all follow the form ‘name verb..."
numeral adjective object’ (e.g., “Thomas has two large flowers”). To simulate a realistic complex listening situation the sentence recordings were convolved with a pair of head-related impulse responses (HRIRs). These HRIRs were measured in a large, reverberant cafeteria using the built-in microphones of a B&K head-and-torso simulator (HATS) and a frontal source at a distance of 1 m from, and at the same height as, the HATS (Kayser et al. 2009). The resultant sentences ranged in length from 2.2 to 3.2 sec. They were stored as .wav files with a sampling rate of 44.1 kHz and a resolution of 16 bit.

For the interfering signal, a HATS recording made in the same cafeteria during a busy lunch hour was selected. On each trial, a 5-sec extract from this recording was randomly chosen and processed to have 50-msec raised-cosine on- and offset ramps. The resultant signal was presented at a nominal sound pressure level of 65 dB. It was mixed with a given target sentence, which started 1.25 sec after the cafeteria noise and which was adjusted in level to produce a given SNR (see below).

**Hearing Aid Processing**

All signal processing was implemented on the Master Hearing Aid (MHA) research platform (Grimm et al. 2006). It included the binaural coherence-based NR scheme\(^1\) used previously, linear amplification, and headphone equalization, and was carried out at a sampling rate of 16 kHz. Prior to any processing, the speech stimuli were therefore resampled to 16 kHz. Following the processing, they were resampled to 44.1 kHz.

**Processing Conditions** • A total of five processing conditions were tested that will be referred to as \(S_{\text{off}}N_{\text{off}}\), \(S_{\text{on}}N_{\text{on}}\), \(S_{\text{on}+g}N_{\text{on}+g}\), \(S_{\text{off}}N_{\text{on}+g}\), and \(S_{\text{on}+g}N_{\text{off}}\). These conditions differed in terms of (1) the type of processing that was applied (‘off’, ‘on’, or ‘on+g’) and (2) if the target speech, \(S\), and background noise, \(N\), were processed in the same manner or if one signal was left unprocessed while the other signal was subjected to ‘on+g’ processing. From Figure 1, which shows a block diagram of the simulated HA, it is apparent that there were
three separate (2-channel) signal paths – one for S, one for N, and one for the signal mixture, SN. Each of these signals was analyzed using a Fast Fourier Transform (FFT) based filterbank with twelve frequency bands covering an 8-kHz bandwidth (see Neher et al. 2013 for details). Using a 40-msec integration time constant, the binaural coherence (or interaural similarity) of SN was first estimated in each frequency band based on the method described by Grimm et al. (2009). The estimates produced by this method can take on values between 0 and 1. A value of 0 corresponds to fully incoherent (or diffuse) sound, while a value of 1 corresponds to fully coherent (or directional) sound. Because of diffraction effects around the head, the binaural coherence is always high below about 1 kHz. At higher frequencies, the coherence is low for diffuse and reverberant signal components, but high for the direct sound from nearby sources. In the current study, NR gains, $G_{L,R}(k)$, in the range of $-30$ to $0$ dB were derived by applying an exponent of 2 (corresponding to the strong NR setting from our previous study) to the coherence estimates. In this manner, gains of 0 dB were obtained for fully coherent SN components, while gains of $-30$ dB were obtained for fully incoherent SN components. These gains could then be applied to S and/or N (processing type ‘on’); alternatively, S and/or N could remain unprocessed (processing type ‘off’). The NR stage was followed by a gain stage that allowed changing the long-term spectrum of S and/or N. Level estimators, $L(k)$, placed in the SN path before and after the NR stage allowed estimating the change in level due to ‘on’ processing across the entire (5-sec) duration of a given SN stimulus. Based on the obtained level estimates, frequency-dependent gains, $\Delta L(k)$, could then be derived and applied following the NR to compensate for any long-term spectral changes (processing type ‘on+g’); alternatively, these gains could all be set to 0 dB (processing types ‘off’ and ‘on’). Thus, by configuring the NR and subsequent gain stage in the S and N paths in the appropriate manner, each of the $S_{off}N_{off}$, $S_{on}N_{on}$, $S_{on+g}N_{on+g}$, $S_{off}N_{on+g}$, and $S_{on+g}N_{off}$ conditions could be obtained. Following the analysis filterbank, NR, and
subsequent gain stage, the signals were spectrally shaped in accordance with the NAL-RP
prescription rule (Byrne et al. 1991) to ensure adequate audibility for each participant. The
resultant signals were then passed through a resynthesis filterbank, and the output signals
were summed. Finally, a 32nd-order finite impulse response filter was applied to the (2-
channel) output stimulus to compensate for the uneven magnitude response of the HDA200
headphones.

Regarding the five processing conditions, it should be noted that the S_{off}N_{off} and S_{on}N_{on}
conditions were virtually identical to the inactive and strong NR settings from our previous
study (see Introduction). Consequently, they served as reference conditions for the current
study. It should also be noted that while it would basically be possible to implement the
S_{off}N_{off}, S_{on}N_{on}, and S_{on+g}N_{on+g} conditions in some actual HAs (based on a single SN path
each), the same is not true for the S_{off}N_{on+g} and S_{on+g}N_{off} conditions (because of their separate
S and N paths). The motivation for including these two (artificial) processing conditions here
was to be able to explore the perceptual consequences of the applied processing for S and N
separately.

Signal Changes

The chosen processing conditions gave rise to a number of signal changes, which are
illustrated in Figure 2 for an example stimulus with an input SNR of 4 dB. The panels on the
left-hand side show, for each processing condition, the waveforms of S and N after NR
processing and the subsequent long-term gain correction stage. The panels on the right-hand
side show the changes in gain over time and frequency caused by the S_{on}N_{on}, S_{on+g}N_{on+g},
S_{off}N_{on+g}, and S_{on+g}N_{off} conditions, all relative to the S_{off}N_{off} condition. A prominent effect of
the S_{on}N_{on} condition is to suppress (incoherent) signal components above about 1 kHz (see
the first panel on the right-hand side of Figure 2). The resultant low-pass characteristic is
compensated in the S_{on+g}N_{on+g} condition, leading to full restoration of the long-term stimulus
Neher et al.

Noise reduction, hearing loss, and working memory

spectrum (see the second panel on the right-hand side of Figure 2). It is important to note, however, that in terms of speech-weighted SNR improvement (‘ΔAI-SNR’) the $S_{on}N_{on}$ and $S_{on+g}N_{on+g}$ conditions were virtually identical. That is, relative to the $S_{off}N_{off}$ condition these two conditions led to $ΔAI$-SNRs of about 3.6 dB at 4 dB SNR (and about 2.9 and 1.6 dB at 0 and $-4$ dB SNR, respectively).

The third and fourth panel on the right-hand side of Figure 2 show corresponding data for the $S_{off}N_{on+g}$ and $S_{on+g}N_{off}$ conditions. Note that these data correspond to the gain changes undergone by, respectively, $N$ and $S$ as a consequence of ‘on+g’ processing. For the stimulus used here, the $S_{off}N_{on+g}$ and $S_{on+g}N_{off}$ conditions gave rise to $ΔAI$-SNRs of 2.7 and 1.0 dB, respectively. Thus, most of the SNR improvement brought about by $S_{on+g}N_{on+g}$ processing was due to the attenuation of $N$. However, $S$ was also amplified (and thus modified) somewhat relative to the $S_{off}N_{off}$ condition. Apart from these effects, $S$ and $N$ were affected by co-modulation distortion, which is apparent in the form of short-term gain fluctuations from 1.25 to about 3.8 sec into the stimulus (see the third and fourth panel on the right-hand side of Figure 2). This was because, at a given instant in time, the NR gain calculated based on $SN$ was determined by the stronger of the two ($S$ or $N$) signals. Analysis of these short-term gain fluctuations revealed that, in terms of overall magnitude, these co-modulation effects were more pronounced in $N$.

Altogether, the five chosen test conditions enabled us to explore the effects of three types of signal changes: restored HFA, SD, and NAD. Specifically, by comparing across the $S_{on}N_{on}$ and $S_{on+g}N_{on+g}$ conditions we could assess the effects of restored HFA. Furthermore, by comparing across the $S_{off}N_{off}$ and $S_{on+g}N_{off}$ conditions and across the $S_{off}N_{on+g}$ and $S_{on+g}N_{on+g}$ conditions we could assess the effects of SD. Finally, by comparing across the $S_{off}N_{off}$ and $S_{off}N_{on+g}$ conditions and across the $S_{on+g}N_{off}$ and $S_{on+g}N_{on+g}$ conditions we could assess the effects of NAD.
Dual-Task Test

As pointed out above, our motivation for including the dual-task test was to re-examine potential group differences in terms of response to NR processing. In accordance with our previous study, we did not expect to find such effects for the speech task, but perhaps for the VRT task due to the modified baseline procedure. Such differences would then be indicative of some groups having to expend less (or more) cognitive resources on the speech task due to certain types of NR processing, thereby leading to a concurrent performance improvement (or decrement) on the VRT task (cf., Sarampalis et al. 2009).

The dual-task test used in the current study was almost identical to that from our previous study. Participants were required to complete a speech recognition task and a concurrent VRT task. They were informed that the two tasks were equally important and that they therefore should pay equal amounts of attention to them. The main differences to our previous study were that we assessed baseline VRT performance for a number of processing conditions in the presence of background noise (rather than for the unprocessed condition in quiet only), and that we carried out all of these measurements at the beginning (rather than at randomly determined points in time throughout the course of the dual-task measurements). In this manner, we aimed to achieve a constant level of arousal and thereby prevent an effect we had observed previously, namely that our two C-groups had unintentionally slowed down in the quiet condition relative to the (supposedly more strenuous) noisy conditions. This then had prevented us from using these baseline scores to correct the ones obtained under noisy conditions, as any differences among our groups would have been distorted.

Speech Recognition Task • Speech recognition was measured at SNRs of −4 and 0 dB. Thus, a total of 10 conditions (5 processing conditions × 2 SNRs) were tested. Since our previous study had revealed good test-retest reliability for these measurements at these SNRs, only one measurement was made per condition. For each measurement, one randomly chosen
test list (consisting of 20 five-word sentences each) was used. Test lists were not repeated. Following the presentation of a speech stimulus, listeners had to repeat the words they had understood, which were scored by an experimenter using a graphical user interface displayed on the experimenter’s screen. To start with, training measurements were made with $S_{on}N_{on}$ processing at 4 dB SNR, and then with $S_{off}N_{on+g}$ processing at 0 dB SNR. These were followed by the 10 actual measurements, which were carried out in randomized order. Participants were required to take a 10-min break halfway through the 12 (training and actual) measurements.

**Visual Reaction Time Task** • For the VRT task, two boxes were displayed on the participant’s screen. The boxes were 6 cm high, 4.5 cm wide, and 1.5 cm apart. At quasi-random intervals, a digit between 1 and 8 appeared in one of the boxes. Using the keyboard provided, participants had to press a key marked by a green sticker if the digit was even or a key marked by a red sticker if the digit was odd. They were told to perform this task as quickly as possible while maintaining a high level of accuracy. Two digits were presented per speech stimulus. The first digit appeared after a randomly chosen delay of between 0.5 and 1.5 sec relative to the start of the cafeteria noise. It remained on the screen until the participant had pressed either key, or for maximally 2 sec. The second digit then appeared after another randomly chosen delay of between 0.5 and 1.5 sec and remained visible until either key was pressed, or for maximally 2 sec. Missed digits, digits responded to correctly, and reaction times for the digits responded to correctly were recorded for each trial. The arithmetic mean of these reaction times was then calculated to obtain the absolute VRT ('VRT_{abs}') score for a given condition.

**Baseline Correction** • Prior to any measurements, all participants practiced the VRT task on its own until they could perform it reliably. Baseline VRT performance was then assessed to allow controlling for inter-individual differences in motor function (under presumably
constant levels of arousal). Our approach was similar to that of Hornsby (2013). That is, based on the dual-task procedure described above, four measurements were made during which the listener was asked to concentrate on the digits and to ignore the concurrently presented audio. The first measurement was made with $S_{\text{on}}N_{\text{on}}$ processing at 4 dB SNR and was carried out for training purposes only. It was followed by three measurements carried out at −4, 0, and 4 dB SNR with $S_{\text{off}}N_{\text{off}}$, $S_{\text{on}}N_{\text{on}}$, or $S_{\text{on}+g}N_{\text{on}+g}$ processing applied. The three SNRs and processing conditions were randomly combined, and their order of presentation was also randomized. The median of the three resultant $V_{\text{VRT}}^{\text{abs}}$ scores was taken as the VRT baseline (‘$V_{\text{VRT}}^{\text{base}}$’) score. Inter-individual differences in absolute VRT performance were then corrected for by transforming the $V_{\text{VRT}}^{\text{abs}}$ scores into relative VRT (‘$V_{\text{VRT}}^{\text{rel}}$’) scores and expressed in terms of a percentage based on the following equation:

$$V_{\text{VRT}}^{\text{rel}}(\%) = 100 \times \left( \frac{V_{\text{abs}} - V_{\text{base}}}{V_{\text{base}}} \right).$$

Positive $V_{\text{VRT}}^{\text{rel}}$ scores therefore indicate slower (i.e., worse) VRT performance.

### Overall Preference Ratings

In addition to the dual-task measurements, we collected paired-comparison ratings in terms of overall preference. Listeners were asked to imagine being inside the cafeteria and wanting to communicate with the speaker of the sentences. They then had to make pairwise comparisons of the different processing conditions and to decide which setting they preferred overall and (unlike in our previous study) also by how much. Test conditions were identical to the ones used for the dual-task test, except that an additional SNR of 4 dB was included and that no digits were presented concurrently with the audio. On each trial, four speech stimuli were generated as described above and then concatenated, resulting in a 20-sec signal. Comparisons were blocked by SNR. Using a graphical user interface (GUI) and a touch screen, listeners could control playback of the (looped) stimuli as well as enter their responses. On the GUI, stimuli were labeled ‘A’ and ‘B’. A 7-point rating scale was used that
Neher et al.  Noise reduction, hearing loss, and working memory

ranged from “A is very much better” to “B is very much better” and that included the option
to indicate no difference in preference between A and B (cf., ITU-T 1996). Six trials
corresponding to the three possible combinations of \( S_{off}N_{off} \), \( S_{on}N_{on} \), and \( S_{on+g}N_{on+g} \)
processing at both \(-4\) and \(4\) \(\text{dB SNR} \) were initially presented in randomized order for
familiarization purposes. They were followed by 60 pairwise comparisons (3 SNRs \(\times\) 10
possible combinations of the five processing conditions \(\times\) 2 presentations each), the order of
which was also randomized. In the subsequent statistical analyses, the first set of ratings was
treated as test data, while the second set was treated as retest data.

9 Listening Effort Scaling

To further explore the perceptual effects of the different processing conditions we also
collected subjective ratings of listening effort. Our procedure was similar to that of Luts et al.
(2010). Using a 9-point rating scale ranging from “completely effortless” to “maximally
effortful” listeners had to indicate how much effort they thought they had to expend to be
able to understand a set of target sentences. Stimuli were generated in the same manner as for
the overall preference measurements. Again, SNRs of \(-4\), 0, and \(4\) \(\text{dB} \) were used. Participants
were required to listen to each (20-sec) stimulus at least once before they could indicate their
response using a GUI and a touch screen. For familiarization purposes, six trials
corresponding to \( S_{off}N_{off} \), \( S_{on}N_{on} \), or \( S_{on+g}N_{on+g} \) at \(-4\) and \(4\) \(\text{dB SNR} \) each were initially
presented in randomized order. They were followed by 30 trials (3 SNRs \(\times\) 5 processing
conditions \(\times\) 2 presentations each), the order of which was also randomized. In the
subsequent statistical analyses, the first set of ratings was treated as test data, while the
second set was treated as retest data.

21
Test Protocol

Most of the 37 participants who had taken part in our previous study attended two 1.5-h visits. At the first visit, the participant’s audiogram was re-measured if the one measured previously was older than six months. Next, the measurements related to the dual-task test were performed. At the second visit, the listening effort and overall preference measurements were made. For the three newly recruited participants there were three 1- to 1.5-h visits. At the first visit, the participant’s audiogram was measured and the RS test was administered. In addition, 30-40 min of systematic training on the dual-task test were completed. At the second visit, the measurements related to the dual-task test were made, while the overall preference and listening effort measurements were carried out at the third visit.

Statistical Analyses

In preparation for the statistical analyses, we divided the speech scores by 100 (giving a range up to 1) and transformed them into rationalized arcsine units (RAU; Studebaker 1985) to normalize the variance across the range of scores. We then carried out mixed-model ANOVAs on the RAU, VRT\textsubscript{rel}, and LES scores. Whenever appropriate, we applied the Greenhouse-Geisser correction to correct for violations of sphericity. Because we could not expect the preference scores to be interval-scaled, we analyzed them using Friedman or Kruskal-Wallis tests and followed any significant effects up with series of Wilcoxon signed-rank tests. In each case, we applied a Bonferroni correction to the significance level to prevent inflation of the type-I error rate.

Furthermore, to assess possible relations between the different outcomes we performed a correlation analysis. For the correlations involving the preference scores we used Spearman’s $\rho$ rank correlation coefficient; for all other correlations we used Pearson’s $r$ correlation coefficient. Again, we adjusted the significance level using a Bonferroni correction.
Given the exploratory nature of the current study, we did not have estimates of test-retest reliability available for the measures and populations of interest and were thus unable to perform an a priori power analysis to determine the sample size needed for ensuring adequate statistical power. However, for all non-significant ANOVA effects we included estimates of observed statistical power (OP), so that, in conjunction with the effect size (i.e., eta squared, $\eta^2$) estimates it would be possible for the reader to judge the relevance for further work.

**RESULTS**

**Speech Recognition**

To analyze the RAU scores we performed a mixed-model ANOVA with SNR and processing condition as within-subject factors and listener group as between-subject factor. We found significant effects of SNR ($F[1,35] = 792.0; p < 0.0001; \eta^2_p = 0.96$), processing condition ($F[4,140] = 25.6; p < 0.0001; \eta^2_p = 0.42$), and listener group ($F[3,35] = 8.1; p < 0.001; \eta^2_p = 0.41$). Furthermore, the interaction between SNR and listener group was statistically significant ($F[3,35] = 4.0; p < 0.05; \eta^2_p = 0.26$). The effects of processing condition $\times$ listener group ($F[12,140] = 1.2; p > 0.05; \eta^2_p = 0.09$), SNR $\times$ processing condition ($F[4,140] = 2.1; p > 0.05; \eta^2_p = 0.06$), and SNR $\times$ processing condition $\times$ listener group ($F[12,140] = 0.9; p > 0.05; \eta^2_p = 0.07$) were all non-significant.

Figure 3a shows mean RAU scores and associated 95% confidence intervals (CIs) for the different groups of listeners as a function of SNR. At the higher SNR, speech recognition performance was better, as expected. Furthermore, the $G_{H+C+}$ and $G_{H-C-}$ groups lay at the high and low end of the observed performance range, respectively. To investigate the significant effect of listener group further we carried out a post hoc analysis with Bonferroni correction.
This revealed that the performance of the GH+C+ group differed significantly from that of the GH-C+ ($p < 0.01$) and G-C+ ($p < 0.001$) groups.

Figure 3b shows mean RAU scores and associated 95% CIs for the different processing conditions averaged across the two SNRs. Relative to SoffNoff (mean: 76.5 RAU), SoffNon+g resulted in better speech recognition (mean: 81.1 RAU), whereas the speech recognition achievable with the other three processing conditions was lower than that achievable with no NR (means: 72.4-73.1 RAU). To investigate the significant effect of processing condition further we carried out a series of planned contrasts. These revealed statistically significant differences between SoffNoff and all other processing conditions (all $p < 0.01$; see Figure 3b).

Furthermore, we observed statistically significant differences between SoffNoff and all other processing conditions (all $p < 0.0001$). Relative to SoffNoff, SoffNon+g therefore led to a performance improvement of about 5 RAU, whereas the other processing conditions led to performance decrements of 3-4 RAU. In terms of the three types of signal changes considered, we observed no effect of HFA on speech recognition (SoffNon vs. SoffNon+g; $p > 0.05$). However, we observed negative effects of SD on speech recognition. Specifically, performance worsened by about 4 RAU when N was left unprocessed (SoffNoff vs. SoffNoff; $p < 0.0001$), and by about 8 RAU when N was subjected to ‘on+g’ processing (SoffNoff vs. SoffNoff; $p < 0.00001$). With respect to NAD, we observed a performance improvement of 4.6 RAU when S was left unprocessed (SoffNoff vs. SoffNoff; $p < 0.0001$), whereas performance remained unaffected when S was subjected to ‘on+g’ processing (SoffNoff vs. SoffNoff; $p > 0.05$).

**Visual Reaction Times**

Prior to performing any statistical analyses on the VRT data, we examined their reliability by calculating the percentage of digits missed and the percentage of digits responded to correctly for each group. The percentage of digits missed ranged from 0.6 to
1.4% across groups (grand average: 1.0%), and the percentage of digits responded to correctly ranged from 97.5 to 98.4% (grand average: 97.8%). Thus, as had been the case in our previous study, our participants were able to perform the concurrent secondary task in a reliable manner. Also in line with our previous study, VRT performance varied considerably within each group of listeners, with the VRT\textsubscript{abs} scores ranging from about 0.6 to 1 sec for each group.

To analyze the VRT\textsubscript{rel} scores we performed a mixed-model ANOVA with SNR and processing condition as within-subject factors and listener group as between-subject factor. We observed significant effects of SNR ($F_{[1,36]} = 6.7; p < 0.05; \eta^2_p = 0.16$) and processing condition ($F_{[3.1,112.6]} = 5.4; p < 0.01; \eta^2_p = 0.13$). Furthermore, the interaction between SNR, processing condition, and listener group was statistically significant ($F_{[10.6,127.3]} = 2.1; p < 0.05; \eta^2_p = 0.15$). The effects of listener group ($F_{[3,36]} = 1.1; p > 0.05; \eta^2_p = 0.09$), SNR $\times$ listener group ($F_{[3,36]} = 2.2; p > 0.05; \eta^2_p = 0.16$), processing condition $\times$ listener group ($F_{[9.4,112.6]} = 0.8; p > 0.05; \eta^2_p = 0.06$), and SNR $\times$ processing condition ($F_{[3.5,127.3]} = 1.3; p > 0.05; \eta^2_p = 0.04$) were all non-significant.

Figure 4a shows mean VRT\textsubscript{rel} scores and associated 95% CIs for the different groups of listeners as a function of SNR. As expected, VRT\textsubscript{rel} scores were generally larger at –4 dB SNR. Furthermore, there was a (non-significant) tendency for the two C+ groups to obtain lower (i.e., better) VRT\textsubscript{rel} scores than the two C- groups. Figure 4b shows mean VRT\textsubscript{rel} scores and associated 95% CIs for the different processing conditions averaged across the two SNRs. Relative to S\textsubscript{off}N\textsubscript{off} (mean: 8.3%), S\textsubscript{on}N\textsubscript{on} resulted in worse VRT performance (mean: 11.6%), whereas the performance achievable with the other three processing conditions was very similar to that achievable with no NR (means: 8.6-8.8%). To investigate the significant effect of processing condition further we carried out a series of planned contrasts. These revealed statistically significant differences between S\textsubscript{on}N\textsubscript{on} and each of the
other processing conditions (all $p < 0.01$; see Figure 4b). Thus, relative to all the other conditions, $S_{on}N_{on}$ led to a performance decrement of about 3% on the VRT task. No other differences in VRT performance were found. In terms of the three types of signal changes considered, we therefore found an VRT improvement of about 3% due to restored HFA ($S_{on}N_{on}$ vs. $S_{on+g}N_{on+g}$; $p < 0.01$). For SD ($S_{off}N_{off}$ vs. $S_{on+g}N_{off}$ and $S_{on+g}N_{on+g}$ vs. $S_{off}N_{on+g}$; all $p > 0.05$) and NAD ($S_{off}N_{off}$ vs. $S_{off+g}N_{on+g}$ and $S_{on+g}N_{on+g}$ vs. $S_{off+g}N_{off}$; all $p > 0.05$) no effects on VRT performance were observable.

To investigate the significant three-way interaction between SNR, processing condition, and listener group further we carried out a post hoc analysis with Bonferroni correction. This revealed that this effect was due to changes in the relative performance of the four groups of listeners in the $S_{off}N_{off}$, $S_{off+g}N_{on+g}$, and $S_{on+g}N_{off}$ conditions across the two SNRs. Closer inspection of these changes showed that they were on the order of a few percentage points and that they did not follow any systematic pattern.

**Overall Preference**

The paired-comparison ratings in terms of overall preference were analyzed as follows (cf., Brons et al. 2013; ITU-T 1996). First, we assigned a value between −3 and 3 to each rating. For instance, if a participant rated stimulus A as being much better than stimulus B, we assigned a value of 3 to that rating etc. We then used the resultant values to assess test-retest reliability. Specifically, we calculated Spearman’s $\rho$ rank correlation coefficient for each combination of listener group and SNR. We found the median of all 12 correlation coefficients to be 0.54 and the range 0.19 to 0.65. Test-retest reliability was lowest at −4 dB SNR (median $\rho$: 0.29; range: 0.19-0.65), intermediate at 0 dB SNR (median $\rho$: 0.46; range: 0.32-0.58), and highest at 4 dB SNR (median $\rho$: 0.61; range: 0.58-0.64). Furthermore, the $G_{H+}$ group gave the least consistent ratings (median $\rho$: 0.32; range: 0.19-0.64), while the $G_{H+C+}$
group gave the most consistent ratings (median $\rho$: 0.62; range: 0.50-0.65). Altogether, these data indicate that reproducibility of the preference ratings was rather poor at $-4$ dB SNR, but reasonable at 0 and 4 dB SNR.

The subsequent analyses of the preference data were based on scores that we derived as follows. First, we assigned the value that we previously had assigned to each (pairwise) rating to the preferred stimulus. Second, we sign-reversed the same value and assigned it to the other stimulus (e.g., $-3$ for stimulus B in the example above). In effect, this produced a scale ranging from ‘much better’ (+3) to ‘much worse’ (−3). Third, we averaged the values obtained for each listener, processing condition, and SNR. As a result of this, the overall mean across processing conditions was zero for each SNR and listener group. In the following analyses, the effects of SNR, listener group, and SNR \times listener group would therefore have been non-significant and were thus ignored. Furthermore, for reasons of clarity, we decided not to consider the three-way interaction between SNR, listener group, and processing condition.

To investigate the main effect of processing condition we averaged the preference scores across 0 and 4 (but not $-4$) dB SNR and analyzed them using a Friedman’s test. To investigate the interaction between processing condition and SNR we performed two Friedman’s tests on the data obtained at 0 and 4 dB SNR. To investigate the interaction between processing condition and listener group we averaged the preference scores across 0 and 4 (but not $-4$) dB SNR and performed a Kruskal-Wallis test on the data from each of the five processing conditions. Following a Bonferroni correction for eight tests, we obtained a significant effect from each Friedman test (all $\chi^2_{[40]} > 32.8$, all $p < 0.0001$) and a non-significant effect from each Kruskal-Wallis test (all $H_{[3]} < 8.6$, all $p > 0.05$).

Figure 5a shows mean preference scores and associated 95% CIs for the different processing conditions as a function of SNR. At $-4$ dB SNR, the five processing conditions
resulted in very similar preference scores. At the higher SNRs, preference for $S_{on}N_{on}$ increased, whereas for $S_{off}N_{off}$ and $S_{on+g}N_{off}$ preference tended to decrease. To investigate the interaction between processing condition and SNR further we analyzed the preference scores obtained at 0 and 4 dB SNR using two series of Wilcoxon signed-rank tests, each with a Bonferroni correction for 10 tests. For both SNRs, we observed significant differences for $S_{off}N_{off}$ vs. $S_{on}N_{on}$, $S_{off}N_{off}$ vs. $S_{on+g}N_{on+g}$, $S_{on}N_{on}$ vs. $S_{off}N_{on+g}$, $S_{on}N_{on}$ vs. $S_{on+g}N_{off}$, $S_{on+g}N_{on+g}$ vs. $S_{on+g}N_{off}$, and $S_{off}N_{on+g}$ vs. $S_{on+g}N_{off}$ (all $p < 0.05$). At 0 dB SNR, we also observed a significant difference for $S_{off}N_{off}$ vs. $S_{on+g}N_{on+g}$ ($p < 0.05$). At 4 dB SNR, we also observed a significant difference for $S_{on}N_{on}$ vs. $S_{on+g}N_{on+g}$ ($p < 0.0001$).

Figure 5b shows mean preference scores and associated 95% CIs for the different processing conditions averaged across 0 and 4 dB SNR. As can be seen, $S_{on}N_{on}$ was most preferred while $S_{off}N_{off}$ and $S_{on+g}N_{off}$ were least preferred. To investigate the significant effect of processing condition further we carried out a series of Wilcoxon signed-rank tests with a Bonferroni correction for 10 tests. Relative to no processing, we observed a clear positive effect of $S_{on}N_{on}$ on overall preference ($S_{off}N_{off}$ vs. $S_{on}N_{on}$: $p < 0.0001$). Restored HFA had a negative impact on preference ($S_{on}N_{on}$ vs. $S_{on+g}N_{on+g}$: $p < 0.001$), although relative to $S_{off}N_{off}$ processing $S_{on+g}N_{on+g}$ was still judged to be better ($p < 0.05$). SD had no effects on overall preference ($S_{off}N_{off}$ vs. $S_{on+g}N_{off}$ and $S_{on+g}N_{on+g}$ vs. $S_{off}N_{on+g}$; all $p > 0.05$). NAD, in turn, had a positive effect on overall preference, especially when $S$ was left unprocessed ($S_{on+g}N_{on+g}$ vs. $S_{on+g}N_{off}$: $p < 0.01$; $S_{off}N_{off}$ vs. $S_{off}N_{on+g}$: $p < 0.0001$).

**Perceived Listening Effort**

Prior to performing any statistical analyses on the listening effort scores, we examined their reliability by calculating Pearson’s $r$ correlation coefficient for each combination of listener group and SNR. We found the median of all correlation coefficients to be 0.69 (range: 0.47-0.86). Furthermore, with the exception of the data from the G$_{HH+C}$ group at −4 dB
SNR ($r = 0.47$) all correlations were larger than 0.59, indicating reasonable reproducibility overall.

For the subsequent analyses, we averaged the two effort ratings per listener, processing condition, and SNR. We then performed a mixed-model ANOVA with SNR and processing condition as within-subject factors and listener group as between-subject factor. We observed significant effects of SNR ($F [1.5,54.9] = 416.6; \ p < 0.00001; \ \eta^2_p = 0.92$) and processing condition ($F [2.4,85.7] = 46.1; \ p < 0.00001; \ \eta^2_p = 0.56$). Furthermore, the interactions between processing condition and SNR ($F [5.6,200.7] = 9.2; \ p < 0.00001; \ \eta^2_p = 0.20$) and processing condition and listener group ($F [7.1,85.7] = 2.6; \ p < 0.05; \ \eta^2_p = 0.18$) were statistically significant. The effects of listener group ($F [3,36] = 0.6; \ p > 0.05; \ \eta^2_p = 0.05$), SNR $\times$ listener group ($F [1.5,54.9] = 0.4; \ p > 0.05; \ \eta^2_p = 0.03$), and SNR $\times$ processing condition $\times$ listener group ($F [16.7,200.7] = 1.1; \ p > 0.05; \ \eta^2_p = 0.08$) were all non-significant.

Figure 6a shows mean effort scores and associated 95% CIs for the different processing conditions as a function of SNR. As expected, perceived effort reduced with increasing SNR, resulting in mean scores of 7.3, 4.8, and 3.0 scale points at $-4$, 0, and 4 dB SNR, respectively. Furthermore, at $-4$ dB SNR the effort scores were rather similar, while at 0 and especially 4 dB SNR $S_{on}N_{on}$ and $S_{off}N_{off+g}$ resulted in the lowest and $S_{off}N_{off}$ and $S_{on+g}N_{off}$ in the highest effort scores. To investigate the interaction between processing condition and SNR further we carried out a post hoc analysis with Bonferroni correction. We obtained confirmation for various changes in relative perceived effort brought about especially (but not exclusively) by $S_{on}N_{on}$ processing across $-4$ and 0 dB SNR (all $p < 0.01$). Across 0 and 4 dB SNR, the five processing conditions resulted in more similar (albeit not identical) patterns of performance (all $p < 0.05$).
Figure 6b shows mean listening effort scores and associated 95% CIs for the different processing conditions averaged across the three SNRs. As can be seen, perceived effort varied with processing condition, $S_{off}N_{off}$ and $S_{on+g}N_{off}$ leading to the highest (means: 5.6 scale points each) and $S_{on}N_{on}$ to the lowest (mean: 4.1 scale points) ratings. To investigate the significant effect of processing condition further we carried out a series of planned contrasts. Except for $S_{off}N_{off}$ vs. $S_{on+g}N_{off}$ ($p > 0.05$), we found statistically significant differences between all processing conditions (all $p < 0.05$; see Figure 6b). Specifically, relative to no processing we observed a large positive effect (corresponding to about 1.5 scale points) of $S_{on}N_{on}$ on perceived listening effort ($S_{off}N_{off}$ vs. $S_{on}N_{on}$: $p < 0.00001$). This improvement decreased substantially (i.e., to about 0.3 scale points) when HFA was restored ($S_{on}N_{on}$ vs. $S_{on+g}N_{on+g}$: $p < 0.00001$). With respect to SD, we observed a negative effect (corresponding to about 0.5 scale points) when N was subjected to ‘on+g’ processing ($S_{on+g}N_{on+g}$ vs. $S_{off}N_{on+g}$: $p < 0.00001$), but not when it was left unprocessed ($S_{off}N_{off}$ vs. $S_{on+g}N_{off}$: $p > 0.05$). NAD, in turn, was found to reduce perceived listening effort slightly (i.e., by about 0.3 scale points) when S was subjected to ‘on+g’ processing ($S_{on+g}N_{on+g}$ vs. $S_{on+g}N_{off}$: $p < 0.01$) and somewhat more (i.e., by about 0.8 scale points) when S was left unprocessed ($S_{off}N_{off}$ vs. $S_{off}N_{on+g}$: $p < 0.00001$). Altogether, therefore, there was good agreement between the effects of the various signal changes on perceived listening effort and overall preference.

To investigate the significant interaction between processing condition and listener group further we carried out another post hoc analysis with Bonferroni correction. We found that this effect was driven by across-group differences in the extent to which $S_{on}N_{on}$ processing improved perceived effort relative to $S_{off}N_{off}$ ($p < 0.01$) and $S_{on+g}N_{off}$ ($p < 0.05$) processing. More specifically, for the G_{H,C} listeners perceived effort improved by about 0.5 scale points due to $S_{on}N_{on}$ processing, whereas for the other groups this improvement was on the order of 1.5 to 2 scale points each. Otherwise, the group ratings were very similar.
Relations Among Outcome Measures

In order to assess the degree of concordance among our outcome measures, we calculated correlation coefficients for all pairwise combinations of the RAU, \( VRT_{rel} \), preference, and effort scores. Because we had found the reproducibility of the preference scores obtained at \(-4 \) dB SNR to be rather poor and because the RAU and \( VRT_{rel} \) scores were available at \(-4 \) and \( 0 \) dB SNR only, we decided to focus on the scores collected at \( 0 \) dB SNR. Thus, our correlation analysis was based on 200 data points (\( 1 \) SNR \( \times \) 5 processing conditions \( \times \) 40 listeners) per outcome measure. Following Bonferroni correction for six tests, we observed statistically significant correlations between the RAU and \( VRT_{rel} \) scores (Pearson’s \( r = -0.28, p < 0.001 \)), the RAU and LES scores (Pearson’s \( r = -0.27, p < 0.001 \)), and the preference and LES scores (Spearman’s \( \rho = -0.33, p < 0.0001 \)).

DISCUSSION

The present study had two main research aims. First, we wanted to explore the perceptual consequences of changes in HFA, SD, and NAD due to strong binaural NR based on measures of speech recognition, \( VRT \) performance, perceived listening effort, and overall preference. Second, we wanted to investigate if response to these changes is affected by PTA and WMC. To that end, four age-matched groups of elderly listeners with either mild or moderate PTA and either larger or smaller WMC were tested. Statistical analyses revealed a mixed influence of HFA on outcome. That is, although restored HFA had a positive effect on \( VRT \) performance, it also reduced overall preference and increased perceived effort. SD had negative effects on speech recognition and perceived effort, particularly when both S and N were processed with the NR gains computed for SN. NAD, in turn, had positive effects on speech recognition, perceived effort, and overall preference, particularly when S was left unprocessed. \( VRT \) performance was unaffected by both SD and NAD. Regarding the possible
influence of PTA and WMC, we found little evidence that these two user characteristics affect response to the three types of signal changes considered. In the following sections, these findings are discussed in more detail.

**Speech Recognition**

As part of the current study, we observed a reduction in speech recognition of about 4% due to $S_{on}N_{on}$ (relative to $S_{off}N_{off}$) processing. This basically replicates the 5% performance decrement due to the strong (relative to the inactive) NR setting that we had found previously (see Introduction). Furthermore, also in line with our previous study, the effect of listener group on speech recognition was significant.

Regarding the perceptual consequences of the various signal changes, we had hypothesized that restored HFA might have – at least for the two H+ groups – a minor positive effect on speech recognition, but this was not the case. Nevertheless, the literature findings that this hypothesis was based on were obtained with low-pass filtered stimuli, which basically had all their energy above a certain cut-off frequency (typically around 3-4 kHz) removed (see Introduction). In contrast, the binaural NR algorithm used here only removed incoherent signal components at mid to high frequencies, and from Figure 2 it is apparent that S still contained considerable high-frequency energy following $S_{on}N_{on}$ processing. Given that we had spectrally shaped our stimuli to ensure adequate audibility for each participant (see Hearing Aid Processing) and that restored HFA did not significantly improve audibility (see Signal Changes), this can probably explain why we did not observe any effects of restored HFA on speech recognition.

Also in concordance with our expectations, we found a negative effect of SD on speech recognition. This effect was particularly pronounced when not only S but also N was subjected to ‘on+g’ processing. Under these ($S_{on+g}N_{on+g}$) conditions, the two signals were processed with the gains computed for SN and thus were co-modulated (see Signal Changes).
In view of the finding that signals with similar amplitude modulation characteristics tend to be fused together by the auditory system (e.g., Moore 2003), it is likely that this co-modulation distortion handicapped our participants in their ability to segregate S from N, thereby resulting in poorer speech recognition. With respect to NAD, we had expected to find positive effects on speech recognition performance. Nevertheless, these were only apparent for the Soff\text{Noff} vs. Soff\text{N}_{on+g} comparison and not for the S_{on+g}N_{on+g} vs. S_{on+g}N_{off} comparison. This was even though subjecting S to ‘on+g’ processing had led to a larger ΔAI-SNR (see Signal Changes). Apparently, however, this SNR advantage was outweighed by the concurrent distortion artifacts.

With respect to the possible influence of WMC on the speech recognition achievable with the various processing conditions, we found no evidence for such a relationship. While this is broadly consistent with our previous study, it is at odds with the results of Arehart et al. (2013) who found smaller WMC to be associated with increased susceptibility to signal degradation due to noise and distortion caused by frequency compression (see Introduction). In addition, given that shorter compression time constants lead to more distortion (e.g., Moore 2008), it is at odds with research showing that listeners with better cognitive function achieve better speech recognition with fast-acting amplitude compression, whereas listeners with poorer cognitive function perform better with slow-acting amplitude compression (Gatehouse et al. 2006b; Lunner & Sundewall-Thorén 2007). Neither did we find any evidence that PTA interacts with the effects of the different processing conditions (or signal changes) on speech recognition. Again, while this is broadly consistent with our previous study, it is inconsistent with other research showing that listeners with relatively mild high-frequency hearing losses perform better with extended-bandwidth stimuli (e.g., Hornsby et al. 2011). At present, the reasons for these conflicting results are not clear, but they are probably related to differences across studies in terms of participants, stimuli, and processing.
Neher et al. Noise reduction, hearing loss, and working memory

conditions used. In this context, it is worth noting that although the two (non-significant) interactions involving listener group and processing condition had been characterized by relatively small effect sizes ($\eta_p^2 = 0.07-0.09$; see above), their observed statistical power had been moderately high (0.52-0.66). A slightly larger sample size could therefore have changed these results.

Interestingly, we observed statistically significant correlations between the RAU scores and both the VRT$_{rel}$ and the effort scores, although the magnitude of these correlations was rather small ($|r| < 0.3$). This finding is broadly consistent with reports from other NR studies that listening effort judgments are only weakly, if at all, associated with speech recognition performance (Brons et al. 2013; Luts et al. 2010), thereby justifying the inclusion of both types of measures in HA research.

**Visual Reaction Times**

In our previous study, we had observed longer VRTs due to strong (relative to inactive) NR processing (see Introduction), and this is also what we observed in the current study for S$_{on}$N$_{on}$ (relative to S$_{off}$N$_{off}$) processing. Furthermore, as in our previous study, the effects of listener group, listener group $\times$ SNR, and listener group $\times$ processing condition were all non-significant. This was in spite of the fact that our listeners had carried out the VRT task in a reliable fashion and that we had corrected the data for inter-individual differences in overall response speed in an attempt to improve sensitivity. In addition, although we had observed a statistically significant interaction between SNR, processing condition, and listener group, these effects had been rather small and unsystematic.

Altogether, these findings suggest that, irrespective of the processing condition, our four groups of listeners expended comparable cognitive resources on the speech task, thereby resulting in similar VRT$_{rel}$ scores. However, since we did not determine baseline speech
recognition performance we cannot rule out the possibility that some listeners traded performance on this task for performance on the VRT task. This general problem with dual-task paradigms means that the effects of interest may be masked, especially if there are between-group differences in baseline performance due to differences in, for example, auditory and cognitive function.

It could also be that, given only 10 participants per group, a secondary task such as the one used here is insufficiently sensitive for revealing differences among groups of elderly listeners differing in terms of WMC (and possibly other salient characteristics). Some support for this assumption is available from the fairly small effect sizes of the three (non-significant) effects involving the factor ‘listener group’ ($\eta^2_p = 0.06-0.16$; see above) coupled with their comparatively low observed statistical power (0.28-0.52). Incidentally, Desjardins and Doherty (2013), who combined a speech recognition task with a task requiring participants to track a moving target on a computer screen, did not observe any significant secondary task performance differences between 15 elderly NH (mean age: 67 yr) and 16 elderly HI (mean age: 68 yr) listeners either (although relative to 15 young NH listeners clear differences were apparent).

Alternatively, the degree of similarity between the secondary task and the cognitive measure used to stratify the participants could play a role for secondary task sensitivity. In the study by Ng et al. (2013) referred to above, the effects of WMC, noise, and binary mask-based NR processing on speech identification and recall performance were investigated. Compared to our participants and those of Desjardins and Doherty (2013), the listeners tested by Ng et al. were somewhat younger (mean age: 59 yr; range: 32-65 yr). For the statistical analyses, these listeners were divided into two (age-matched) groups with either better or worse performance on the RS test ($N=13$ each). Results showed a significant group difference in terms of performance on the (secondary) recall task. That is, as a result of NR
processing speech recall improved for persons with larger WMC but not for listeners with smaller WMC. Unlike in our study and the one of Desjardins and Doherty (2013), the secondary task of Ng et al. (2013) resembled the RS test and therefore could have taxed WMC in a more similar manner, which could be the reason for the significant group difference observed here.

With respect to the perceptual consequences of the various signal changes, we observed a positive effect of restored HFA on VRT performance, which was in line with our expectations. More precisely, participants were able to reach the performance level they had achieved with S_{off}N_{off} processing, thereby overcoming the VRT performance decrement caused by S_{on}N_{on} processing. Previously, we had speculated that this decrement could be explained based on the findings of McCreery and Stelmachowicz (2013) summarized above.

Broadly speaking, the effects of S_{on}N_{on} processing can be likened to those of a (shallow) low-pass filter that reduces HFA (see Signal Changes). Presumably, reduced HFA makes it more difficult to discern consonant sounds, resulting in a need for more top-down processing (as indexed by the longer reaction times) to be able to understand a given speech signal. We consider it likely that both the increase in VRTs due to S_{on}N_{on} processing and the decrease in VRTs due to S_{on+g}N_{on+g} processing can be interpreted in this manner.

Unlike the other outcomes, VRT performance was unaffected by SD and NAD. In our previous study, we had observed good test-retest reliability for the VRT measure. It therefore seems unlikely that too much random variability in the data from the current study was the reason for the absent effects. Given that, relatively speaking, the perceptual effects of SD and NAD were not as prominent as those of HFA (see Figure 2) and that group differences were not apparent in the VRT scores either, we tentatively conclude that our VRT measure was insufficiently sensitive with respect to these effects.
Regarding the results from our correlation analysis, we found the VRT_{rel} scores to be correlated with the RAU scores, but neither with the effort nor the preference scores. Although dual-task paradigms are commonly assumed to provide indirect indices of perceived effort, recent studies indicate that they do not tap the same aspects of listening effort as subjective ratings (Desjardins & Doherty 2013; Gosselin & Gagné 2011). Our results are in agreement with this. Future research should therefore focus on the development of objective measures that are able to capture perceived listening effort more accurately.

**Overall Preference**

In our previous study, we had found smaller WMC to be associated with a preference for strong over moderate NR processing. We had hypothesized that this was indicative of our C- listeners being more disturbed by the noise than by the distortions introduced by the binaural NR algorithm, whereas for our C+ listeners the opposite would be the case. In the current study, we had therefore expected the C- groups to more strongly prefer NAD than the C+ groups, and the C+ groups to more strongly dislike SD than the C- groups. However, our analyses did not reveal such effects – neither with respect to WMC nor with respect to PTA.

In order to maximize the effects of the signal changes under investigation, we had based our processing conditions on the strong rather than the moderate NR setting used previously. In our previous study, all of our groups had preferred some over no NR processing, and there had been a tendency for the C+ listeners to prefer moderate over strong NR processing across the range of SNRs used in the current study. We therefore consider it likely that the lack of any group effects in the preference data from the current study was due to the fact that our participants had not been able to express a preference for weaker NR processing, which would have led to less noise attenuation but at the same time to less target speech degradation.
Neher et al. Noise reduction, hearing loss, and working memory

With respect to the perceptual consequences of the three types of signal changes, we observed that restored HFA clearly reduced overall preference. This implies that previous related research findings obtained with music and movie stimuli (Franks 1982; Ricketts et al. 2008; see Introduction) do not necessarily generalize to speech-in-noise signals. In the case of our stimuli, restored HFA provided comparatively little additional target speech audibility while increasing the level of the cafeteria noise (see Figure 2). As a result, listening comfort presumably worsened, resulting in considerably weaker preference for $S_{on+g}N_{on+g}$ (relative to $S_{on}N_{on}$) processing. Regarding NAD, the results we obtained were qualitatively very similar to those seen in the RAU scores. To recapitulate, NAD led to an increase in overall preference when both $S$ and $N$ were subjected to ‘on+g’ processing, and to an even clearer increase in overall preference when $S$ was left unprocessed. These results are consistent with the interpretation based on the influence of co-modulation distortion provided above in relation to the RAU scores.

We also found a correlation between the preference and LES scores. Out of all the correlations observed this one was the largest, although it was still relatively weak ($r = -0.33$). Nevertheless, this result implies that preference for the different processing conditions was partly driven by the perception of effort. This interpretation is also supported by the finding that, qualitatively speaking, the effects of the three types of signal changes were rather similar for the preference and effort scores.

**Perceived Listening Effort**

Our motivation for including (subjective) listening effort scaling was to have a benchmark for our (objective) VRT measure. Consistent with literature data, the scores from these two measures were unrelated, implying that they capture different perceptual aspects. Furthermore, our analyses revealed no effects of listener group ($\eta^2_p = 0.05$; observed power =
0.16). This is at odds with other effort rating data obtained with speech-in-noise stimuli which showed that, for a given SNR, elderly HI listeners experience more effort than elderly NH listeners (Larsby et al. 2005). Nevertheless, these data were obtained at one SNR only and the possibility for differences in cognitive skills among groups was not considered, thus making a direct comparison difficult.

As pointed out above, the perceptual consequences of the various signal changes on perceived effort were very similar to those seen in the RAU and preference scores. One notable exception to this was that for the effort scores we observed a statistically significant interaction between processing condition and listener group. To recapitulate, this was due to the G_{Hi-C} group benefitting less from S_{on}N_{on} processing than the other groups of listeners. Furthermore, even though the (non-significant) interaction between listener group, processing condition, and SNR was characterized by a fairly small effect size ($\eta^2_p = 0.08$) its observed statistical power was rather high (0.72). Again, therefore, a slightly larger sample would probably have changed these results.

Implications

The current study indicated a negligible influence of PTA and WMC on response to binaural NR effects. This was at odds with our earlier study, which had revealed a dependency of preferred NR strength on WMC. It would thus be relevant to investigate the influence of NR strength on these results. It would also be relevant to perform a priori power analyses for adequate sample size determination. Furthermore, we observed that restored HFA was beneficial for VRT performance, but harmful for perceived effort and overall preference. In the current study, we chose to restore HFA fully in order to maximize its effects. It is possible that partial HFA restoration would result in a more favorable balance across outcome measures. Furthermore, our results showed that co-modulation distortion reduces binaural NR benefit. Consequently, it would be of interest to investigate ways of
minimizing this distortion, for example by pre-processing the stimuli with a directional microphone that reduces the level of (non-frontal) interfering signal components and thus their impact on the resultant NR gains. Future work should also consider other HA algorithms and user factors. In the current study, we only considered one particular type of (binaural) NR and one particular measure of cognitive function. It is possible that other types of HA signal processing (e.g., monaural NR) and other cognitive factors (e.g., processing speed) would produce different results.

ENDNOTES

1 Currently, this algorithm is not available in commercial HAs. However, due to its modest requirements regarding the exchange of information across HAs it could in principle be implemented in current binaurally linked devices.

2 Full (rather than partial) compensation of the long-term stimulus spectrum was carried out to maximize the effects of restored HFA in the S_{on+g}N_{on+g} condition and to ensure that in the S_{off}N_{on+g} and S_{on+g}N_{off} conditions outcome would not be affected by the processed (‘on+g’) signals being less audible than the unprocessed (‘off’) signals at mid to high frequencies.

3 ΔAI-SNR was calculated by first estimating the SNR improvement in one-third octave bands and then taking the scalar product of these estimates and the one-third octave band importance function from the Speech Intelligibility Index (ANSI 1997).

4 To quantify the relative magnitude of co-modulation distortion in S and N, we computed the standard deviation of the absolute values of the gain changes corresponding to the 1.25 to 3.8 sec segment of the stimulus shown in the third and fourth panel of Figure 2. For the data in the third panel (i.e., the gain changes applied to N) we obtained a value of
2.1 dB, while for the data in the fourth panel (i.e., the gain changes applied to S) we obtained a value of 0.9 dB.

The means (and standard deviations) of the VRT_{base} scores were 0.68 sec (0.06 sec), 0.74 sec (0.08 sec), 0.71 sec (0.08 sec), and 0.66 sec (0.04 sec) for the G_{H+C+}, G_{H-C+}, G_{H+C-}, and G_{H-C-} groups. The corresponding values for the VRT_{abs} scores were 0.73 sec (0.10 sec), 0.79 sec (0.09 sec), 0.81 sec (0.11 sec), and 0.73 sec (0.10 sec).

ACKNOWLEDGEMENTS

The authors thank their colleagues at the Hörzentrum Oldenburg for their help with recruiting participants and Sanja Rennebeck for her help with the data collection. This research was supported by the DFG Cluster of Excellence EXC 1077/1 “Hearing4all”.

REFERENCES


Neher et al.  Noise reduction, hearing loss, and working memory


Neher et al. Noise reduction, hearing loss, and working memory


Neher et al. Noise reduction, hearing loss, and working memory


46
FIGURES

Figure 1. Block diagram of the simulated hearing aid. L, left channel; R, right channel; S, target speech; N, background noise; SN, signal mixture; L_{SNR}, gain factor for signal-to-noise ratio (SNR) adjustment; NR, noise reduction; k, frequency band index; L(k), level estimator; \( \Delta L_{LR}(k) \), gains for long-term spectrum restoration; G_{LR}(k), NR gains; eq., equalization.
Figure 2. Graphical illustration of the effects of $S_{off}N_{off}$, $S_{on}N_{on}$, $S_{on+g}N_{on+g}$, $S_{off}N_{on+g}$, and $S_{on+g}N_{off}$ processing on (one channel of) an example stimulus with an input signal-to-noise ratio of 4 dB. The panels on the left-hand side show the time waveforms of the target speech, S (black) and the cafeteria noise, N (gray). The panels on the right-hand side show the difference in gain between $S_{off}N_{off}$ and either $S_{on}N_{on}$, $S_{on+g}N_{on+g}$, $S_{off}N_{on+g}$, or $S_{on+g}N_{off}$ as a function of time and frequency. a.u., arbitrary units.
Figure 3. (a) Mean rationalized arcsine unit (RAU) scores for the four groups of listeners as a function of signal-to-noise ratio (SNR). (b) Mean RAU scores for the five processing conditions averaged across the two SNRs. Error bars show 95% confidence intervals. Horizontal bars denote statistically significant processing conditions with the primary experimental contrasts being shown in black (* \( p < 0.05 \), ** \( p < 0.01 \), *** \( p < 0.001 \), **** \( p < 0.0001 \), ***** \( p < 0.00001 \)). SD, speech distortion; NAD, noise attenuation and distortion.
Figure 4. (a) Mean relative visual reaction time (VRT$_{rel}$) scores for the four groups of listeners as a function of signal-to-noise ratio (SNR). (b) Mean VRT$_{rel}$ scores for the five processing conditions averaged across the two SNRs. Error bars show 95% confidence intervals. Horizontal bars denote statistically significant processing conditions with the primary experimental contrasts being shown in black (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

HFA, high-frequency audibility.
Figure 5. (a) Mean overall preference scores for the five processing conditions as a function of signal-to-noise ratio (SNR). (b) Mean overall preference scores for the five processing conditions averaged across 0 and 4 dB SNR. Error bars show 95% confidence intervals. Horizontal bars denote statistically significant processing conditions with the primary experimental contrasts being shown in black (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$, ***** $p < 0.00001$). HFA, high-frequency audibility; NAD, noise attenuation and distortion; PC, processing condition.
Figure 6. (a) Mean listening effort scaling scores for the five processing conditions as a function of signal-to-noise ratio (SNR). (b) Mean listening effort scaling scores for the five processing conditions averaged across the three SNRs. Error bars show 95% confidence intervals. Horizontal bars denote statistically significant processing conditions with the primary experimental contrasts being shown in black (* $p < 0.05$, ** $p < 0.01$, **** $p < 0.001$, ***** $p < 0.0001$). HFA, high-frequency audibility; SD, speech distortion; NAD, noise attenuation and distortion; PC, processing condition.
TABLES

Table 1. Group means in terms of age, pure-tone average hearing loss (PTA), and reading span (RS). Data in parentheses correspond to minimum and maximum values. \( N = 10 \) per group.

<table>
<thead>
<tr>
<th>Group</th>
<th>Age (yr)</th>
<th>PTA (dB HL)</th>
<th>RS (%-correct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G_{H+C+} )</td>
<td>73.0 (60-80)</td>
<td>36.8 (23-46)</td>
<td>44.6 (35-54)</td>
</tr>
<tr>
<td>( G_{H-C+} )</td>
<td>75.0 (71-84)</td>
<td>55.4 (51-65)</td>
<td>44.1 (35-56)</td>
</tr>
<tr>
<td>( G_{H+C-} )</td>
<td>76.6 (69-82)</td>
<td>38.8 (29-49)</td>
<td>28.1 (20-33)</td>
</tr>
<tr>
<td>( G_{H-C-} )</td>
<td>75.5 (66-82)</td>
<td>58.7 (50-69)</td>
<td>27.4 (19-33)</td>
</tr>
</tbody>
</table>