Auditory acclimatization to bilateral hearing aids: Effects on sentence-in-noise processing times and speech-evoked potentials

Julia Habicht*1,2, Mareike Finke2,3 and Tobias Neher1,2

1 Medizinische Physik, Oldenburg University, Oldenburg, Germany
2 Cluster of Excellence “Hearing4all”
3 Department of Otolaryngology, Hannover Medical School, Hannover, Germany

*Corresponding author:
Julia Habicht
Department of Medical Physics and Acoustics
Carl-von-Ossietzky University
D-26111 Oldenburg
Germany

Phone: +49 (0)441 2172-310
Fax: +49 (0)441 2172-350

Email: julia.habicht@uni-oldenburg.de

Date of 2nd resubmission: 19.05.2017
Statement of Conflicts of Interest and Funding

The authors report no conflicts of interest. This research was funded by the German Research Foundation (DFG) Cluster of Excellence “Hearing4all” (grant number EXC 1077/1).
Abstract

Objectives: Using a longitudinal design, the current study sought to substantiate indications from two previous cross-sectional studies that hearing aid (HA) experience leads to improved speech processing abilities as quantified using eye-gaze measurements. Another aim was to explore potential concomitant changes in event-related potentials (ERPs) to speech stimuli.

Design: Groups of elderly novice (novHA) and experienced (expHA) HA users matched in terms of age and working memory capacity participated. The novHA users were acclimatized to bilateral HA fittings for up to 24 weeks. The expHA users continued to use their own HAs during the same period. The participants’ speech processing abilities were assessed after 0 weeks (novHA: N = 16; expHA: N = 14), 12 weeks (novHA: N = 16; expHA: N = 14), and 24 weeks (N = 10 each). To that end, an eye-tracking paradigm was used for estimating how quickly the participants could grasp the meaning of sentences presented against background noise together with two similar pictures that either correctly or incorrectly depicted the meaning conveyed by the sentences (the “processing time”). Additionally, ERPs were measured with an active oddball paradigm requiring the participants to categorize word stimuli as living (targets) or non-living (non-targets) entities. For all measurements, the stimuli were spectrally shaped according to individual real-ear insertion gains and presented via earphones.

Results: Concerning the processing times, no changes across time were found for the expHA group. After 0 weeks of HA use, the novHA group had significantly longer (poorer) processing times than the expHA group, consistent with previous findings. After 24 weeks, a significant mean improvement of ~30% was observed for the novHA users, leading to a performance comparable to that of the expHA group. Concerning the ERPs, no changes across time were found.
Conclusion: The results from this exploratory study are consistent with the view that auditory acclimatization to HAs positively impacts speech comprehension in noise. Further research is needed to substantiate them.

Key words

Hearing aids, acclimatization, speech comprehension, eye-tracking, event-related potentials
A. Introduction

Age-related hearing loss is a common chronic health condition that often remains untreated. Despite substantial advancements in hearing aid (HA) technology over the last decades, less than 25% of the hearing-impaired population over the age of 60 uses HAs (e.g. Feder et al., 2015; Lin et al., 2011; Popelka et al., 1998). In the last years, evidence has been accumulating that untreated hearing loss can have important consequences for a listener’s speech processing abilities. For example, Campbell and Sharma (2014) observed an association between cortical reorganization and poorer speech-in-noise performance in the early stages of hearing impairment. Based on a review of the literature, Peelle and Wingfield (2016) concluded that untreated hearing loss can affect the cognitive systems underlying speech comprehension. Together, these findings indicate that the auditory system adapts to the reduced auditory input that arises from untreated hearing loss, particularly at higher (or late) levels of processing.

For listeners who are fitted with HAs, there may also be a period of adaptation or acclimatization as the novice users learn to make use of the new auditory input that the HAs provide (e.g. Cox and Alexander, 1992; Gatehouse, 1992; Horwitz and Turner, 1997; Kuk et al., 2003; Munro and Lutman, 2003; Yund et al., 2006). To illustrate, Gatehouse (1992) tested four participants with bilateral hearing losses who were fitted with HAs in one ear. Following 12 weeks of HA use, he found a significant improvement in aided speech recognition for the fitted but not the unfitted ears of his participants. Similarly, Munro and Lutman (2003) tested 16 participants with bilateral hearing losses who were fitted with HAs in one ear. Following 12 weeks of HA use, they also observed improved speech recognition for the fitted but not the unfitted ears.

While the results of Gatehouse (1992) and Munro and Lutman (2003) are consistent with an auditory acclimatization effect, those of other studies are not (e.g. Bentler et al., 1993; Dawes and Munro, 2016; Dawes et al., 2014a; Humes and Wilson, 2003; Humes et al., 2002; Saunders and Cienkowski, 1997). To
illustrate, Humes et al. (2002), who performed one of the largest acclimatization studies to date, tested a group of 134 novice bilateral HA users in terms of speech recognition one, six, and 12 months post-fitting. In spite of the considerable experimental effort, these authors were not able to demonstrate any improvements in speech recognition with HA use. Recently, Dawes et al. (2014a) tested groups of novice HA users with bilateral hearing losses who were either unilaterally or bilaterally fitted ($N = 16$ each) after one and 12 weeks of HA use. Consistent with the findings of Humes et al. (2002), they did not observe any improvements in speech recognition over time either.

Apart from speech recognition measurements, event-related potentials (ERPs) have been used to investigate possible changes in hearing abilities due to HA use (e.g. Bertoli et al., 2011; Dawes et al., 2013; Gatehouse and Robinson, 1996; Munro et al., 2007). Due to their high temporal resolution, ERPs can provide detailed insights into the neural processing of auditory signals. Thus, in contrast to speech recognition measurements, they enable a differentiation between early (sensory) and late (cognitive) processes (e.g. Luck, 2014). In a case study, Gatehouse and Robinson (1996) recorded early ERPs in response to different tone stimuli in a long-term unilateral HA user with bilateral hearing loss. Focusing on the analysis of the so-called N1 and P2 ERP components, they observed larger amplitudes for some of their stimuli in the fitted compared to the unfitted ear of the participant. Gatehouse and Robinson (1996) interpreted these across-ear ERP differences in terms of an acclimatization effect due to HA use. To follow up on this finding, Bertoli et al. (2011) carried out similar ERP measurements with experienced unilateral and bilateral HA users ($N = 10$ each), all of whom had symmetric hearing losses. Based on the findings of Gatehouse and Robinson (1996), they expected to find larger amplitudes and shorter latencies for the N1 and P2 components for the participants’ fitted ears relative to the unfitted ears. However, they did not find any clear differences between the unilateral and bilateral groups in terms of across-ear ERP differences, which would have been supportive of an acclimatization effect. Recently, Dawes et al. (2014b) performed an even more comprehensive ERP study. They fitted novice ($N = 42$) and
experienced \((N = 17)\) HA users with bilateral hearing losses according to ‘National Acoustic Laboratories-Non-Linear 1’ prescription targets (e.g. Dillon, 2012) and then performed ERP measurements with tone stimuli before and after 12 weeks of HA use. Consistent with the results of Bertoli et al. (2011), they did not observe any changes in the N1 and P2 components as a result of HA use.

A possible explanation for the lack of conclusive evidence for improved hearing abilities due to HA experience could be that the applied outcome measures tapped into more sensory processes reflecting sound perception rather than late processes reflecting cognitive-linguistic abilities. Presumably, the latter types of processes are particularly prone to neural adaptation (see above). According to this view, experimental paradigms for the assessment of speech comprehension would be suited for revealing effects of HA acclimatization. In this context, two recent cross-sectional studies (Habicht et al., 2016; Wendt et al., 2015) used an eye-tracking paradigm to compare experienced and inexperienced HA users matched in terms of age, pure-tone average hearing loss (PTA), and working memory capacity in terms of their speech comprehension abilities. In this paradigm, the participant has to grasp the meaning of acoustic sentence stimuli that may differ in linguistic complexity and that are presented together with two similar pictures that either correctly (target) or incorrectly (distractor) depict the sentence meanings. The participant has to indicate the target picture by pressing a button as quickly as possible after the sentence presentation (the “response time”). Additionally, the time taken by the participant to start fixating the target picture automatically (or pre-consciously) during the sentence presentation (the “processing time”) is determined using an eye-tracker. In the studies of Wendt et al. (2015) and Habicht et al. (2016), despite comparable stimulus audibility and speech recognition performance, the processing times were longer (poorer) for the inexperienced HA users than for the experienced HA users, whereas the response times did not differ across the two listener groups. Furthermore, there were no interactions between HA use and linguistic complexity. These findings suggest that, irrespective of
linguistic complexity, acclimatization to HAs leads to faster sentence comprehension in noise as measured by processing times.

In the current study, we used a longitudinal design to substantiate these cross-sectional findings. We fitted an experimental group of inexperienced HA users with bilateral HAs and performed the eye-tracking measurements before and after several months of HA use. In addition, we had a control group of experienced HA users complete the same test protocol with their own HAs. We hypothesized that the processing times of the experienced HA users would remain stable over time, whereas those of the novice HA users would initially be longer and then shorten with HA use duration. In addition to the processing time measurements, we performed ERP measurements while the participants performed an auditory oddball task. Oddball paradigms are commonly used for investigating central-auditory and cognitive processing (e.g. Polich, 2007). Typically, participants are presented with a sequence of frequent non-target stimuli (e.g. low-frequency tones) interspersed with infrequent target stimuli (e.g. high-frequency tones). A negative (N2) and a positive (P3) deflection is evoked in response to the infrequent target stimuli over central (N2) and parietal scalp (P3) regions (Luck, 2014). Finke et al. (2016a) used ERPs to investigate auditory-cognitive processing in normal-hearing listeners and cochlear implant users. In contrast to previous studies conducted with simple (e.g. tonal) stimuli, participants were required to discriminate actively between infrequent target words (living entities) and frequent non-target words (non-living entities). Results showed prolonged late (N2, P3) processing in cochlear implant users compared to normal-hearing listeners. In line with the eye-tracking results of Wendt et al. (2015) and Habicht et al. (2016) and the lack of acclimatization effects in previous ERP studies conducted using tone stimuli (see above), this suggests that late speech processing (involving processes such as lexical information access, semantic categorization, and sentence recognition) is sensitive to changes due to hearing loss and the compensation with hearing devices. In the current study, we therefore used the paradigm of Finke et al. (2016a) to further explore acclimatization effects in HA users. We hypothesized
that inexperienced HA users would differ from experienced HA users in their late ERP components (N2, P3). More specifically, we expected increased amplitudes and decreased latencies with increasing HA use in our experimental group but no such changes in our control group, consistent with an adaptation to the new auditory input provided by the HAs.

B. Methods

Ethical approval for all experimental procedures was obtained from the ethical review board of the University of Oldenburg. Prior to any data collection, written informed consent was obtained from all participants. The participants were paid on an hourly basis for their participation.

1. Participants

A total of 33 participants were recruited from a large database of hearing-impaired listeners available at the Hörzentrum Oldenburg GmbH. Fifteen of the participants were experienced HA (expHA) users with at least one year of bilateral HA experience (mean: 8.3 yr; standard deviation: 5.6 yr), whereas the other 18 participants were novice HA (novHA) users with no previous HA experience. In order to minimize procedural training effects, an effort was made to recruit individuals who had previously participated in the studies of Wendt et al. (2015) or Habicht et al. (2016). For all but four participants (all novHA users), this was the case. Additional inclusion criteria were (1) age from 60 to 80 yr, (2) bilateral, sloping, sensorineural hearing loss in the range from 40 to 80 dB HL between 3 and 8 kHz, and (3) self-reported normal or corrected-to-normal vision. The novHA users were instructed to wear their HAs for at least 6 h/day for a total duration of 12 weeks, whereas the expHA users were asked to continue using their HAs as normal (but at least for 6 h/day) over the same time frame. For the novHA group, we used the data-logging functionality of the HAs to obtain this information. Due to the wide range of HAs (see Hearing aid fittings and stimulus amplification), this was not possible for the expHA group, which is why
we relied on self-reports. One expHA user could not complete the study due to illness. Furthermore, two
novHA users did not wear their HAs the required 6 h/day and were therefore excluded from the study.
Thus, there were 14 expHA users and 16 novHA users fulfilling all requirements. Near the end of the first
12-week acclimatization period, all 30 participants were asked to continue in the study for another 12
weeks. Twenty participants (i.e., 10 per group) volunteered to extend the study for another 12 weeks,
resulting in a total of 24 weeks of HA use for them.

Table 1 (left) summarizes the main characteristics of the two groups of participants who took part for 12
weeks. Figure 1 shows mean hearing threshold levels. Pure-tone average hearing loss (PTA) as calculated
across ears for the standard audiometric frequencies from 0.5 to 4 kHz (PTA0.5-4k) amounted to 43 and
38 dB HL for the expHA and novHA group, respectively. An independent t-test revealed no group
difference in terms of age ($t_{28} = -0.5, p > 0.5$). Another independent t-test showed that there was a
significant group difference in terms of PTA0.5-4k ($t_{28} = -2.8, p < 0.01$), with the novHA group having a
lower mean PTA0.5-4k. Table 1 (right) also summarizes the main characteristics of the two groups of
participants who took part for 24 weeks. Again, the two groups differed in terms of PTA0.5-4k ($t_{18} = -3.2,
p < 0.01$) with the novHA group having a lower mean PTA0.5-4k, but not in terms of age ($t_{18} = -0.8, p > 0.5$).

2. Test protocol

The novHA participants attended four or five visits. At the first visit, the HAs were fitted. The second visit
comprised the individual real-ear insertion gain (REIG), working memory capacity, and speech
recognition measurements, and at the third and fourth visit the response times, processing times, and
ERPs were measured. Those participants who agreed to take part in the 12-week extension of the study
(see Participants) repeated these measurements another time at the fifth visit. Between the third and
fourth visit and the fourth and fifth visit, the participants used their HAs for about 12 weeks (see Table 1
for overall HA use duration). The first and second visit took 1 h each, whereas the remaining visits took 2 h each. The expHA users followed the same test protocol, except that they did not attend the first visit.

3. Apparatus

All measurements took place in a soundproof booth. The (visual) working memory capacity measurements were performed with a computer connected to an additional monitor. The speech recognition measurements were performed with a computer equipped with a RME Fireface UCX (Audio AG, Haimhausen, Germany) soundcard.

For the electroacoustic measurements, we used a computer equipped with a Siemens Unity 2 (Sivantos, Erlangen, Germany) measurement box. For the response time and processing time measurements, the setup of Wendt et al. (2014) was used (EyeLink 1000 desktop system, EyeLink CL high-speed camera, SR Research Ltd., Samsung 2253BW monitor). The visual stimuli were presented on a 22” multiscreen color computer monitor with a resolution of 1680 × 1050 pixels. Participants were seated such that their eyes were 60 cm in front of the monitor. Calibration of this setup was carried out at the start of each block of measurements (using a 9-point fixation stimulus procedure of the manufacturer of the eye-tracker). Furthermore, before each stimulus presentation, participants had to fixate a point in the center of the screen for drift correction. The behavioral response (i.e. a button press) of the participant was collected using a keyboard (Logitech Deluxe 250, Logitech Europe S.A., Lausanne, Switzerland). All acoustic stimuli were presented via free-field equalized Sennheiser (Wennebostel, Germany) HDA200 headphones. The acoustic stimuli were calibrated with a Brüel & Kjær (B&K; Nærum, Denmark) 4153 artificial ear, a B&K 4134 ½” microphone, a B&K 2669 preamplifier, and a B&K 2610 measurement amplifier.

The ERPs were recorded in an electrically shielded booth with a Biosemi ActiveTwo system and Biosemi ActiView software 6.03 (BioSemi, Amsterdam, Netherlands). The recordings were digitized at 1024 Hz
without any additional filtering. Before the recordings, we made sure that all electrode impedances were lower than 10 kΩ. Stimuli were controlled using MATLAB R2006b (Mathworks, Natick, MA), presented using an RME (Audio AG, Haimhausen, Germany) DIGI 96/8 PAD soundcard, converted via an RME ADI-8 DS digital-analog converter (sampling rate: 48 kHz), and presented over Etymotic Research (Elk Grove Village, Illinois) ER-2 insert earphones.

To ensure adequate audibility, all speech stimuli were amplified according to individual REIGs (see Hearing aid fittings) using the Master Hearing Aid research platform (Grimm et al., 2006).

4. Hearing aid fittings and stimulus amplification

The novHA users were fitted bilaterally with Sivantos (Erlangen, Germany) pure micon 7mi receiver-in-the-canal (RIC) devices. These HAs are equipped with multi-channel dynamic range compression, noise management, and feedback cancellation. Acoustic coupling was achieved via standard double click domes or, if ear canals were too small, closed click domes. The HAs were fitted according to ‘National Acoustic Laboratories – Non-Linear 1’ prescription targets (NAL-NL1; e.g. Dillon, 2012). Target gains were verified via REIG measurements in accordance with recommended procedures (British Society of Audiology and the British Academy of British, 2007). Active noise (e.g. wind) management was activated and its strength adjusted if requested by the participants. Volume controls were switched off. The novHA users were allowed to get used to their devices for up to three days. Gains were adjusted only if the novHA users felt that they could not tolerate the prescribed amplification for the duration of the study (mean deviation from target = 2.7 dB; standard deviation = 2.5 dB; see Figure 2). Afterwards, no further adjustments were made to the HA fittings, and participants were not able to alter the amplification themselves.
The expHA participants continued to use their own devices in their own settings for the duration of the study. They were instructed not to make any alterations to them. The HAs were all non-linear devices fitted with click domes. Specifically, they were Phonak Audeo ($N = 4$), Widex Bravo ($N = 2$), Oticon Alta ($N = 1$), Oticon Nera2 ($N = 1$), Siemens Pure ($N = 1$), Resound Air ($N = 1$), Unitron Shine ($N = 1$), Audio Service Riva ($N = 1$), and Oticon Agil ($N = 2$) devices. REIG measurements were also performed for the expHA group to document the amplification characteristics. Figure 2 shows mean prescription target REIGs and mean user REIGs for the left and right ears of the novHA and expHA users as measured using the International Speech Test Signal (ISTS; Holube et al., 2010) at a 65-dB-SPL input level. As can be seen, for the novHA users the insertion gains were close to the targets up to 4 kHz. In contrast, the expHA users were generally underfitted across all frequencies.

For a subset of the 30 participants (i.e., 4 novHA users and 3 expHA users), REIG measurements were repeated after 12 weeks to verify the stability of the prescribed amplification over time. The mean difference in user REIGs was 0.2 dB (expHA: -0.1 dB; novHA: 0.4 dB). An independent $t$-test indicated no change in gain over time ($t_6 = 1.4, p > 0.05$).

Based on the individual REIG measurements, we spectrally shaped all speech stimuli and presented them via earphones during the response time, processing time, and ERP measurements. Our motivation for doing so was that we wanted to approximate the individual HA fittings as closely as possible, while at the same time achieving very good reproducibility of the acoustical conditions across the different visits. An approach commonly adopted in HA research is to perform such measurements in the sound field with the participant wearing the HAs. However, this approach is susceptible to various confounding influences, for example due to head movements, standing waves, variability in the position of the participant and the HAs across different measurements, as well as uncontrollable interactions with the HA signal processing (cf. Billings, 2013). To avoid these confounds, we presented all stimuli via...
headphones or insert earphones with individualized spectral shaping to ensure a close correspondence with the HA fittings the listeners were accustomed to.

5. **Working memory capacity**

To characterize our participants in terms of cognitive function we administered the reading span test after Daneman and Carpenter (1980) using the implementation of Carroll et al. (2015). The reading span test is a measure of visual working memory capacity that is rather widely used in hearing research (e.g. Lunner, 2003; Neher et al., 2014) and that includes three sub-tasks. First, participants have to read aloud sentence segments displayed successively on a screen and have to answer “yes” if the three previous segments made up a meaningful sentence (e.g. “Das Mädchen–sang–ein Lied”; “The girl–sang–a song”) or “no” if they did not make up a meaningful sentence (e.g. “Die Flasche–trank–Wasser”; “The bottle–drank–water”). After a block of three to six sentences, the participant has to repeat as many of the first or final words of the last block of sentences as possible in any order. Altogether, there were three training and 54 test sentences. As the performance measure, we used the percentage of correctly recalled target words for the 54 test sentences.

6. **Speech recognition measurements**

Prior to the eye-tracking measurements, we assessed baseline speech recognition performance using the “Oldenburg corpus of Linguistically and Audiologically Controlled Sentences” (OLACS; Uslar et al., 2013). The OLACS consists of seven grammatically correct sentence structures that vary in linguistic complexity. For our measurements, we used two sentence structures (see Table 2): (1) subject-verb-object sentences with a canonical word order and therefore ‘low’ linguistic complexity, and (2) object-verb-subject sentences with a non-canonical word order and therefore ‘high’ linguistic complexity. In each sentence,
there are two characters (e.g., a dragon and a panda), one of which (the subject) performs a given action
with the other (the object). In the German language, the linguistic complexity of these sentences is
determined by relatively subtle grammatical or acoustic cues, e.g. “Der müde Drache fesselt den großen
Panda” (meaning: “The tired dragon ties up the big panda”; low linguistic complexity) vs. “Den müden
Drachen fesselt der große Panda” (meaning: “The big panda ties up the tired dragon”; high linguistic
complexity).

The task of the participants was to repeat the words they had understood, which an experimenter then
scored. Initially, a training measurement based on 40 sentences with low and high linguistic complexity
was performed to familiarize the participants with the sentences and the procedure. Using 40 additional
sentences with low and high linguistic complexity, we then estimated the SNR corresponding to 80%-
correct speech reception (‘SRT80’). For all measurements, we used stationary speech-shaped noise at a
nominal level of 65 dB SPL and amplified the speech-in-noise stimuli according to the individual REIGs.

7. Response time and processing time measurements

Stimuli and response time measurements

As acoustic stimuli, we used OLACS sentences with low and high linguistic complexity in speech-shaped
noise. For each participant and level of linguistic complexity, we presented the stimuli at the individual
SRT80 with linear amplification according to the individual REIGs. In each case, we presented the noise at
a nominal level of 65 dB SPL. As visual stimuli, we used the picture sets developed by Wendt et al. (2014)
that complement the sentences of the OLACS. Each picture set consists of two similar pictures that are
displayed next to each other and that are presented together with the corresponding acoustic sentence.
One of the pictures (the target) correctly depicts the situation conveyed by a given sentence. The other
picture (the competitor) depicts the same situation but with interchanged roles of subject and object.
For each sentence, there are two picture sets. In one picture set, the left picture is the target; in the other picture set, the right picture is the target.

During the eye-tracking measurements the acoustic and visual stimuli were combined as follows (cf. Wendt et al., 2014). As shown in Figure 3, the visual stimulus is initially presented on its own (stimulus segment 1) and then the acoustic sentence-in-noise stimulus is added (stimulus segments 2-5). The participant has to identify the target picture by pressing a corresponding button on the hardware controller as quickly as possible after the presentation of the acoustic stimulus (stimulus segment 6). On each trial, the time taken to press the button relative to the end of the spoken sentence was recorded. In the following, we will refer to this as the response time.

**Processing time measurements**

To measure the processing times, we followed the procedure of Wendt et al. (2015). That is, we determined the eye-fixations toward three different regions of interest on the computer monitor: (1) the target picture, (2) the competitor picture and (3) other regions. Based on the recorded eye-fixation data, we first determined the so-called single target detection amplitude (sTDA) for each participant and weeks of HA use (0 weeks, 12 weeks, 24 weeks) across all sentences of a given level of linguistic complexity. The sTDA is a quantitative, normalized measure across time of the eye-fixation rate of a participant towards the target picture in relation to the eye-fixation rate towards the competitor picture and other regions on the screen (cf. Wendt et al., 2015). Figure 3 shows the sTDA of an example participant as a function of time for an example sentence with low (Figure 3a) and high (Figure 3b) linguistic complexity. If at a given point in time the target picture is fixated more than the competitor picture (or any other region on the screen), the sTDA is positive. If the competitor picture is fixated more, the sTDA is negative. Since the sentence recordings differ in terms of their durations, the eye-fixations were time-aligned by segmenting them in a consistent fashion (cf. Wendt et al., 2014). The processing
time was estimated on the basis of the point of target disambiguation and the decision moment. The point of target disambiguation corresponds to the onset of the first word that allows for disambiguation to occur, i.e. the moment from which the target picture can in principle be identified by the participant. For the sentences used here, the point of target disambiguation always corresponded to the start of stimulus segment 3 (see Figure 3). The decision moment is defined as a relative criterion threshold corresponding to the 42%-point of the sTDA maximum of each test condition (for details see Habicht et al., 2016; Müller et al., 2016). The processing times were then derived by taking the difference (in milliseconds) between the points of target disambiguation and decision moments. Using this approach, we estimated the processing time for each participant and level of linguistic complexity.

Procedure

The eye-tracking data were recorded in four blocks. Each block included 37 trials. Specifically, there were 30 trials with 15 sentences of low and high linguistic complexity and 7 catch trials. These catch trials were included to force the participants to look at both pictures each time and to prevent other task-solving strategies. We used two types of catch trials. First, either the target or competitor picture was depicted on both the left and the right side of the screen. Hence, either both or neither picture matched the spoken sentence. Second, two additional sentence types with different structures (and thus levels of linguistic complexity) from the OLACS were included: subject-relative clauses (e.g. “Der Lehrer, der die Models bestiehlt, zittert” – “The teacher who is stealing from the models is shivering”) and object-relative clauses (e.g. “Der Maler, den die Vampire beschatten, gähnt” – “The painter whom the vampires are shadowing is yawning”). These types of sentences were included to prevent the participants from getting used to specific sentences structures, thereby forcing them to continuously attend to them. The test blocks were presented in randomized order across participants. Each participant carried out 148 test trials (37 trials x 4 test blocks). One test block took about 7 min to complete (28 min in total).
Preparatory analyses

Initially, we assessed the ability of our participants to identify the correct picture by determining the picture recognition rate for each level of linguistic complexity and HA use duration. Although the picture recognition rate cannot be compared directly with the SRT80, one would expect picture recognition rates of around 80% and higher due to the availability of both acoustic and visual information. All participants \((N = 30)\) achieved picture recognition rates of 80% and above.

Since the response times were logarithmically distributed, we applied a logarithmic transformation to them. We then used Kolmogorov’s test and Levene’s test to check the assumptions of normality and homogeneity of variance. All datasets fulfilled the requirements for normality and homogeneity of variance (all \(p > 0.05\)). The same was true for the processing times. Thus, we used parametric statistics to analyze the log response times and processing times.

8. Event-related potential (ERP) measurements

Method

We used an active auditory oddball paradigm to evaluate the sensory and cognitive processing of speech stimuli by means of ERPs (Finke et al., 2016a). The task of the participants was to respond to the target stimuli \((p = 0.2)\) by pressing a button. The stimuli were two-syllabic German nouns with living beings as targets and non-living entities as non-targets (Rufener et al., 2014), all presented in quiet at a nominal presentation level of 65 dB SPL with linear amplification applied according to individual REIGs. Each stimulus word lasted for 800 ms. Individual stimuli were separated in time by 1.5 s with ±50 ms of jitter to avoid phase-locking. We presented 350 stimuli in three predefined lists, whereby 7 target stimuli and
28 non-target stimuli were repeated 10 times in random order. At least two non-target stimuli were presented between two target stimuli. The electroencephalogram (EEG) was continuously recorded from 68 scalp electrodes. The electrodes were placed according to the international 10-20 system (Jasper, 1958). In addition to the ERPs, hit rates and response times were measured.

**General analysis**

The EEG data were analyzed using EEGLAB 13.4.4b (Delorme and Makeig, 2004). Raw data were re-referenced to the average of the left and right mastoids (Luck, 2014; Woodman, 2010), downsampled to 250 Hz, and filtered from 1 to 40 Hz using sinc finite impulse response filters with a Hanning window (Widmann and Schröger, 2012). Dummy epochs were then created by extracting data segments from 0 to 2000 ms relative to the auditory stimulus onset. Epochs containing unique, non-stereotypical artifacts (e.g. swallowing, electrode cable movements) greater than three standard deviations over all epochs were rejected from further analyses. Epochs containing repeatedly occurring, stereotypical artefacts (e.g. eye-blinks, horizontal eye-movements, and electrical heartbeat) were corrected using independent component analysis (ICA; Jung et al., 2000). Computed ICA components were applied to re-referenced, filtered (0.1-30 Hz), and epoched (−200 to 1500 ms re. stimulus onset) data. For baseline correction, the pre-stimulus interval (~200 to 0 ms) was used. For all further analyses, only trials to which participants had responded correctly (i.e. hits for target stimuli and correct rejections for non-target stimuli) were included.

Next, single-subject averages were visually inspected for the different conditions (targets, non-targets) and HA use durations (0 weeks, 12 weeks, 24 weeks) to verify that the expected peaks could be reliably identified in each sub-average. Regions of interest (ROIs) were determined for fronto-central (N1 and P2), central (N2) and parietal (P3) waves. Each ROI included the channels with the largest deflection.
observed in the grand average response. For further ERP analysis, 11 channels (Fz, F1, F2, FCz, FC1, FC2, FC3, FC4, Cz, C1, C2) for the N1 and P2 component, 12 channels (Fz, FCz, FC1, FC2, FC3, FC4, Cz, C1, C2, C3, C4, CPz) for the N2 component (see Figure 5), and 11 channels (CPz, CP1, CP2, Pz, Pz1, Pz2, P3, P4, P0z, PO3, PO4) for the P3 component were included. For the statistical analyses, the means of the absolute signed (positive and negative) area under the ERP waveform for the given latency range was used for quantifying the amplitudes (Luck, 2014). The latency ranges were determined as 50 to 250 ms for the N1 component, 150 to 400 ms for the P2 component, 570 to 780 ms for the N2 component and 650 to 950 ms for the P3 component by verifying that all single-subject averages showed a corresponding wave within each latency range. For the latency analyses, the 50% area time point which divides the signed area into one-half was quantified. For the P3 response, the analyses were performed on the difference wave which was computed by subtracting the non-target waveforms from the target waveforms for each participant and HA use duration (0 weeks, 12 weeks, 24 weeks; Finke et al., 2016a).

Preparatory analyses

In preparation for the ERP analysis, we excluded one expHA user due to missing channels in the raw data and one novHA user whose amplitudes and latencies deviated by more than three times the interquartile range from the lower and upper quartiles of the novHA datasets. Furthermore, based on the visual inspection of the ERPs, we excluded three novHA users for whom reliable N1 and P2 responses could not be identified in any single-subject average. Thus, we had data from 25 participants (expHA: N = 13; novHA: N = 12) for the final analyses. According to Shapiro-Wilk’s test and Levene’s test, the amplitude and latency data fulfilled the requirements for normality and homogeneity of variance (all p > 0.05).
C. Results

1. Working memory capacity

On average, the expHA and novHA groups could recall 43.8% and 39.6% of all target words, respectively \((N = 30)\). These results are in good agreement with other comparable studies (e.g. Arehart et al., 2013; Neher et al., 2014). An independent \(t\)-test revealed no significant difference in terms of reading span between the two groups \((t_{28} = -1.2, p > 0.2)\). For the participants taking part 24 weeks \((N = 20)\), the expHA and novHA groups could recall 45.6% and 43.3% of all target words, respectively. Another independent \(t\)-test revealed no significant difference in terms of reading span between these two groups either \((t_{18} = -0.6, p > 0.5)\). The results of the reading span test are summarized in Table 1.

2. Speech recognition measurements

On average, the two groups of participants achieved very similar SRT80s for sentences with low \((-1.6 \text{ vs. } -1.5 \text{ dB})\) and high \((-1.1 \text{ vs. } -1.3 \text{ dB})\) linguistic complexity \((N = 30)\). Two independent \(t\)-tests showed no differences among the expHA and novHA groups for the two sentence types (both \(t_{28} < 0.2\), both \(p > 0.6\)). Furthermore, there was no difference in mean SRT80s for the two sentence types \((t_{29} < 3.3, p > 0.05)\). For the participants taking part 24 weeks \((N = 20)\), the SRT80s for sentences with low \((-1.8 \text{ vs. } -1.5 \text{ dB})\) and high \((-1.2 \text{ vs. } -1.3 \text{ dB})\) linguistic complexity did not differ across groups (both \(t_{18} < 1.1\), both \(p > 0.05\)). However, across sentence types there was a marginal difference in the SRT80s (i.e. \(-1.7 \text{ vs. } -1.3 \text{ dB})\) in favor of the sentences with low linguistic complexity \((t_{19} < 7.2, p = 0.02)\). The results of the SRT80 measurements are also summarized in Table 1.
3. **Response times**

To analyze the response time measurements, we performed two analyses of variance (ANOVAs) – one for the participants who took part for 12 weeks and another one for the participants who took part for 24 weeks – with participant group (expHA, novHA) as between-subject factor and linguistic complexity (low, high) and HA use duration (0 weeks, 12 weeks or 0 weeks, 12 weeks, 24 weeks) as within-subject factors on the log-transformed data. None of the main effects or interactions was significant (all $p > 0.05$). (We also performed a corresponding analysis on the response times after normalizing them based on a per-subject grand average response time, but this did not change the results.) Table 3 shows median response times for the different levels of linguistic complexity, listener groups, and HA use durations.

4. **Processing times**

To analyze the processing times, we performed a mixed-model ANOVA with listener group ($N = 30$) as between-subject factor and linguistic complexity (low, high) and HA use duration (0 weeks, 12 weeks) as within-subject factors. This revealed significant effects of listener group ($F_{(1,28)} = 21.0$, $p < 0.0001$, $\eta^2_p = 0.43$) and linguistic complexity ($F_{(1,28)} = 55.6$, $p < 0.00001$, $\eta^2_p = 0.67$), but no interactions (all $p > 0.05$).

Figure 4 shows mean processing times with standard deviations for the listener groups after 0 (left panel) and 12 (middle panel) weeks of HA use. After 0 weeks, the novHA group had longer processing times than the expHA group (mean listener group difference: 450 ms). After 12 weeks, the expHA users had processing times very similar to those obtained at the beginning of the study (0 weeks), whereas for
the novHA users there was a trend towards shorter (better) processing times (mean listener group difference: 339 ms).

Additionally, we performed a mixed-model ANOVA with participant group ($N = 20$) as between-subject factor and linguistic complexity (low, high) and HA use duration (0 weeks, 12 weeks, 24 weeks) as within-subject factors. This revealed significant effects of listener group ($F_{(1,18)} = 7.6, p = 0.013, \eta^2_p = 0.3$) and linguistic complexity ($F_{(1,18)} = 53.8, p < 0.00001, \eta^2_p = 0.75$). Furthermore, we observed a significant interaction between listener group and HA use duration ($F_{(2,36)} = 3.5, p = 0.042, \eta^2_p = 0.16$). No further interactions were observed (all $p > 0.05$). Post hoc analyses with Bonferroni correction showed that the processing times of the two groups differed in favor of the expHA group after 0 weeks ($F_{(1,18)} = 7.9, p = 0.012, \eta^2_p = 0.31$) and after 12 weeks ($F_{(1,18)} = 4.2, p = 0.046, \eta^2_p = 0.19$), but not after 24 weeks ($F_{(1,18)} = 1.4, p = 0.72, \eta^2_p = 0.01$). Additionally, the processing times of the novHA users after 24 weeks of HA use differed significantly from those after 0 weeks ($F_{(1,18)} = 6.8, p = 0.018, \eta^2_p = 0.27$) and 12 weeks ($F_{(1,18)} = 5.6, p = 0.029, \eta^2_p = 0.24$) of HA use. In contrast, no changes in the processing times of the expHA users were observed across the three HA use durations (all $p > 0.05$). Figure 4 (right panel) shows mean processing times with standard deviations for the two groups of participants after 24 weeks of HA use. As can be seen, after 24 weeks of HA use the trend observed after 12 weeks continued, with the novHA group performing on a par with the expHA group. Altogether, the processing times of the novHA group therefore improved from 1283 ms to 927 ms (i.e. by almost 30%) after 24 weeks of HA use.

Because two of our novHA users had not participated in the studies of Wendt et al. (2015) or Habicht et al. (2016) whereas the majority had done so (see Participants), we repeated the above analysis with the ‘untrained’ listeners excluded ($N = 18$). The results were essentially unchanged. That is, we still observed significant effects of listener group ($F_{(1,16)} = 5.3, p = 0.035, \eta^2_p = 0.2$) and linguistic complexity ($F_{(1,16)} = 41.6, p < 0.00001, \eta^2_p = 0.72$) and a significant interaction between HA use duration and...
participant group \( F_{(2,32)} = 1.7, p = 0.049, \eta^2_p = 0.1 \), consistent with the hypothesis that HA experience leads to shorter processing times.

5. ERP measurements

The analyses of the ERPs showed no effects of HA use duration (all \( p > 0.05 \)). For conciseness, we will therefore analyze and present the ERP data averaged across HA use duration (i.e. 0 weeks and 12 weeks to maximize \( N \)) in the following sections.

Hit rates and response times

On average, the expHA and novHA groups had hit rates of 97.7% and 96.3% and response times of 1124 ms and 1194 ms, respectively. Two independent \( t \)-tests revealed no group differences in terms of these two measures (both \( t_{23} < -0.8 \), both \( p > 0.05 \)).

ERP amplitudes and latencies

To analyze the N2 amplitudes and latencies, we performed ANOVAs with participant group (expHA, novHA) as between-subject factor and stimulus type (target, non-target) as within-subject factors. The expHA users showed a larger N2 wave than the novHA users, as indicated by a significant main effect of listener group \( F_{(1,23)} = 4.29, p = 0.05, \eta^2_p = 0.16 \). The N2 amplitudes were significantly smaller for the novHA users than for the expHA users (means: 96 vs. 132 \( \mu V*ms \)). No further main effects or interactions for N2 amplitudes and latencies were observed (all \( p > 0.05 \)). Figure 5 shows the ERPs for the expHA users and novHA users and the two stimulus types averaged across the N2 ROI.
To analyze the P3 amplitudes and latencies of the (target minus non-target) difference waves (data not shown), we performed two independent t-tests. These revealed no significant effects (both \( p > 0.05 \)).

**D. Discussion**

The main purpose of the current study was to substantiate indications from two earlier cross-sectional experiments that HA experience offers an advantage in terms of grasping the meaning of sentences presented in background noise. To that end, we measured processing times (i.e., the time to grasp the meaning of sentences presented against noise together with two pictures that either correctly or incorrectly depict the meaning of the sentences) based on eye-gaze measurements as well as behavioral response times (i.e., the time taken to press a button after the end of the spoken sentence) using the paradigm of Wendt et al. (2014). Additionally, we measured speech-evoked ERPs (as well as response times) with an active oddball paradigm to investigate whether concomitant changes would be apparent in late ERPs (i.e. the N2 and P3 responses) to word stimuli. Participants were groups of experienced and inexperienced HA users. Consistent with our hypotheses, the processing times of the expHA group remained stable over time, whereas those of the novHA group improved. Despite the reduction in processing times, there was no corresponding reduction in response times. Concerning the ERPs, we observed smaller N2 amplitudes for the novHA users than for the expHA users. However, we observed no group differences in the N2 and P3 latencies, and neither did we find any ERP changes across time. Below, we discuss these findings in more detail.

**Does acclimatization to HAs lead to shorter processing times?**

Concerning the processing times, the effect of participant group that we observed is consistent with the earlier results of Wendt et al. (2015) and Habicht et al. (2016). In contrast to our earlier work, the results
reported here were obtained using a longitudinal design. As such, they provide a direct indication that HA use can have a positive influence on speech comprehension abilities (and that the better processing times of the expHA users observed in the two earlier studies were in fact due to changes in auditory processing as a result of HA provision). Interestingly, the positive effect of auditory acclimatization first emerged after 24 weeks of HA use. This suggests that the effects of acclimatization on cognitive-linguistic processes may take several months to manifest themselves. It is also worth noting that, in spite of having a smaller mean PTA_{0.5-4k} (38 vs. 45 dB HL; see Table 1), the novHA users initially had longer processing times than the expHA users. This could imply that consistent HA use can at least partly outweigh the added negative effects of a greater hearing loss.

An alternative explanation for the apparent acclimatization (or reversed auditory deprivation) effect could be that the novHA users initially had longer processing times because of their unfamiliarity with the provided amplification. As the novHA users then became accustomed to the altered auditory input, their processing times reduced to levels comparable with those of the expHA users. In an effort to shed light on these two possible explanations, Habicht et al. (2016) compared the effects of three different HA conditions on the processing times of experienced and inexperienced HA users. These conditions included the clinically established (and thus relatively common) prescription formula ‘National Acoustic Laboratories-Revised’ (NAL-R; e.g. Dillon, 2012) as well as two other conditions. The latter two conditions were chosen to result in different spectral envelope characteristics, e.g. one with substantial high-frequency amplification that is very atypical of clinically fitted HAs. In spite of this, Habicht et al. (2016) observed a group difference in favor of the experienced users for all three HA conditions. This would seem to suggest that the effects of HA use are persistent to unfamiliar gain characteristics, that is, that the results of Habicht et al. (2016) reflect a relative benefit of experience with amplified listening, regardless of the type of amplification. According to this view, the improvement in the processing times
of the novHA users observed in the current study would be due to reversed auditory deprivation. However, further measurements are necessary to confirm this assumption.

Regarding the practical implications of the observed improvement in sentence-in-noise processing times, one could argue that since there was no concomitant improvement in behavior (i.e. response times) there should not be any real-world consequences of this. However, if it takes a person longer to process a sentence (as in the current study), then in a complex conversation with multiple talkers the practical consequence of this would be to fall behind in the conservation, potentially leading to full withdrawal from it. In a complex conversation, there typically will not be time for any ‘offline’ analyses. Thus, a person has to be able to process the speech input ‘online’, that is, the ability the eye-tracking paradigm used here taps into.

It should also be noted that since we measured the response times after the sentence presentation (see Stimuli and response time measurements), it is currently unclear if the greater experimental effort required for performing the processing time measurements is in fact needed to reveal effects of, for example, participant group and HA use duration. Further research would be required to investigate if allowing the participants to behaviorally indicate the target picture during the sentence presentation would be just as sensitive to such effects.

Does acclimatization to HAs lead to greater ERP amplitudes and shorter latencies?

A further aim of the current study was to investigate whether the changes observed in the processing times would essentially also manifest themselves in speech-evoked ERPs, that is, in shorter latencies and greater amplitudes of late ERP components (N2, P3). Contrary to our expectations, we did not find longer N2 and P3 latencies in our novice users compared to our experienced users, irrespective of HA use duration. In general, the latencies that we measured were rather long, which was likely a result of the
speech stimuli that we used (e.g. Finke et al., 2016a; Luck, 2014). Due to the usage of word stimuli whose meaning had to be extracted, our participants were forced to engage relatively late cognitive processes, which are assumed to be reflected in the N2 and N4 waves (e.g. Finke et al., 2016a; Wang and Dong, 2013). Concerning the ERP amplitudes, we observed significantly smaller N2 amplitudes for the novHA users than for the expHA users, irrespective of HA use duration. However, we did not find any significant differences or changes in the P3 amplitudes (although there was a trend towards smaller P3 amplitudes in the novHA users). Hence, our ERP results provided only limited evidence for an effect of HA experience on higher-level processes related to speech processing.

One reason for the small between-group differences and the lack of any acclimatization effects in our ERP measurements could be that we presented our stimuli in quiet. Finke et al. (2016a) found the largest ERP differences between their two groups of participants (normal-hearing listeners and cochlear implantees) for stimuli presented in noise. It would therefore be of interest to repeat our measurements under noisy conditions.

Another possible reason could be that the expHA users were underfitted, with an average target-to-user REIG difference of 6.0 dB (see Hearing aid fittings and stimulus amplification). This was in contrast to the novHA users for whom audibility was better restored, with an average target-to-user REIG difference of 2.7 dB. Greater stimulus intensity leads to larger ERP amplitudes and shorter ERP latencies (e.g. Billings et al., 2007). Thus, it could be that the resultant differences in stimulus audibility had a confounding effect on the ERPs of our two groups.

### E. Summary & Conclusions

In the current study, we used a longitudinal design to explore the effects of auditory acclimatization to bilateral HAs on sentence processing times using the eye-tracking paradigm of Wendt et al. (2014). In addition, we complemented these measurements with speech-evoked ERPs as measured with an active...
oddball paradigm. For all measurements, the stimuli were spectrally shaped according to individual
REIGs and presented via earphones. We acclimatized elderly inexperienced HA users to bilateral HA
fittings for up to six months. Before acclimatization, these participants had significantly longer processing
times than participants matched in terms of age and working memory capacity with at least one year of
bilateral HA experience. This is consistent with the findings from two earlier cross-sectional studies and
suggests poorer sentence-in-noise processing. After 24 weeks of HA use, the processing times of the
novice users improved by almost 30% and were thus comparable to those of the experienced users. In
contrast, we found no acclimatization effects on ERPs related to higher-level speech processing.

Altogether, these results provide some support for the idea that HA use has a positive influence on
important communication abilities. Nevertheless, given the relatively small sample size used here, it
would be necessary to confirm these findings with a larger group of participants. Additionally, it is still
somewhat unclear if the observed improvement was due to reversed auditory deprivation or if
unfamiliarity with the provided amplification had a role to play. Furthermore, it would be of interest to
investigate more precisely which changes in neuronal processing HA (non-)use gives rise to. Therefore, a
next step could be to measure anatomical differences in cortical activity during sentence comprehension
in novHA and expHA users. This research could perhaps also shed some light on individual differences in
acclimatization to HAs.

F. Acknowledgements

The authors thank Sivantos GmbH, Erlangen, Germany for the provision of the hearing aids, Birger
Kollmeier (University of Oldenburg) for general support and advice, Martin Meyer (University of Zurich)
for providing the speech stimuli for the EEG measurements, Carlos Filipe Da Silva Souto (University of
Oldenburg) for helping with the EEG setup, Timo Gerdes (University of Oldenburg) for performing parts
of the measurements, and Stefan Debener (University of Oldenburg) for providing input to the ERP
analysis. This research was funded by the DFG Cluster of Excellence EXC 1077/1 “Hearing4all”.

G. References


British Society of Audiology and the British Academy of Audiology (2007). Real ear measurement to
verify the fitting of digital signal processing hearing aids of Podcast. www.thebsa.org.uk/wp-
content/uploads/2014/04/REM.pdf.

*PloS one* 9(2):e90594.

German reading span test with dual task design for application in cognitive hearing research.


**Figure captions**

Figure 1. Mean hearing threshold levels averaged across both ears of the expHA \((N = 14)\) and novHA \((N = 16)\) participants.

Figure 2. Mean user REIGs (black) and mean target REIGs (gray) based on NAL-NL1 with standard deviations for the expHA (top, \(N = 14\)) and novHA (bottom, \(N = 16\)) groups.

Figure 3. Illustration of the processing time and response time measurements. Shown is the hypothetical eye-fixation rate (sTDA, black line) for an example sentence (“Der müde Drache fesselt den großen Panda” - “The tired dragon ties up the big panda”) with the two corresponding pictures (top: target; bottom: distractor) over the course of the acoustic and visual stimulus presentation (segments 1-6). The shaded area illustrates 95% confidence intervals. The gray dot denotes the point of target disambiguation (PTD), which defines the onset of the first word that allows matching the acoustic sentence to the target picture (upper picture). The gray + symbol denotes the decision moment (DM), where the eye-fixation rate exceeds the criterion threshold [42%-point of the sTDA maximum, dashed line]. The horizontal gray bar corresponds to the processing time, i.e. the time difference between PTD and DM. The black x symbol denotes the time point of the button press and the black bar corresponds to the response time.

Figure 4. Mean processing times with standard deviations for the two participant groups (expHA, novHA) after 0 weeks (\(N = 30\)), 12 weeks (\(N = 30\)), and 24 weeks (\(N = 20\)) of HA use.

Figure 5. ERPs for non-targets (left) and targets (right) for the two groups of participants (expHA: \(N = 13\), black lines; novHA: \(N = 12\), gray lines), averaged across all ROI channels of the N2 analysis and HA use duration (0 weeks, 12 weeks). Note that 0 ms corresponds to the onset and 800 ms to the offset of the
auditory stimulus (vertical lines). The gray rectangle highlights the time range (570-780 ms) used for the analysis of the amplitudes and latencies. The voltage maps at N2 latency show grand averages for each participant group across stimuli and HA use durations.