Hearing aid noise suppression and working memory function

Tobias Neher\textsuperscript{a,b,c}, Kirsten C. Wagener\textsuperscript{c} & Rosa-Linde Fischer\textsuperscript{d}

\textsuperscript{a} Medizinische Physik and Cluster of Excellence “Hearing4all”, Dept. of Medical Physics and Acoustics, Carl-von-Ossietzky University, D-26111 Oldenburg, Germany

\textsuperscript{b} Institute of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, DK-5230 Odense, Denmark, Email: tneher@health.sdu.dk

\textsuperscript{c} Hörzentrum Oldenburg GmbH, Marie-Curie-Str. 2, D-26129 Oldenburg, Germany, Phone: +49 441 2172-143, Email: k.wagener@hoerzentrum-oldenburg.de

\textsuperscript{d} Sivantos GmbH, Henri-Dunant-Str. 100, D-91058 Erlangen, Germany, Phone: +49 9131 308-3550, Email: rosa-linde.fischer@sivantos.com

\textsuperscript{e} Corresponding author

Running title: Noise suppression and working memory

Date of re-submission: 10\textsuperscript{th} October 2017
Neher: Noise suppression and working memory

Acronyms

ANOVA = Analysis of variance

BTE = Behind-the-ear

HA = Hearing aid

PTA4 = Pure-tone average hearing loss calculated across 0.5, 1, 2 and 4 kHz and left and right ears

SNR = Signal-to-noise ratio

ΔAI-SNR = Speech-weighted SNR improvement
Neher: Noise suppression and working memory

Keywords

Hearing aids; noise reduction; cognition; working memory; individual differences
Neher: Noise suppression and working memory

Abstract

Objective: Research findings concerning the relation between benefit from hearing aid (HA) noise suppression and working memory function are inconsistent. The current study thus investigated the effects of three noise suppression algorithms on auditory working memory and the relation with reading span.

Design: Using a computer simulation of bilaterally fitted HAs, four settings were tested: (1) unprocessed, (2) directional microphones, (3) single-channel noise reduction and (4) binaural coherence-based noise reduction. Settings 2-4 were matched in terms of the speech-weighted signal-to-noise ratio (SNR) improvement. Auditory working memory was assessed at +6 dB SNR using listening span and N-back paradigms.

Study sample: Twenty experienced HA users aged 55-80 years with large differences in reading span.

Results: For the listening span measurements, there was an influence of HA setting on sentence-final word recognition and recall, with the directional microphones leading to ~6% better performance than the single-channel noise reduction. For the N-back measurements, there was substantial test-retest variability and no influence of HA setting. No interactions with reading span were found.
Conclusions: HA noise suppression may affect the recognition and recall of speech at positive SNRs, irrespective of individual reading span. Future work should improve the reliability of the auditory working memory measurements.
Neher: Noise suppression and working memory

1 Introduction

Over the last decade, audiologists have become increasingly interested in the relation between speech perception and working memory. This is because working memory has been found to be an effective predictor of the speech-in-noise abilities of hearing-impaired listeners (e.g. Akeroyd, 2008; Rönnberg et al, 2010). Briefly, working memory relates to the ability to manipulate and store sensory information and is required for complex cognitive tasks such as speech-in-noise reception (e.g. Baddeley, 1986). The processes underlying this ability are thought to draw upon a common set of mental resources, which can be allocated according to the task demands at hand but which are also limited in capacity. Consequently, if more processing is required for a given task, less capacity can be allocated to information storage and vice versa.

Research has shown that verbal working memory capacity, which is commonly measured using a reading span test (e.g. Daneman & Carpenter, 1980), can vary widely among hearing-impaired individuals (see Akeroyd, 2008 for a review). This has raised the question of possible consequences for auditory rehabilitation, including the fitting of hearing aids (HAs). To date, several studies have dealt with the influence of reading span on outcome from different HA processing strategies including noise suppression (for reviews, see Lunner et al, 2009; Souza et al, 2015). The purpose of noise suppression is to improve the signal-to-noise ratio (SNR) – and thus speech perception – by attenuating noise-dominated signal components. Not only will this lead to the removal of noise, it will also distort the target signal whenever speech-dominated components are erroneously processed (e.g. Kates, 2008). In general, noise attenuation can be expected to release mental resources because less noise will make it easier for the cognitive system to handle the target signal, whereas speech distortions will degrade the target signal and thus tax the cognitive system (e.g. Rabbitt, 1966).
Neher: Noise suppression and working memory

It has been suggested that individually available working memory resources may determine the extent to which speech distortions affect benefit from HA signal processing (Lunner et al, 2009). That is, individuals with smaller working memory capacity may be more susceptible to speech distortions than individuals with larger working memory capacity. While some support for this exists in the context of fast-acting dynamic range compression (which also introduces speech distortions), the picture is unclear for noise suppression (Souza et al, 2015). To illustrate, Ng et al (2013) evaluated the effects of a binary mask-based noise suppression algorithm with a group of HA users (ages 32-65; mean: 59 years) who differed in reading span. Outcome was assessed using an auditory memory test dubbed ‘sentence-final word identification and recall’ that required the participants to recognise the final words of a sequence of sentences presented in background noise and to recall them afterwards. Data analyses showed a main effect of reading span on recall performance, with better recall being related to longer reading span. Furthermore, an interaction between reading span and noise suppression benefit was observed. That is, participants with longer reading span recalled more words than participants with shorter reading span as a result of noise suppression processing. In other words, the noise suppression benefit in word recall was larger for participants with long reading span than for participants with short reading span.

In a follow-up study based on essentially the same experimental paradigm and algorithm, Ng et al (2015) replicated the main effect of reading span on recall performance using a presumably different group of HA users (ages 56-65; mean: 62 years). Furthermore, they observed an interaction between reading span, noise suppression and serial word position. That is, participants with shorter reading span achieved better recall due to noise suppression for words that occurred towards the end of the sequence of sentences, whereas participants with
longer reading span achieved better recall irrespective of where in the sequence the words occurred.

While the findings of Ng et al speak in favour of a relation between reading span and noise suppression outcome, those of other researchers do not. To illustrate, Arehart et al (2015) tested a group of hearing-impaired listeners in terms of speech recognition and speech quality with the algorithm also used by Ng et al. Even though reading span was a significant predictor of overall speech recognition, no interaction with noise suppression outcome was found. Using a dual-task paradigm that combined speech recognition with a visual tracking task, Desjardins & Doherty (2014) investigated the effects of modulation-based noise suppression with a group of HA users. While in some conditions noise suppression improved visual tracking performance, no relation with reading span was found. In a series of studies, Neher and co-workers (2014; 2014a; 2014b) investigated the effects of binaural coherence-based noise suppression using measures of speech recognition, visual response time and preference. Their participants were groups of elderly HA users with carefully controlled variation in reading span. Across the three studies, there was a main effect of reading span on overall speech recognition, but no interaction with noise suppression outcome emerged.

To summarise, several studies have addressed the link between benefit from HA noise suppression and working memory function. In these studies, different algorithms were tested using a wide range of outcome measures. In the studies that used auditory working memory tasks for assessing the effects of noise suppression interactions with reading span were found, whereas in the studies that relied on other tasks (e.g. visual tracking or response tasks) this was not the case. A possible explanation for this could be that the resemblance of the auditory working memory and reading span tasks boosted the association between the two types of measures,
Neher: Noise suppression and working memory

whereas the other outcome measures engaged other processing abilities and thus did not lead to associations with reading span.

The purpose of the current study was to extend the findings summarised above in a number of ways. First, we investigated the effects of several different noise suppression schemes on speech recognition and auditory verbal working memory as well as the interaction with reading span. The noise suppression schemes that we used were carefully matched in terms of SNR improvement. The auditory working memory measurements were based on two different paradigms. Second, we also collected test-retest data for the auditory working memory measurements. In this manner, we wanted to address the generalizability and reliability of any observable effects. Using a computer simulation of bilaterally fitted HAs, we tested three different algorithms with a group of HA users who exhibited large differences in reading span. The algorithms included a directional microphone setting that attenuated off-axis noise while leaving the on-axis target speech essentially untouched. The other two (non-directional) algorithms resulted in approximately the same amount of noise attenuation as the directional microphone setting but also added some speech distortions. For the assessment of auditory working memory, we employed two measures: listening span and $N$-back. Listening span is essentially the auditory equivalent of reading span. $N$-back is an established measure for assessing the executive component of working memory (e.g. Jaeggi et al, 2010). Doherty & Desjardins (2015) used both types of measures to investigate the benefit of amplification for 24 listeners aged 50-74 years with mild-to-moderate hearing losses. For all of their participants, they observed that performance on the two measures improved with amplification, thereby demonstrating the basic feasibility of making such measurements with hearing-impaired adults. We therefore adopted the measures of Doherty and Desjardins for our purposes.
In view of the results summarised above, we expected to find an interaction between reading span and benefit from the different noise suppression schemes on auditory working memory performance, as assessed using the listening span and N-back measures. Given that the listening span task that we used resembled the paradigm of Ng et al (2013; 2015), we particularly expected to find such an interaction for the listening span measurements. Furthermore, we anticipated that the directional microphone setting would generally improve auditory working memory (due to it attenuating noise without the added expense of concurrent speech distortions) and that the two other noise suppression algorithms would perhaps do so, too.

2 Methods

The current study was approved by the ethics committee of the University of Oldenburg. All participants provided written informed consent and received financial compensation.

2.1 Participants and reading span measurements

The participants were recruited from a cohort of several hundred hearing-impaired listeners belonging to the database of the Hörzentrum Oldenburg, Germany. Our target was a group of 20 participants fulfilling the following selection criteria: (1) native German speakers, (2) sloping bilateral sensorineural hearing loss, (3) asymmetry in air-conduction thresholds of no more than 15 dB HL across ears for the standard audiometric frequencies from 0.5 to 4 kHz, (4) air-bone gaps of no larger than 15 dB HL at any audiometric frequency between 0.5 and 4 kHz, (5) at least nine months of bilateral HA experience, (6) self-reported normal or corrected-to-normal vision, (7) no history of any psychiatric disorders (e.g. depression), and (8) a speech recognition score of at least 80%-correct on an initial screening test (Sect. 2.4).
In addition to these criteria, our goal was to achieve as large a spread as possible in reading span. Using the German reading span test implementation of Carroll et al. (2015), we presented our participants with short sentence segments one at a time at a rate of one segment per 0.8 sec. After three such segments, there was a pause of 1.75 sec, during which the participants had to respond either ‘yes’ if the previous three segments made up a semantically correct sentence (e.g. ‘The girl–sang–a song’) or ‘no’ if the previous three segments made up a semantically absurd sentence (e.g. ‘The bottle–drank–water’). Following a sequence of three, four, five or six sentences, the participants were prompted to recall the first or final words of all three, four, five or six previous sentences in any order. As the outcome measure, we used the percentage of correctly recalled words across a total of 54 trials.

The participants chosen for the current study were aged 55-80 years (mean: 72 years). Figure 1 shows their audiograms averaged across left and right ears together with the grand average audiogram. In terms of pure-tone average hearing loss calculated across 0.5, 1, 2 and 4 kHz and left and right ears (PTA4), the participants ranged from 33-61 dB HL (mean: 49 dB HL). Their reading span scores ranged from 26-70% correctly recalled words (mean: 44%), which constitutes a rather large spread (cf. Souza et al, 2015, their Tables 1-3). The age, PTA4 and reading span data were not correlated with each other (all absolute Pearson’s $r$ correlation coefficients $< 0.23$, all $p > 0.3$).

[Insert Figure 1 about here]

2.2 Physical test setup

The auditory tests were carried out in a soundproof booth. Audio playback was via an Auritec (Hamburg, Germany) Earbox Highpower soundcard and a pair of Sennheiser
Neher: Noise suppression and working memory

(Wennebostel, Germany) HDA200 headphones. Calibration was carried out using a Brüel & Kjær (B&K; Nærum, Denmark) 4153 artificial ear, a B&K 4134 1/2” microphone, a B&K 2669 preamplifier and a B&K 2610 measurement amplifier.

The reading span test was carried out in a quiet well-lit room with the computer screen displaying the stimuli positioned ~0.5 m in front of the participant's face.

2.3 Hearing aid settings

The HA settings were based on a computer simulation of bilaterally fitted HAs. This simulation was realised with the help of the Master Hearing Aid research platform of Grimm et al (2006). It included noise suppression, linear amplification and headphone equalisation, and was carried out at a sampling rate of 16 kHz. Before presentation, all stimuli were resampled to 44.1 kHz.

2.3.1 Input signals

For the creation of the HA input signals, we employed publicly available recordings made in a large cafeteria with a head-and-torso simulator equipped with two behind-the-ear (BTE) HA dummies (Kayser et al, 2009). Specifically, we spatialised our target signals by convolving them with impulse response measurements made with the front and rear BTE microphones of the two HA dummies for a source positioned at 0° azimuth and 1-m distance. For the masker signal, we used a recording made with the front and rear BTE microphones during a busy lunch hour. This recording, which is several minutes in length, is characterised by spatially diffuse speech babble, occasional parts of intelligible speech from nearby speakers as well as sporadic transient sounds from cutlery, dishes and chairs. To obtain the HA input signals, we mixed the resultant target and masker signals at a given SNR (Sect. 2.4).
2.3.2 Noise suppression

In total, we tested four HA settings: *unproc*, *dirmic*, *scnr* and *bcf*. The settings were based on three noise suppression algorithms that differed in their design principles and thus their effects on the input signals.

The *unproc* setting served as a reference condition. It corresponded to the two front BTE microphone signals without any noise suppression.

The *dirmic* setting corresponded to two unilateral forward-facing directional microphones with cardioid polar patterns. These were realised by processing the front and rear BTE microphone signals of each HA dummy with a simple delay-and-sum beamformer algorithm (e.g. Dillon, 2012). To compensate for the high-pass characteristic that is typical of directional microphones, we applied a finite impulse response filter to the output of each directional microphone. This filter ensured that the *dirmic* setting was spectrally matched to the *unproc* setting in the 0° direction. We then applied another (two-channel) finite impulse response filter to the output signals of the two directional microphones. This filter ensured that the *dirmic* setting was also matched to the *unproc* setting in terms of its interaural (or spatial) characteristics in the 0° direction (see Neher, 2014 for details).

The *scnr* setting corresponded to a single-channel noise reduction algorithm after Gerkmann & Hendriks (2012). This algorithm, which was applied independently to the two front BTE microphone signals, derives short-time Fourier transform-domain estimates of the noise power spectral density and the speech power by means of a speech presence probability estimator and temporal cepstrum smoothing (Breithaupt et al, 2008). Based on the resultant speech and noise power estimates, gains are derived and applied to *remove noise from* the input signal, after which the time-domain signal is resynthesised using a standard overlap-add
Neher: Noise suppression and working memory

procedure. As is typical of any single-channel noise reduction algorithm (e.g. Kates, 2008), the
scnr setting caused speech distortions whenever it misclassified target signal components as
noise.

The bcf setting corresponded to the binaural coherence filter-based noise reduction
algorithm tested by Neher et al (2014; 2014a; 2014b). This algorithm, which was applied jointly
to the two front BTE microphone signals, estimates the binaural coherence (or interaural
similarity) of the left and right input signals across time and frequency (Grimm et al, 2009).
Based on these estimates, it derives gains for the attenuation of interaurally dissimilar (or
incoherent) signal components and applies them jointly to the left and right input signals. Due to
diffraction effects around the head, the interaural coherence will always be high below ~1 kHz,
which is why the efficacy of this algorithm is restricted to higher frequencies. Similar to the scnr
setting, the bcf setting caused speech distortions whenever target signal components exhibited
reduced binaural coherence and were thus attenuated by the algorithm.

To enable a comparison of the effects of the dirmic, scnr and bcf settings, we
parameterised the scnr and bcf algorithms such that they resulted in approximately the same
speech-weighted SNR improvement (ΔAI-SNR) as the dirmic setting. For the scnr algorithm, we
therefore set the maximal attenuation for noise-dominated signal components to ~8 dB. For the
bcf algorithm, we chose a gain function that resulted in a maximal attenuation of ~20 dB for
fully incoherent sounds (cf. Neher & Wagener, 2016, their Figure 1). To quantify ΔAI-SNR, we
generated ~3 min of stimulus material per test condition and applied the gains calculated based
on the signal mixture separately to the speech and noise signals (the so-called shadow-filtering
method). We then estimated the SNR improvement relative to the unproc condition in one-third
octave bands and took the scalar product of these estimates and the one-third octave band
Neher: Noise suppression and working memory

importance function from the Speech Intelligibility Index (ANSI, 1997). Table 1 summarises the results. On average, the three noise suppression settings led to ΔAI-SNRs of ~3 dB. Across the two target speech materials used here (Sect. 2.4), there were differences in ΔAI-SNR of up to 1 dB. These were caused by spectro-temporal differences among these materials.

[Insert Table 1 about here]

Figure 2 illustrates the physical effects of the dirmic, scnr and bcf settings for an example stimulus. The panels on the left-hand side show the waveforms of the speech and noise signals at the output of each HA setting. The panels on the right-hand side show the level differences between the unproc condition and each of the other three HA conditions. As can be seen, the dirmic, scnr and bcf settings differed markedly in the applied gains. The dirmic setting attenuated off-axis signals irrespective of their spectro-temporal properties. The scnr algorithm resulted in a rather constant amount of attenuation for noise-dominated signal segments. The bcf algorithm attenuated incoherent signal components in the mid-to-high frequency range, with the applied gains varying with the estimated binaural coherence.

[Insert Figure 2 about here]

2.3.3 Linear amplification and headphone equalisation

To provide audibility similar to many clinical HA fittings, we spectrally shaped all stimuli according to the ‘National Acoustic Laboratories – Revised Profound’ fitting rule (National Acoustic Laboratories, Sydney, Australia; e.g. Dillon, 2012). Specifically, for each participant we determined the gain required at 0.25, 0.5, 1, 1.5, 2, 3, 4 and 6 kHz and mapped the resultant
values onto the filterbank of the Master Hearing Aid using interpolation techniques. Furthermore, we processed the left and right output channels with a finite impulse response filter that compensated for the uneven magnitude response of the headphones used for the stimulus presentation (Sect. 2.2).

2.4 Auditory working memory measurements

To assess auditory working memory, we employed listening span and N-back measures akin to those used by Doherty & Desjardins (2015). For both types of measurements, we calibrated the target speech (see below) and cafeteria noise (Sect. 2.3.1) signals to nominal sound pressure levels of, respectively, 71 and 65 dB in the unproc condition. As a consequence, we performed all measurements at a fixed SNR of +6 dB. In their study, Doherty & Desjardins (2015) had used an SNR of +8 dB. We decided to use a slightly lower SNR to accommodate the SNR improvement brought about by our noise suppression settings (Table 1).

To ensure that +6 dB SNR led to good speech recognition for all participants, we initially carried out a screening test. Following a training run based on the dirmic setting and 12 trials per target speech material, we measured speech recognition with the unproc setting using 24 trials per target speech material. To be included in the study, a participant had to achieve at least 80%-correct speech recognition for each type of target speech material. On average, the 20 participants included here achieved 94%-correct recognition for the listening span material (range: 83-100%) and 97%-correct recognition for the N-back material (range: 88-100%).

To assess test-retest reliability, we performed two measurements per auditory working memory measure and test condition.
2.4.1 Listening span measurements

For the listening span measurements, we used the (German) sentence material from the Basle sentence test (Tschopp & Züst, 1994; Niklès & Tschopp, 1996) as target signals. The Basle sentence test resembles the Revised Speech Perception in Noise test (Bilger et al, 1984). It consists of 150 low-context sentences for which the sentence-final word (always a monosyllabic noun) is not predictable based on the preceding words (e.g. ‘The party is against a strike’) and 150 high-context sentences for which the sentence-final word is predictable (e.g. ‘The windows are made of glass’). Each set of 150 sentences is distributed across 10 test lists of 15 sentences each. Half of these lists are spoken by a male speaker and the other half by a female speaker. The sentence-final words were chosen from a German word frequency dictionary of spoken words (Ruoff, 1990). The sentences were generated by following various criteria to maximise the perceptual homogeneity of the material (Tschopp & Ingold, 1992). For both the low- and the high-context sentences, this resulted in sentence-final words with comparable intelligibility for normal-hearing listeners when the entire sentences were masked by noise with a constant presentation level (as also done in the current study).

In the current study, we tested each HA setting with both low- and high-context sentences. For each measurement, we distributed 12 sentences from a given test list across three blocks of 2, 4 or 6 sentences, consistent with the approach of Doherty & Desjardins (2015). We then presented these blocks of sentences with 4-sec inter-sentence intervals against continuous cafeteria noise. Following the presentation of a sentence, the participants had to repeat aloud the entire sentence and to memorise the sentence-final word for later recall. At the end of a block, the noise stopped and the participants were prompted to recall as many of the sentence-final words as possible in any order. An experimenter scored the responses in terms of both
Neher: Noise suppression and working memory

recognition and recall accuracy. In case a participant reported a word during the recall phase that was erroneously reported during the repetition phase, the recall response was scored as correct. Performance was calculated as the percentage of sentence-final words recognised or recalled correctly. If a given sentence-final word was not repeated at all by a participant, this word was not included in the percent-correct calculations for the recall measure.

Prior to the actual measurements, all participants completed a practice run in the unproc condition that comprised 12 sentences distributed across three blocks of 2, 4 or 6 sentences. After completing this run, all participants were able to perform the task reliably. The sentences used for the training were not reused for the actual measurements.

2.4.2 N-back measurements

For the N-back measurements, we used the monosyllabic digits (i.e. 1, 2, 3, 4, 5, 6, 8 and 9) from the German digit triplet test (Buschermöhle et al, 2014), which were spoken by a female speaker, as target signals. For each measurement, we presented 25 randomly chosen digits with 2-sec inter-digit intervals against continuous cafeteria noise. The task of the participant was to memorise each digit and to repeat aloud either the previous digit (1-back task) or the second-to-last digit (2-back task) while the presentation of the digits continued. An experimenter scored the participant’s responses. Performance was calculated as the percentage of digits repeated correctly.

Prior to the actual measurements, all participants completed several training runs. All of these runs were performed in the unproc condition and comprised 10 digits. Initially, the participants practiced the 1-back task until they felt comfortable with it, after which they continued with the 2-back task. Despite multiple attempts, two participants were unable to master any of the N-back measurements. An additional participant could only master the 1-back
Neher: Noise suppression and working memory

task, while a fourth participant could only carry out the 2-back task during the retest but not the
test measurements (Sect. 2.5). As a consequence, we performed the 1-back measurements with
18 participants and the 2-back measurements with 17 participants.

2.5 Test protocol

The various measurements were distributed across three visits of maximally 1.5 hr each,
including breaks. At the first visit, the audiogram was measured and the reading span (Sect. 2.1)
and screening measurements (Sect. 2.4) were carried out. At the second and third visit, the
listening span and N-back measurements were performed. The measurements at the third visit
served as retest measurements of those made at the second visit. The third visit was scheduled to
take place approximately one week after the second visit.

2.6 Preparatory analyses and statistical tests

For the listening span measurements, there were relatively few numerical outcomes,
especially for a block size of 2 (i.e. 0, 1 or 2 correctly recognised/recalled words). Given that
these data also exhibited ceiling effects (cf. Figure 3), we excluded them from the statistical
analyses. Concerning the data obtained for block sizes of 4 and 6, we pooled them across the
low- and high-context sentences to increase the range of possible numerical outcomes and to
normalise the variance of the datasets. To assess test-retest reliability, we then calculated
Pearson’s r correlation coefficient for the two sets of sentence-final word recall scores per HA
setting. We observed correlation coefficients of 0.24, 0.44, 0.49 and 0.59 (block size of 4) and
0.60, 0.65, 0.71 and 0.47 (block size of 6) for the unproc, dirmic, scnr and bcf setting,
respectively. Thus, for a block size of 6 there were moderate-to-strong test-retest correlations,
whereas for a block size of 4 the correlations were overall somewhat weaker.
Neher: Noise suppression and working memory

For the N-back measurements, we observed ceiling effects for the 1-back task (cf. Figure 4) and substantial test-retest variability for the 2-back task. Across our participants, the difference in the percentage of errors made between the test and retest measurements ranged from 0-14% for the 1-back task (mean: 5%) and from 10-43% for the 2-back task (mean: 20%). The rather poor reliability of the 2-back data was also reflected in the weak correlation coefficients that we observed for the two sets of correctly repeated digits (0.26, 0.23, -0.16 and -0.19 for the unproc, dirmic, scnr and bcf settings, respectively).

For all subsequent analyses, we averaged the two available measurements per participant and test condition. To analyse the listening span data, we performed repeated-measures analyses of variance (ANOVAs) with block size (4, 6) and HA setting (unproc, dirmic, scnr, bcf) as within-subject factors. To determine whether any observable effects were driven by individual differences in verbal working memory capacity, we included reading span as a covariate. Furthermore, we initially also included age and PTA4 to control for potentially confounding effects due to these characteristics. Because age did not contribute significantly to the statistical models, we excluded it from all subsequent analyses. Prior to these steps, we centred each covariate by subtracting the overall sample mean from the individual data points to leave the within-subject factor sum of squares unaltered (e.g. Fidell & Tabachnick, 2006). Because we observed no violations of the sphericity requirement, we left the ANOVA results uncorrected. To follow up on any significant effects, we performed either post hoc t-tests with Bonferroni correction (within-subject factors) or correlation analyses (covariates).

To analyse the N-back data, we performed two Friedman’s ANOVAs on the 1-back and 2-back measurements with HA setting (unproc, dirmic, scnr, bcf) as within-subject factor. To
Neher: Noise suppression and working memory

compare performance across the two tasks (1-back, 2-back), we performed a Wilcoxon signed-rank test.

3 Results

3.1 Listening span measurements

Table 2 summarises the results from the ANOVA carried out on the sentence-final word recognition scores. There were significant main effects of block size, HA setting, reading span and PTA4 (all $p < 0.05$), but no interactions (all $p > 0.2$). Our follow-up analyses revealed that sentence-final word recognition was better for a block size of 4 than for a block size of 6 (means: 84.4% vs. 72.6%; $p < 0.00001$) and with the dirmic setting than with the scnr setting (means: 81.6% vs. 75.5%; $p = 0.037$). Furthermore, we observed associations between sentence-final word recognition and both reading span ($r = 0.74$, $p < 0.001$) and PTA4 ($r = -0.47$, $p = 0.039$).

Figure 3a shows means and 95% confidence intervals of the recognition scores for the different block sizes and HA settings.

[Insert Table 2 and Figure 3 about here]

Table 2 also summarises the results from the ANOVA performed on the sentence-final word recall scores. There were significant main effects of block size, HA setting and reading span (all $p < 0.05$), but no interactions (all $p > 0.3$). Our follow-up analyses revealed that sentence-final word recall was better for a block size of 4 than for a block size of 6 (means: 72.6% vs. 51.7%; $p < 0.00001$) and with the dirmic setting than with the scnr setting (means: 64.8% vs. 59.5%; $p = 0.010$). Furthermore, we observed an association between sentence-final
word recall and reading span ($r = 0.63, p < 0.003$). Figure 3b shows means and 95% confidence intervals of the recall scores for the different block sizes and HA settings.

As pointed out in Sect. 1, Ng et al (2015) observed an effect of serial word position on the relation between reading span and noise suppression benefit. We therefore performed an additional ANOVA, for which we partitioned the recall data into primacy (first two list positions) and recency (last two list positions) components. We found significant effects of list position ($F_{1,17} = 42.1, p < 0.00001, \eta^2_p = 0.71$) and HA setting ($F_{3,51} = 3.6, p = 0.019, \eta^2_p = 0.19$), but no effect of reading span ($p > 0.6$) nor any interactions (all $p > 0.3$). Post-hoc analyses showed that sentence-final word recall was better for recency than for primacy list positions (means: 71.3% vs. 48.7%; $p < 0.00001$) and for the dirmic setting than for the scnr setting (means: 62.8% vs. 57.1%; $p = 0.004$).

3.2 N-back measurements

The two Friedman’s ANOVAs carried out on the N-back data revealed no influence of HA setting (1-back: $\chi^2_{(3)} = 0.45, p > 0.9$; 2-back: $\chi^2_{(3)} = 5.0, p > 0.1$). The Wilcoxon signed-rank test revealed better performance on the 1-back task than on the 2-back task (means: 95.8% vs. 75.5% correctly repeated digits; $T = 3.6, p < 0.001$). Figure 4 shows boxplots of the N-back data for the different tasks and HA settings.

[Insert Figure 4 about here]

4 Discussion

In the current study, we examined the relation between noise suppression benefit and working memory function. To that end, we evaluated the effects of three different noise
Neher: Noise suppression and working memory

suppression settings (dirmic, scnr, bcf) at +6 dB SNR using two measures of auditory working memory (listening span, N-back). We parameterised the algorithms such that they all resulted in ΔAI-SNRs of ~3 dB. Our participants were 20 experienced HA users aged 55-80 years with substantial differences in reading span. While all of them were able to carry out the listening span task, four participants struggled with the N-back task. Furthermore, while there were moderate-to-strong test-retest correlations for the listening span measurements made with a block size of 6, the correlations were somewhat weaker for the listening span measurements made with a block size of 4 and substantially weaker for all N-back measurements. Concerning the influence of HA setting, we found that the dirmic setting led to ~6% better recognition and recall of sentence-final words than the scnr setting. For the N-back measurements, we found no influence of HA setting. Furthermore, although reading span was positively associated with sentence-final word recognition and recall, it did not interact with the influence of HA setting. Below, we provide a discussion of these findings.

4.1 Influence of noise suppression

Consistent with our expectations, we observed an effect of the dirmic setting on final word recall. Specifically, the dirmic setting provided for a significant (~6%) improvement in memory performance relative to the scnr setting, whereas performance with the unproc, dirmic and bcf settings was comparable. As explained above, directional microphones can provide for noise attenuation without speech distortion, whereas single-channel noise reduction algorithms typically also degrade the target speech. Most likely, these well-known effects were responsible for the influence of HA setting observed here. The fact that we found a negative effect of the scnr setting but not the bcf setting suggests that the latter introduced fewer or less harmful artefacts into our stimuli.
Interestingly, we observed a very similar effect of HA setting on the final word recognition scores. This was despite the fact that we performed all measurements at a clearly positive SNR. With standard test materials and procedures, it is typically impossible to perform sensitive speech reception measurements in the positive SNR range because performance will be at ceiling. To address this issue, researchers have developed new materials and procedures characterised by increased task difficulty, for example by time-compressing the target speech (Schlueter et al, 2014) or by maximising the similarity between the target and interfering signals (Nielsen et al, 2014). In our case, we presumably observed an effect in the recognition scores because we collected these data in parallel with the recall scores. From the field of cognitive psychology, it is well-known that a secondary task (such as memorising sentence-final words) can reduce performance on a primary task (such as recognising these words). This is the basic premise of the dual-task paradigm, which is based on the theory that the brain has a limited capacity to process sensory information and that this capacity is allocated across tasks on an as-needed basis (Kahneman, 1973). If an individual has to carry out two tasks simultaneously, fewer resources will be available per task. In general, this can be expected to reduce performance relative to single-task measurements.

The close correspondence between our recall and recognition scores is consistent with the idea that if a speech signal can be accurately recognised and thus unambiguously matched to a phonological representation stored in long-term memory, this will facilitate retrieval (e.g. Rönnberg et al, 2008). Conversely, if the input signal is degraded because of noise or distortions, this will hamper identification and thus also memory performance. However, as discussed below, our results do not support the view that individually available working memory capacity can help overcome these effects, as postulated elsewhere.


4.2 Influence of reading span

For the recognition and recall data from the listening span measurements, we found a positive association with reading span (Sect. 3.1). This finding is consistent with several other studies that observed a general influence of working memory capacity on speech recognition or memory performance (cf. Souza et al., 2015). As such, our results underline the predictive power of reading span with respect to the speech-in-noise abilities of hearing-impaired adults.

Concerning the potential interaction between reading span and noise suppression benefit, our results did not lend any support to this. As apparent from Sect. 1, while this is broadly in line with the findings of a number of other studies that were obtained using other types of outcome measures, it is at odds with those of Ng et al. (2013; 2015) that were obtained using a listening span measure. There are a number of possible explanations for this. First, it could be that the influence of reading span is constrained to the effects of the binary mask-based algorithm tested by Ng et al. Depending on the underlying design principles, noise suppression algorithms can have rather different acoustic and perceptual effects (e.g. Smeds et al., 2010; Brons et al., 2013). In principle, the abilities indexed by the reading span test could impact a listener’s response to binary mask-based processing but not to other types of noise suppression (although we consider this unlikely). Second, it is possible that other methodological differences between our study and those of Ng et al. played a role here. For instance, our participants were generally older (means: 72 vs. ~60 years). Also, we presented our stimuli at a single fixed SNR, whereas Ng et al. used individualised SNRs targeting 95%-correct speech recognition. On average, however, our participants achieved a very similar performance level in the screening test (94% and 97% accuracy; Sect. 2.4). Also, our test SNR (+6 dB) was comparable with the mean test SNRs used in the two studies of Ng et al. (+4.1 and +7.5 dB). It is currently unclear which of the above
Neher: Noise suppression and working memory

factors, if any, were responsible for the incongruent outcomes, and so further research would be needed to resolve this issue. Nevertheless, our results caution against the assumption of a general link between individual reading span and noise suppression benefit.

It is also worth pointing out that, although all our participants had notable bilateral HA experience (Sect. 2.1), they were not accustomed to the sound of the specific HA settings tested here. Ng et al (2014) obtained data consistent with the idea that novice HA users with long reading span can quickly adapt to unfamiliar HA settings, whereas novice users with short reading span need up to 6 months to do so. Habicht et al (2016) obtained data consistent with the idea that ≥1 year of bilateral HA experience positively impacts speech comprehension in noise, irrespective of the precise HA gain characteristics. However, they did not test for a potential interaction with reading span. Given these findings, further research is needed to investigate how HA setting-specific acclimatisation influences the relation between individual reading span and noise suppression benefit.

4.3 Applicability of listening span and N-back measures

In the current study, we relied on the listening span and N-back measures of Doherty & Desjardins (2015) to assess auditory working memory. This was because these authors had found these two measures to be sensitive to the effects of aiding for all the participants they had tested (Sect. 1). For the listening span measure, we observed ceiling effects for a block size of 2, test-retest correlations of mostly moderate magnitude for a block size of 4, and moderate-to-strong correlations for a block size of 6. The N-back measure, in turn, was problematic for almost 20% of our participants, with some individuals being completely unable to carry out the task. Furthermore, the test-retest reliability of the N-back measurements was poor. In view of the results of Doherty & Desjardins (2015), these findings were mostly unexpected. In general, they
Indicate a need for more methodological work, so that the applicability of the listening span and N-back measures for research studies with older HA users can be improved.

In this context, it is worth noting that, on average, our participants were older (72 vs. ~63 years) and had a larger PTA4 (49 vs. ~25 dB HL) than those of Doherty & Desjardins (2015). These differences could be an explanation for the problems that we encountered. It is also worth noting that the N-back procedure that we used differed from a standard N-back procedure as used in cognitive psychology (e.g. Jaeggi et al, 2010). In the standard procedure, the task of the participant is to determine whether the current item matches the item N steps back in the series on some criterion variable and to give a ‘yes’ or ‘no’ response via a button press (instead of repeating aloud the item N steps back in the series, as in the current study). Furthermore, in the standard procedure targets and lures (i.e. items that recur but not at the designated interval) are controlled across conditions, which enables signal detection theory to be applied to the collected data. It is possible that a standard N-back procedure would increase the reliability of the collected data.

Finally, it is worth noting that we used two different measures to assess verbal working memory, one visual (reading span) and one auditory (listening span). The correlation between sentence-final word recall and reading span performance was $r = 0.63$ (Sect. 3.1). In other words, there was notable unshared variance between these datasets, implying that the two measures do not assess precisely the same construct.

5 Conclusion

The results from the current study indicate that the listening span measure used here is in principle suited for assessing auditory working memory in elderly HA users, particularly if a relatively large block size is used for the data collection. The N-back measure used here is
Neher: Noise suppression and working memory

generally more problematic. Future work should ideally investigate how the reliability of these measures can be improved. Furthermore, the results indicate that HA noise suppression can affect both recognition and recall of speech at positive SNRs if the two types of data are collected in parallel. Finally, our results provide confirmation for a positive association between reading span and speech-in-noise performance, but do not lend support to the idea that reading span modulates noise suppression benefit.

Endnotes

1 In the study of Ng et al (2015), no effect of in- or excluding final-sentence word repetition on final-sentence word recall was found.

Acknowledgments

This work was supported by the German Research Foundation [grant DFG Cluster of Excellence “Hearing4all” EXC 1077/1] and by Sivantos GmbH, Germany. Parts of it were presented at the 2017 Annual Meeting of the German Audiological Society, Aalen, Germany, Feb. 22-25. We thank Matthias Vormann (Hörzentrum GmbH) and Giso Grimm (HörTech gGmbH) for their help with the test setup.

Declaration of interest

This research was co-funded by Sivantos GmbH, Germany. Author RLF is an employee of that company. Otherwise, there are no conflicts of interest to declare.
Neher: Noise suppression and working memory

References

Akeroyd M.A. 2008. Are individual differences in speech reception related to individual
differences in cognitive ability? A survey of twenty experimental studies with normal and

National Standards Institute*, Retrieved:


fidelity, hearing loss, and working memory for digital noise suppression. *Ear Hear*, 36, 505-
516.


based on selective cepstro-temporal smoothing *IEEE International Conference on Acoustics,
Speech and Signal Processing (ICASSP)*, pp. 4897-4900.

Brons I., Houben R. & Dreschler W.A. 2013. Perceptual effects of noise reduction with respect
to personal preference, speech intelligibility, and listening effort. *Ear Hear*, 34, 29-41.

Triplets Test (Part I): Implementations for telephone, internet and mobile devices. *Z Audiol
(Audiol Acoust)*, 53, 139-145.
Neher: Noise suppression and working memory


Neher: Noise suppression and working memory


Neher: Noise suppression and working memory


Neher: Noise suppression and working memory


Table 1: Speech-weighted SNR improvement ($\Delta$AI-SNR) for the dirmic, scnr and bcf settings relative to the unproc setting. Data are shown for the listening span and $N$-back target materials.

<table>
<thead>
<tr>
<th>HA setting</th>
<th>$\Delta$AI-SNR re. unproc (dB)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Listening span</td>
<td>$N$-back</td>
</tr>
<tr>
<td>dirmic</td>
<td>3.4</td>
<td>3.4</td>
</tr>
<tr>
<td>scnr</td>
<td>3.3</td>
<td>2.2</td>
</tr>
<tr>
<td>bcf</td>
<td>3.2</td>
<td>2.9</td>
</tr>
</tbody>
</table>
Table 2: Significant effects from the ANOVAs performed on the sentence-final word recognition scores (rows 2-5) and sentence-final word recall scores (rows 6-8) from the listening span measurements. d.f. denotes degrees of freedom. $\eta_p^2$ denotes partial eta-squared.
Individual audiograms averaged across left and right ears (grey curves) and grand average audiogram (black curve) of the participants.

144x110mm (300 x 300 DPI)
Graphical illustration of the effects of unproc, dirmic, scnr and bcf on an example stimulus from the listening span measurements. The panels on the left-hand side show time waveforms of the target speech, S (black) and the cafeteria noise, N (grey). The panels on the right-hand side show the difference in level between unproc and either dirmic, scnr or bcf calculated for the signal mixtures. a.u. denotes arbitrary units.

177x127mm (300 x 300 DPI)
Means and 95% confidence intervals of (a) the sentence-final word recognition scores and (b) the sentence-final word recall scores from the listening span measurements for the different span sizes and HA settings.
Boxplots of the N-back data for the different tasks and HA settings.

132x102mm (300 x 300 DPI)
Neher: Noise suppression and working memory

**Figure captions**

Figure 1: Individual audiograms averaged across left and right ears (grey curves) and grand average audiogram (black curve) of the participants.

Figure 2: Graphical illustration of the effects of unproc, dirmic, scnr and bcf on an example stimulus from the listening span measurements. The panels on the left-hand side show time waveforms of the target speech, S (black) and the cafeteria noise, N (grey). The panels on the right-hand side show the difference in level between unproc and either dirmic, scnr or bcf calculated for the signal mixtures. a.u. denotes arbitrary units.

Figure 3: Means and 95% confidence intervals of (a) the sentence-final word recognition scores and (b) the sentence-final word recall scores from the listening span measurements for the different span sizes and HA settings.

Figure 4: Boxplots of the N-back data for the different tasks and HA settings.