Perceptual evaluation of noise reduction in hearing aids

Brons, I.

Citation for published version (APA):
Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort

Inge Brons, Rolph Houben, Wouter A. Dreschler

Adapted from *Ear and Hearing, 2013, 34(1), pp.29-41.*
3.1 Introduction

One of the main reasons for hearing aid dissatisfaction is the difficulty of listening to speech in noisy environments. Consequently, most currently marketed hearing aids use a single-microphone noise-reduction system to make listening in noisy environments more comfortable. The noise-reduction algorithm continuously analyzes the input signal to estimate the ratio between speech and noise, and reduces the gain for a frequency band if the band is dominated by noise (Chung 2004). Unfortunately, details regarding the properties of noise reduction (e.g., gain-reduction strength, signal-to-noise ratio (SNR) dependency, time constants) in hearing aids are rarely provided by manufacturers. Furthermore, it is unknown whether the perceptual effects of noise reduction (e.g., intelligibility, listening effort, and preference) differ among hearing aids or even among listeners. Consequently, clinicians have no guidelines for selecting the best noise-reduction system and settings. However, if more information were available regarding noise reduction and its effect on the perception of the user, clinicians could actively select the best individual noise-reduction system and settings, thereby increasing hearing aid satisfaction.

Numerous studies have examined the effects of noise reduction on speech intelligibility in noisy environments. These studies reveal that single-microphone noise-reduction system does not improve speech intelligibility (Nordrum et al. 2006; Loizou and Kim 2011).

Despite its inability to increase speech intelligibility, noise reduction has reportedly benefited several listeners using hearing aids (Bentler 2005). Therefore, researchers are increasingly seeking explanations for this perceived benefit. Apparently, benefit from noise reduction can be expected in terms of listening comfort, effort, and personal preference. However, conclusions from studies that have evaluated noise reduction on these outcomes differ. For instance, some studies found that listeners preferred to have noise reduction on compared with noise reduction off (Boymans and Dreschler 2000; Ricketts and Hornsby 2005). Although this preference for noise reduction suggests that noise reduction increases listening comfort and sound quality, other studies based on rating scales for listening comfort and sound quality found no difference between noise reduction on and noise reduction off (Alcàntara et al. 2003; Bentler et al. 2008). These two studies also included a rating scale for listening effort. Alcàntara et al. (2003) found equal listening-effort ratings when noise reduction was on and off, whereas Bentler et al. (2008) found a reduction in listening effort when noise reduction was on.

We identified two factors that might have contributed to the apparently conflicting results regarding the perceptual effects of noise reduction: (1) differences between the
Chapter 3

noise-reduction systems used and (2) individually different weightings of factors underly-
ing the overall preference. We designed a perceptual experiment to further inves-
tigate these factors.

3.1.1 Differences between noise-reduction systems

One possible explanation for the diverging results regarding the perceptual effects of
noise reduction in hearing aids may be that different hearing aids, and thus different
noise-reduction systems, were used in the studies. Each study compared noise reduc-
tion on with noise reduction off within one type of hearing aid. However, technical
measurements have shown that noise-reduction systems from different hearing aids
can differ substantially in the amount of gain reduction used (Bentler and Chiou 2006;
Hoetink et al. 2009). Currently, there are no perceptual data from comparisons of noise
reduction from different hearing aids, but we hypothesize that the differences in gain
reduction will also have perceptual consequences.

The first goal of this study was to test whether noise-reduction systems on the market
differ perceptually. Because noise reduction aims to reduce background noise while
retaining speech quality and intelligibility, we determined whether subjects attributed
differences among noise-reduction systems in terms of noise annoyance and speech
naturalness. This led to the first research question:

Q1. Are there differences regarding noise annoyance or speech naturalness (a) be-
tween noise reduction on and noise reduction off within the same hearing aid
and (b) among noise-reduction systems from different hearing aids?

Because we did not know whether differences between noise-reduction systems are
perceptually relevant, we used normal-hearing subjects in this initial study. Normal-
hearing subjects form a more homogeneous group of listeners than hearing-impaired
listeners do. For instance, because of suprathreshold deficits such as reduced fre-
quency resolution and impaired modulation detection, hearing-impaired listeners
could differ from one another in which noise-reduction effects are perceptible, com-
plicating the interpretation of the results. In addition, the use of normal-hearing sub-
jects allowed us to compare noise-reduction systems without compensation for an in-
dividual hearing loss. Thus, we did not have to use frequency-dependent linear gain
or dynamic range compression. Of course, if the present study shows that there are
perceptually relevant differences among noise-reduction systems for normal-hearing
subjects, the next step should focus on effects for hearing-impaired subjects, taking all
these complicating factors into account.
Aside from noise annoyance and speech naturalness, we also determined the overall preference, intelligibility scores, and perceived listening effort of the subjects. Thus, our second research question was as follows:

**Q2.** Does noise reduction influence the preference, intelligibility, or perceived listening effort of normal-hearing subjects compared with (a) no noise reduction and (b) noise reduction from other hearing aids?

### 3.1.2 Factors underlying individual preferences

A second possible reason for the diverging results regarding the perceptual effects of noise reduction in hearing aids is that most studies concentrate on group results. However, recent work in our laboratory provided evidence that even normal-hearing listeners differ significantly in their preference for noise-reduction strength (Houben et al. 2012). This study used paired comparisons to determine the preferred setting for noise-reduction strength in an algorithm that was designed for hearing aids. Five of the 10 normal-hearing subjects had an optimized noise-reduction strength that differed significantly from that of the averaged group data. Although the study provided no decisive answer concerning the factors underlying the differences in preference, the results suggest that an individual approach is required in the investigation of the perceptual effects of noise reduction.

The second goal of our study was to determine which factors might influence the preference of the subject for a specific type of noise reduction or for no noise reduction. We looked for correlations between the overall preference and noise annoyance, speech naturalness, intelligibility, and listening effort. Aside from group results, we also determined whether individuals differed in their preference, and we examined the factors related to this preference. Our third research question therefore was:

**Q3.** Is the overall preference of normal-hearing subjects related to the intelligibility scores, perceived listening effort, noise annoyance, or speech naturalness obtained for the same noise-reduction conditions?

It was also useful to include normal-hearing subjects for this objective. If we were to find substantial differences among subjects even in this homogeneous group of listeners, these differences would be caused by individual differences because there would be no differences in hearing ability to confound the results.
3.2 Methods

3.2.1 Method for the comparison of noise reduction from different hearing aids

In Chapter 2 we developed and evaluated a method that allows for direct comparison of noise-reduction systems of different hearing aids, without the confounding effects of other hearing aid characteristics. Briefly, we made recordings from linearly fitted hearing aids with all the processing features deactivated. On the basis of the difference between the input and output, we designed an equalization filter for each individual hearing aid. This filter was intended to remove perceptual differences among recordings from different hearing aids with signal processing turned off.

Once such an inverse filter is available for a specific linearly fitted hearing aid, it can also be applied to recordings from the same hearing aid with noise reduction turned on. The only difference among hearing aids is then caused by the noise reduction because all hearing aids are perceptually equal when noise reduction is turned off.

In Chapter 2 several tests were described to verify whether our methods for hearing aid fitting, recording, and filtering indeed removed all the perceptual differences among recordings from different hearing aids with all the processing features deactivated. First, we verified the linearity of the hearing aid gain by electroacoustical measurements. All the hearing aids had a linear response for input levels between 50 and 95 dB SPL for frequencies up to 6 kHz. Second, we calculated the objective hearing aid speech quality index (HASQI; Kates and Arehart 2010) for their hearing aid recordings of speech-in-noise signals. The equalization filter improved the HASQI score compared with band-pass filtering (the mean linear HASQI index was 0.863 for the band-pass-filtered signals and 0.945 for the equalized signals) and reduced the differences in the HASQI score among hearing aids (the maximum HASQI index difference between two hearing aids was 0.02 after band-pass filtering only, and 0.006 after equalization). These HASQI results show that the equalization filter minimized differences among hearing aids. Third, we performed listening experiments to determine whether the recordings from different hearing aids were perceptually similar after filtering. After filtering, all six normal-hearing subjects were unable to detect perceptual differences among recordings from different hearing aids. Thus, both the objective and subjective evaluations showed that the equalization filter removed perceptual relevant differences among the recordings from different hearing aids if the noise-reduction feature was turned off, without affecting the sound quality.
3.2.2 Hearing aid fitting and recordings

The hearing aids tested in this study were of four different brands of frequently used behind-the-ear hearing aids (Phonak Exélia M, ReSound Azure AZ80-DVI, Starkey Destiny 1200, and Widex Mind 440). This selection was a representative sample of the commercial hearing aids available at the time of the study. The hearing aid numbers used in this study were randomly assigned to the test hearing aids and are different from the numbers used in Chapter 2.

We applied the same methods of hearing aid fitting and recording as described in Chapter 2. In fact, we took the hearing aids, hearing aid settings, and equalization filters from the test set in Chapter 2, so that we were sure that our verification of the method was also applicable to the new recordings. We turned off all the signal-processing features in the hearing aids (directionality, feedback control, noise reduction, compression, frequency transposition, etc.) and carefully adjusted their gains to obtain the same insertion gain for all hearing aids. The target insertion gain was based on the NAL-RP prescription (Dillon 2001) for a conductive hearing loss of 30 dB at 500 Hz and 15 dB at 2 kHz. This resulted in an insertion gain of approximately 10 dB in the low and mid frequencies (between 125 Hz and 2 kHz), decreasing to 0 dB for the higher frequencies.

We recorded the hearing aid output with the use of a B&K Head and Torso Simulator (HATS Type 4128C), which was fitted with a custom-made tight-fitting earmold without venting. The recordings were made in a sound-treated double-walled booth (2.20 × 2.53 × 2.00 m). The speaker was placed 62 cm in front of the recording microphone (on axis) to minimize the influence of room reflections. All the free-field hearing aid input signals were corrected for the speaker response.

We designed an inverse filter for each hearing aid to remove any differences in frequency response that remained among hearing aids despite the careful adjustment of the hearing aid gain. To obtain the required filter response, we compared the hearing aid output to the output of a reference microphone. We used the Matlab function “fir2” to calculate the filter coefficients (500 taps) based on the required response. In addition, the frequency response was limited to 100 Hz to 5.8 kHz with elliptical filters of the seventh-order.

As described earlier, when a filter has been designed for a specific linearly fitted hearing aid, it can also be applied to recordings from the same hearing aid with noise reduction turned on to examine the isolated effect of noise reduction. We selected in each hearing aid the strongest available noise-reduction setting. For the purpose of exploring possible perceptual effects of noise reduction, investigators often use the
maximum setting (Ricketts and Hornsby 2005; Mueller et al. 2006; Nordrum et al. 2006; Palmer et al. 2006). In some hearing aids, the maximum noise-reduction setting was not the setting recommended by the manufacturer for an initial fit. However, if there are no perceptual differences between setting the noise reduction on and off or among different noise-reduction systems even if maximum noise reduction is applied, it is reasonable to hypothesize that there will also be no perceptual differences for lower settings.

3.2.3 Stimuli

We made hearing aid recordings of Dutch female speech (Versfeld et al. 2000) in a multitalker babble noise (Luts et al. 2010). The speech material consists of unrelated, low-context sentences, containing five to nine words per sentence. The signals were presented to each hearing aid with a noise level of 70 dB(A) and four different speech levels (63, 66, 70, and 74 dB(A)) to form stimuli at SNRs of -7, -4, 0, and +4 dB. These levels were well within the linear range of the hearing aids (50-95 dB SPL). The negative SNRs were chosen to prevent ceiling effects for intelligibility. We rated the listening effort and paired comparisons at -4 dB SNR as well so that we had one common SNR across all outcomes. Because hearing-impaired listeners have more difficulty with low SNRs, they are less likely to listen in a setting with -4 dB SNR during their daily activities. Therefore, we additionally measured listening effort at 0 and +4 dB SNR, and noise annoyance, speech naturalness and overall preference at +4 dB SNR. These SNRs are more relevant for hearing-impaired listeners and thus more relevant in evaluating hearing aid processing. The noise was continuous while the speech was paused for approximately 1 sec between sentences. One list (36 sec) preceded the stimulus lists in each condition to allow the hearing aid to adapt to the input signals.

We applied the inverse filters to all the hearing aid recordings. This step resulted in five different conditions. One condition represented the situation in which the noise reduction was turned off (hereafter the “unprocessed” condition). The four additional conditions represented the situations in which the noise reduction (NR) was turned on for each hearing aid (randomly coded NR1-NR4).

Stimuli for all the measurements consisted of single sentences with 0.5 sec of noise before and after the sentence. The stimuli were presented diotically with Sennheiser HDA200 headphones, which had been calibrated with a B&K Artificial Ear Type 4153. The noise level was 70 dB(A) for all the stimuli in the unprocessed condition.
3.2.4 Noise-reduction processing

Although we had no details on the implementation of the different noise-reduction systems, acoustical analyses gave insight into how they reacted to our input signals. We plotted the time signals and difference spectrograms to study the dynamic characteristics of the noise-reduction conditions. Figure 3.1 shows the results of this study. For each noise-reduction system, the upper plot shows the time signal of the hearing aid output with the noise reduction off (dark gray) compared with the output with the noise reduction on (light gray). In addition, the spectrogram-like color plots show the difference between noise reduction on and off (i.e., the gain reduction caused by the noise reduction) as a function of time and frequency. Note that each plot starts and ends with 0.5 sec noise only, with the sentence in between. The more negative the gain value, the stronger the noise reduction. Thus, red areas correspond to the maximum noise reduction (>10 dB reduction of the gain), whereas blue areas correspond to an inactive noise reduction. Figure 3.2 shows the long-term average gain reduction (averaged over 13 sentences) due to the four noise-reduction conditions as a function of frequency. Table 3.1 summarizes these plots by giving the median and the 5th and 95th percentile of the gain values (the difference between noise reduction on and off, in dB) across all time-frequency bins. The 5th percentile provides an estimate of the maximum gain reduction applied by the noise reduction. Again, a more negative value indicates stronger suppression. Similarly, the 95th percentile is an estimate of the minimum gain reduction.

3.2.5 Subjects

The number of subjects chosen for this study was based on a power calculation for speech intelligibility and listening effort. We used a within-subject standard deviation of approximately 13.9% taken from Bosman (1989). The slope of the psychometric function at the speech-reception threshold (SRT₅₀; the SNR at which 50% of the sentences are correctly repeated by the subject) was 16% per dB. To be able to detect a difference of 16% (thus 1 dB change in SRT₅₀) nine subjects should be included for α = 0.05 and 1-β = 0.8. For a power calculation for listening-effort rating, we used the within-subject standard deviation found by Luts et al. (2010), which averaged 0.88 points on the 7-point listening effort scale. With α = 0.05 and 1-β = 0.8, nine subjects are sufficient to detect a 1-point difference on the listening-effort rating scale. For the outcomes measured with paired comparisons, we had no appropriate data available for a power calculation. However, for these data we were especially interested in individual differences, so that the number of subjects was of less importance for these outcome measures. On the basis of the power calculations, we decided to include 10 normal-hearing subjects.
Ten normal-hearing subjects between 19 and 23 years of age (average = 20.8 years) participated in this study. Their hearing thresholds were 15 dB HL or better at 0.25, 0.5, 1, 2, 3, 4, 6, and 8 kHz.

Figure 3.1: Acoustical effects of the four noise-reduction systems on speech in babble noise at -4 dB SNR (left column) and +4 dB SNR (right column). The time signal of the hearing-aid output with noise reduction off (dark background signal) and on (light foreground signal) for each processing condition is shown. The changes in gain caused by noise reduction (the difference between noise reduction on and off) as a function of time and frequency are also shown.
Figure 3.2: Gain-reduction spectra for the four noise-reduction conditions. The difference between noise reduction on and off as a function of frequency is shown, averaged over 13 sentences in noise at two different input SNRs.

Table 3.1: Median and 5th and 95th percentile values of the gain difference between noise reduction and unprocessed conditions for the time-frequency bins that are presented in Figure 3.1.

<table>
<thead>
<tr>
<th></th>
<th>-4 dB Signal-to-noise ratio</th>
<th>+4 dB Signal-to-noise ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median</td>
<td>5th percentile</td>
</tr>
<tr>
<td>NR1</td>
<td>-4.1</td>
<td>-11.5*</td>
</tr>
<tr>
<td>NR2</td>
<td>-7.0</td>
<td>-12.7*</td>
</tr>
<tr>
<td>NR3</td>
<td>+0.8</td>
<td>-9.3*</td>
</tr>
<tr>
<td>NR4</td>
<td>-0.4</td>
<td>-4.5</td>
</tr>
</tbody>
</table>

* Calculated over time-frequency bins between 100 and 750 Hz. Taking the whole frequency range into account resulted in an underestimation of the maximum gain reduction, because maximum gain reduction was acquired in the lower frequencies (≤ 750 Hz), where the number of time-frequency bins was much lower than in the higher frequencies.

3.2.6 Paired-comparison rating

We used paired-comparison rating (a two-interval, seven-alternative forced choice paradigm) to measure noise annoyance, speech naturalness, and overall preference. This method was based on a standard of the International Telecommunication Union (ITU-T P.835, ITU-T 2003), according to which subjects must give separate ratings for the speech signal, background noise, and overall quality. The ITU standard uses a rating scale to measure quality. We chose to use paired comparisons instead because these are more sensitive to subtle differences in conditions (Böckenholt 2001).
For each pair of stimuli (the same sentences in two different processing conditions),
the subjects answered three questions. The first time they listened to the two frag-
ments A and B, the subjects were asked to concentrate on the speech and to rate in
which of the two fragments the speech was more natural and to indicate the strength
of the difference. After they made a choice, they listened to the same fragments again,
now concentrating on the annoyance of the noise and selecting the least annoying frag-
ment. The subjects could listen to both fragments again before they answered the third
question, but this was not required. For the third question, the subjects were asked
which fragment that they would prefer for prolonged listening. For each question,
there were seven possible answers, ranging from “A is much more natural/much less
annoying/much better” to “B is much more natural/much less annoying/much bet-
ter.” The seven choice categories were derived from the comparison category rating
method described in ITU-T P.800 (ITU-T 1996). The subjects were able to indicate no
difference between A and B. They were allowed to listen to the fragments as often as
they preferred before they answered each question.

All five conditions were paired with each of the other conditions, which resulted in 10
different stimulus pairs. Three runs of 10 comparisons were performed at both -4 and
+4 dB SNR, which resulted in a total of 60 comparisons per subject (10 Pairs x 3 Runs
x 2 SNRs). The choice for three runs of comparisons was based on previous studies
that used paired comparisons to determine preference for noise-reduction algorithms
or settings (Boymans and Dreschler 2000; Ricketts and Hornsby 2005; Luts et al. 2010).
All the subjects started with four training pairs. Subsequently, five subjects started
with all the comparisons at -4 dB SNR, and the other five subjects started at +4 dB
SNR.

3.2.7 Intelligibility
We measured intelligibility as the percentage of words that the subjects repeated cor-
rectly at -4 and -7 dB SNR. Each subject started with 13 training sentences containing
all five processing conditions, starting at +4 dB SNR. After every three sentences, the
SNR decreased one step (4 dB for the first two steps and 3 dB for the last step), ter-
minating with an SNR of -7 dB for the last four sentences. After this training list, we
used one list per processing condition per SNR to determine the intelligibility scores.
We balanced the order of conditions across all the subjects to minimize the possible
effects of training on the group data. We also balanced the measurement lists across
the conditions, to minimize possible effects of lists. Every new combination of process-
ing condition and SNR started with 3 training sentences, followed by 10 sentences that
were used to calculate the percentage of correct words. All words, not only key words,
were included in this calculation.
### 3.2.8 Listening-effort rating

The subjects rated the listening effort on a 9-point rating scale that ranged from “no effort” to “extremely high effort.” This test is similar to the test used by Luts et al. (2010) but differed in that our scale used five labeled buttons instead of seven. The labels were based on ITU-T P.800 methodology (ITU-T 1996). The subjects gave ratings for the five processing conditions at three SNRs (-4, 0, and +4 dB), thus for 15 different conditions. Each subject started with a practice run of 15 conditions. This practicing run was followed by three additional runs that we used for analysis.

### 3.3 Results

#### 3.3.1 Paired-comparison rating

Figure 3.3 shows the average rating score for each processing condition. We assigned scores from -3 to 3 for each condition, according to the ITU-T recommendation P.800 (ITU-T 1996). For instance, if the subject rated condition A slightly better than condition B, we assigned a score of 1 to condition A and a score of -1 to condition B. The scale for the noise annoyance is inverted in Figure 3.3. For each outcome, a symbol plotted above the zero line means a better performance on that judgment criterion.

Because we could not expect the scorings to represent a linear interval scale, we used the log-linear modeling approach for ordinal paired comparisons described by Dittrich et al. (2004) for the statistical analysis of the paired-comparison rating data. The model is a log-linear representation of the Bradley-Terry model (Bradley and Terry 1952) and is extended for paired-comparison data with multiple response categories, including a “no difference” option. By fitting this model to the paired-comparison data, we obtained estimates of the “worth” parameters, which describe the location of the five processing conditions on the subjects preference scale. This scale can be interpreted similarly to a ratio scale, thus providing not only the ranking of preference for the five conditions but also information regarding the strength of preference.

We estimated the worth parameters separately for noise annoyance, speech naturalness, and overall preference. We fitted a model for each individual run of 10 comparisons, which resulted in three models per subject per SNR per judgment criterion. We tested the goodness-of-fit for all of the models by comparing the obtained model with a saturated model (a model reproducing the data perfectly). All the p values were above 0.95, indicating a high agreement with the saturated model, and thus, all models could be accepted.
Figure 3.3: Mean rating scores derived from the paired-comparison data for the three judgment criteria and two SNRs. Scores from -3 to +3 were assigned with 0 indicating no difference; -1 and +1 indicating a minor difference; -2 and +2 indicating a moderate difference and -3 and +3 indicating a major difference. Error bars show the 95% confidence interval among subjects (without Bonferroni correction). Higher values indicate better performance. Horizontal bars indicate which processing conditions differ significantly from each other after Bonferroni correction for 10 comparisons.

We performed repeated-measures analysis of variance (ANOVA) on the estimated worth parameters for each judgment criterion separately (noise annoyance, speech naturalness, and overall preference) with SNR and processing condition as fixed effects and subject as a random effect. The resulting F statistics and p values are presented in Table 3.2. We found a significant effect of processing condition for each of the three judgment criteria. In addition, we found significant interactions between
processing condition and SNR for all three criteria. Because of the significant interaction between processing condition and SNR, we performed a subsequent repeated-measures ANOVA for each SNR separately, with processing condition as a fixed effect and subject as random effect. The resulting values for F and p are also given in Table 3.2. The effect of processing condition was significant for each judgment criterion at both SNRs, except for the speech naturalness at +4 dB SNR. The horizontal lines in Figure 3.3 indicate which conditions differed significantly from each other after Bonferroni correction.

Table 3.2: Main analysis of variance outcomes for the paired-comparison results.

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>F</th>
<th>p</th>
<th>F</th>
<th>p</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise annoyance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processing condition</td>
<td>4</td>
<td>24.57</td>
<td>&lt; 0.001</td>
<td>3.21</td>
<td>0.023</td>
<td>5.61</td>
<td>0.001</td>
</tr>
<tr>
<td>SNR</td>
<td>1</td>
<td>0.02</td>
<td>0.892</td>
<td>2.68</td>
<td>0.136</td>
<td>1.34</td>
<td>0.276</td>
</tr>
<tr>
<td>Processing condition x SNR</td>
<td>4</td>
<td>8.23</td>
<td>&lt; 0.001</td>
<td>4.42</td>
<td>0.002</td>
<td>8.79</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>+4 dB signal-to-noise ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processing condition</td>
<td>4</td>
<td>18.26</td>
<td>&lt; 0.001</td>
<td>8.21</td>
<td>&lt; 0.001</td>
<td>3.65</td>
<td>0.014</td>
</tr>
<tr>
<td>+4 dB signal-to-noise ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processing condition</td>
<td>4</td>
<td>24.29</td>
<td>&lt; 0.001</td>
<td>0.33</td>
<td>0.854</td>
<td>8.04</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

3.3.2 Intelligibility

The left panel in Figure 3.4 shows the percentage of words correctly repeated averaged over all 10 subjects. For statistical analysis, we transformed these percentages to rationalized arcsine units (Studebaker 1985) and subsequently performed a repeated-measures ANOVA on the transformed data with SNR and processing condition as fixed effects and subject as a random effect. We found significant effects of SNR (F[1, 9] = 39.0, p < 0.001) and processing condition (F[4, 36] = 4.4, p = 0.005). Post hoc Bonferroni-corrected pairwise comparisons showed that the scores for NR2 were significantly worse than those for NR4 (uncorrected p = 0.0045). The differences between the conditions are shown in the right panel of Figure 3.4, in which the scores of each noise-reduction condition are plotted relative to the scores of the unprocessed condition.

3.3.3 Listening-effort rating

The left panel in Figure 3.5 shows the mean listening-effort ratings assigned by the 10 subjects. Note that a higher value means that the listening effort was lower. To satisfy the ANOVA criteria, we transformed the listening-effort ratings with an arcsine transformation.
Figure 3.4: Left panel: Mean and 95% confidence interval of the percentage of words correctly repeated by the 10 subjects at -7 and -4 dB SNR. Right panel: Mean results from both SNRs relative to the unprocessed condition. Horizontal bars indicate which processing conditions differ significantly from each other after Bonferroni correction for 10 comparisons.

Figure 3.5: Left panel: mean and 95% confidence intervals of the listening effort ratings assigned by the 10 subjects at -4, 0 and +4 dB SNRs. Right panel: mean results from all 3 SNRs relative to the unprocessed condition. Horizontal bars indicate which processing conditions differ significantly from each other after Bonferroni correction for 10 comparisons.

We performed a repeated-measures ANOVA with SNR and processing condition as fixed effects and subject as a random effect. We found significant effects of SNR (F[2, 18] = 155.4, p < 0.001) and processing condition (F[4, 36] = 6.0, p < 0.001). Post hoc pairwise comparisons with Bonferroni correction showed that NR1 and NR4 involved significantly lesser effort than NR2 and NR3 did (uncorrected p < 0.05). The right panel of Figure 3.5 shows the differences between each processing condition and the unprocessed condition averaged across all three SNRs.
3.3.4 Relations between outcome measures

Because all the outcomes were measured at -4 dB SNR, we used the (transformed) data from this SNR to determine whether the overall preference was related to the other outcome measures. For noise annoyance, speech naturalness, overall preference (worth estimates), and listening effort, we calculated for each subject the average of the three repeats. Thus, we obtained one value per condition per subject for each outcome measure (and thus 5 processing conditions x 10 subjects = 50 values per outcome measure). We standardized the data by subtracting the mean and dividing by the standard deviation so that all the outcome measures had zero mean and a standard deviation of 1. Figure 3.6 shows the standardized results for all the outcome measures at -4 dB SNR.

We calculated Pearsons correlation coefficients between the overall preference and each outcome. After Bonferroni correction, we found significant correlations between the overall preference and the noise annoyance ($r = 0.48, n = 50, p < 0.001$) and between the overall preference and the speech naturalness ($r = 0.50, n = 50, p < 0.001$). Noise annoyance and speech naturalness were not significantly correlated to each other ($r = -0.04, n = 50, p = 0.76$). Overall preference was not significantly correlated to intelligibility scores ($r = 0.01, n = 50, p = 0.94$) or listening effort ($r = 0.26, n = 50, p = 0.07$).

To determine whether overall preference could be predicted by noise annoyance and speech naturalness, we applied linear regression analysis to the worth estimates. We performed this analysis both on the group and individual levels. We used the data from each individual run of 10 comparisons, which resulted in 15 values (5 processing
conditions x 3 runs) for noise annoyance, speech naturalness, and overall preference per subject per SNR. For the group analysis, this process resulted in 150 values per judgment criterion (10 subjects x 15 worth values). We used backward stepwise regression with thresholds of 0.05 for entering or removing terms. The dependent variable was the overall preference, and the independent variables were the noise annoyance and speech naturalness. In this way, we could determine whether the overall preference could be predicted by either noise annoyance or speech naturalness, by these factors together, or by none of these factors.

The left panel in Figure 3.7 shows the standardized regression coefficients ($\beta$) for noise annoyance and speech naturalness for each individual and for the entire group at -4 dB SNR. Because the data were standardized, a higher coefficient indicates a greater effect of that variable on the overall preference. Thus, for all of the subjects together (“group”), both noise annoyance and speech naturalness contributed equally to the overall preference ($\beta$ noise annoyance = 0.54 and $\beta$ speech naturalness = 0.57), together explaining 56% of the variance in overall preference ($R^2 = 0.56$). For subjects 1, 2, 4, and 8 noise annoyance and speech naturalness were also both included in the best model (nonzero $\beta$ coefficients, see Figure 3.7). The $R^2$ values for subjects 1, 2, 4, and 8 were, 0.79, 0.88, 0.91, and 0.76, respectively. In contrast, for subject 3, neither noise annoyance nor speech naturalness was included. For the remaining five subjects, only one of the variables remained in the best regression model: speech naturalness for subjects 5, 6, and 10 (with $R^2$ 0.83, 0.60, and 0.44, respectively) and noise annoyance for subjects 7 and 9 (with $R^2$ 0.41 and 0.49, respectively).

**Figure 3.7:** Standardized regression coefficients ($\beta$) for noise annoyance (vertical axis) and speech naturalness (horizontal axis) at -4 dB SNR (left panel) and +4 dB SNR (right panel). Mean and 95% confidence intervals were given for each individual subject (circles) and for the group results (diamonds).
Because worth estimates for noise annoyance, speech naturalness, and overall preference were also measured at +4 dB SNR, we performed the same regression analysis at this SNR. This process yielded the coefficients represented in the right panel of Figure 3.7. For the group, the noise annoyance was weighted more than the speech naturalness ($\beta_{\text{noise annoyance}} = 0.77$ and $\beta_{\text{speech naturalness}} = 0.29$), together explaining 78% of the variance in overall preference ($R^2 = 0.78$). For 5 of the 10 subjects, noise annoyance was the only explaining variable, and for 2 subjects, speech naturalness was the only explaining variable. For the remaining three subjects, both factors were included in the best model. $R^2$ values for the individual regression models at +4 dB SNR ranged between 0.58 and 0.98 and were all higher than at -4 dB SNR (except for subject 8, for whom $R^2$ was 0.04 lower at +4 dB SNR than at -4 dB SNR).

3.4 Discussion

With respect to our research questions, we can summarize our findings as follows:

Q1. Hearing aid noise reduction is able to reduce the annoyance of babble noise for normal-hearing listeners at -4 dB SNR (all four hearing aids) and at +4 dB SNR (three of the four hearing aids). At -4 dB SNR, however, two of the four noise-reduction systems also reduce speech naturalness. The noise-reduction systems differ from one another in how strongly they reduce noise annoyance and preserve speech naturalness.

Q2. Normal-hearing listeners prefer noise reduction over no noise reduction within two of the four hearing aids at +4 dB SNR. For our selection of hearing aids, noise reduction provides no statistically significant benefit in terms of intelligibility or listening effort compared with no noise reduction. Compared with each other, however, noise-reduction systems differ mutually in terms of all the outcome measures.

Q3. The overall preference of normal-hearing listeners correlates to noise annoyance and speech naturalness, but not to intelligibility or listening effort. There are differences among individual listeners in whether they place more weight on noise annoyance or on speech naturalness in determining their overall preference.

3.4.1 Noise annoyance, speech naturalness, and overall preference

This study is, to our knowledge, the first in which the perceptual effects of noise-reduction systems from different hearing aids are directly compared with each other. The recording and filtering technique developed in Chapter 2 allowed us to compare all combinations of noise-reduction systems directly with each other in a paired-
comparison design. Even within the small sample of hearing aids included, the measurements show that noise-reduction systems differ perceptually, as was previously suggested based on technical differences (Bentler and Chiou 2006; Hoetink et al. 2009).

The paired-comparison data (Figure 3.3) show that noise-reduction systems were able to reduce the noise annoyance at +4 dB SNR without affecting the speech naturalness. In the more difficult listening situation of -4 dB SNR, however, the results reveal a trade-off between noise reduction and speech distortion. Apparently, at this lower SNR, it was harder for the noise-reduction system to differentiate between speech and noise, so the reduction of noise was accompanied by distortion of the speech signal. Indeed, Figure 3.1 shows that the gain applied by the four noise-reduction systems differs between the two input SNRs (-4 and +4 dB). For instance, in the right column of Figure 3.1 we see less gain reduction (blue areas) in the spectrograms of NR1 and NR2 during the moments that speech is present (speech presence is visible from the time-signal plots shown above the spectrogram), indicating that the noise reduction does not reduce the gain during speech presence. In the left column, these blue areas for NR1 and NR2 are smaller than in the right column, indicating that at -4 dB SNR the noise reduction had more difficulty recognizing the speech in the speech-plus-noise mixture. This implies that the main cause for the reduced speech naturalness at -4 dB SNR is the unwanted suppression of the speech signal. In addition, the quick changes in gain introduced by NR1 and NR2 may have caused distortions to the speech signal. The underlying mechanisms for NR3 and NR4 seem to differ strongly from those of NR1 and NR2. NR3 reduces the gain for frequencies of up to 500 Hz, which seems to be independent of the presence of speech. This low-frequency gain reduction by NR3 is much stronger at -4 dB SNR than at +4 dB SNR (Table 3.1, Figure 3.1 and Figure 3.2: at -4 dB SNR the gain is reduced up to 9 dB, whereas at +4 dB SNR the maximum lies at approximately 6 dB). This low-frequency gain reduction is likely the cause of the reduced speech naturalness for NR3 at -4 dB SNR. NR4 only reduces the gain between 1 kHz and 2 kHz at -4 dB SNR and does not reduce the gain at +4 dB SNR. The small amount of gain reduction by NR4 at -4 dB SNR seemed to be able to reduce noise annoyance, although to a lesser extent than that by NR1 and NR2. The advantage of NR4 at -4 dB SNR was that the speech naturalness was kept intact, so that it was not less preferred than NR1 and NR2. However, at +4 dB SNR, NR4 lacks the advantage of reduced noise annoyance, which was found most clearly for NR1 and NR2.

Our results may help to understand the diverging results from previous studies. First, we found that subjects preferred noise reduction on over noise reduction off in only two of the four hearing aids. The four studies mentioned in the Introduction used another type of hearing aid, which may have contributed to their apparently conflicting
results. Second, our results showed significant preferences for noise reduction on over noise reduction off only at +4 dB SNR and not at -4 dB SNR. Indeed, the two studies that found a significant preference for noise reduction on over noise reduction off both measured preference at positive SNRs (Boymans and Dreschler 2000, at +5 dB SNR; Ricketts and Hornsby 2005, at +1 and +6 dB SNR), as opposed to Alcântara et al. (2003), who measured mainly at negative SNRs and did not find changes in quality or comfort because of noise reduction. In the study by Bentler et al. (2008), the measurement conditions differed per listener so that no systematic effect of the SNR could be derived. Although numerous other factors (for instance, the type of noise, other hearing aid characteristics, and hearing ability), play a role, we conclude that differences among noise-reduction systems and their dependency on SNR should be considered in the interpretation of noise-reduction studies.

3.4.2 Intelligibility
Our finding that none of the noise-reduction systems changed intelligibility scores compared with unprocessed corresponds to previous findings (Bentler 2005; Nordrum et al. 2006). We also found that intelligibility differed slightly between two noise-reduction systems. Scores for NR2 and NR4 differed on average 8.8%, which was just significant (p = 0.0045, with Bonferroni-corrected \( \alpha = 0.005 \)). One should keep in mind that this difference was measured in a laboratory situation with only one noise type and at the most sensitive point of the psychometric function (50%). Thus, it remains to be seen whether this improvement leads to an actual benefit in real-life situations.

Nordrum et al. (2006) also compared the effect of noise-reduction systems from different hearing aids on intelligibility but did not find significant differences among noise-reduction systems. The authors did not equalize other hearing aid characteristics, so results from different hearing aids could not directly be compared with each other. Furthermore, all previous intelligibility measurements with hearing aid noise reduction were performed with hearing-impaired listeners. The difference among listeners complicates comparison with our results because we measured at lower SNRs.

Luts et al. (2010) compared single-microphone noise-reduction systems that were developed and optimized for use in hearing aids. For normal-hearing subjects and the same sentence material and background noise as we used, the authors found an SRT\(_{50}\) (the SNR at which 50% of the sentences were correctly repeated) of -5.2 dB SNR in the unprocessed condition for normal-hearing listeners. The two single-microphone noise-reduction systems they evaluated did not change this value. Although the measurement methods differed (i.e., sentence scoring in an adaptive procedure versus word scoring at a fixed SNR), the results of Luts et al. were in agreement with ours.
3.4.3 Listening effort

Noise reduction did not change listening effort compared with the unprocessed condition. If noise reduction was on, two of the noise-reduction systems required slightly more effort than the other two systems. In the Introduction, we mentioned two studies that determined listening effort for listening with noise reduction on and noise reduction off within a single hearing aid. One study found a reduction in listening effort because of noise reduction (Bentler et al. 2008), and the other study found no effect (Alcàntara et al. 2003). On the basis of our results, it seems reasonable that, other than factors like noise type and SNR, the differences among noise-reduction systems contributed to differences among the results from different studies.

Although there is a common assumption that noise reduction may reduce listening effort compared with no noise reduction, there is little evidence confirming this assumption, and our data did not confirm it. The main difficulty in studying listening effort is the lack of a proper method to measure listening effort (Bentler 2005; Edwards 2007; Lunner et al. 2009). Such a test would ideally be able to evaluate noise reduction in situations relevant for the user, thus at SNRs where the speech is highly intelligible. In these situations speech-intelligibility tests suffer from ceiling effects, but the effort required to obtain the same intelligibility score may differ among situations. For instance, Sarampalis et al. (2009) used a dual-task paradigm to measure listening effort and found decreasing response time (indicating decreasing listening effort) with increasing SNR, even at SNRs where speech was highly intelligible. Although the dual-task paradigms such as in the example of Sarampalis et al. seem promising, the use of a secondary task in the nonauditory domain complicates the interpretation of the results and often requires a specialized test equipment. The disadvantage of our method was that we found a ceiling effect for listening effort at the SNR where speech was highly intelligible (+4 dB). Thus, a more appropriate method is required to measure the potential effects of noise reduction on listening effort.

3.4.4 Relations between all outcome measures

NR4 performed better than the other noise-reduction conditions in terms of intelligibility and listening effort (Figure 3.6), although it had the weakest gain reduction (Table 3.1). In contrast, NR2 applied the strongest gain reduction (Table 3.1) and reduced noise annoyance more than any other condition but did not perform well with respect to intelligibility and listening effort (Figure 3.6). This difference may be explained by the fact that the stronger reduction in noise by NR2 also affected the speech more severely. Considering these results, it is surprising that NR3, which reduced speech naturalness, preference, and listening effort the most, did not have the worst intelli-
gibility score. However, from Figure 3.1c and Figure 3.2c, it is clear that NR3 reduces gain for lower frequencies (≤ 500 Hz), while increasing the gain for higher frequencies (> 500 Hz), with no clear relation to the presence or absence of speech. Thus, the reduction in speech naturalness for NR3 was most likely caused by the low-frequency gain reduction, in contrast to the reduced speech naturalness for NR2, which was caused by the suppression of different speech fragments and quick changes in gain. Although our normal-hearing subjects did not prefer the spectral shaping caused by NR3 and they did not benefit from it either in terms of speech intelligibility, this might be different for hearing-impaired listeners. For the normal-hearing subjects, the audibility of all stimuli was maintained. However, hearing-impaired listeners could benefit from the small increase of gain that NR3 applies for frequencies above 1 kHz. These frequencies have been shown to be important for speech perception in noise (Smoorenburg 1992), so maintaining or increasing the audibility of this part of the signal could be beneficial for hearing-impaired subjects.

3.4.5 Relations among preference, noise annoyance, and speech naturalness

We found that the overall preference of our subjects was related to noise annoyance and speech naturalness. This finding corresponds with results of other studies. For instance, Hu and Loizou (2007b) evaluated different speech-enhancement algorithms using the ITU-T P.835 methodology, according to which subjects give ratings for the background intrusiveness, the signal distortion, and the overall quality of sound samples. The authors also performed linear regression on the ratings, which resulted in a noise coefficient of 0.37 and a speech coefficient of 0.57. The authors concluded that listeners integrate the effects of both signal and background distortions when assigning ratings for overall quality, but that they seem to place more emphasis on the speech distortion than on the background noise. Marzinzik (2000) used paired comparisons with the same three judgment criteria and also concluded that speech distortions counterbalanced the reduction of noise in the overall preference judgments. However, he performed no further analysis to determine the weighting between both parameters.

To date, studies have only looked at group results and provided no insight into possible individual differences. However, Houben et al. (2012) performed paired comparisons on different settings for the strength of noise-reduction in an algorithm designed for hearing aids and found significant interindividual differences in preferences among normal-hearing subjects. Because stronger noise reduction introduces more speech distortion, the authors hypothesized that listeners differ in their preference for a trade-off between noise annoyance and speech distortion. Our regression results support this hypothesis. Although the preference of some subjects seems to be a balanced weighting between noise annoyance and speech naturalness, for other subjects one of these
Chapter 3

factors was clearly more decisive than the other. Thus, we hypothesize that whereas some subjects accepted a degradation of speech quality to reach a less-noisy situation, others rejected noise reduction as soon as the reduction was at the cost of speech naturalness. The finding of individual weighting of background noise and speech quality agrees with findings of Versfeld et al. (1999), where individual subjects differed in which factor was most decisive for their overall preference for different types of signal processing and distortion: intelligibility was balanced against clarity, and distortion of the signal was balanced against the amount of added background noise.

The differences between the left and right panels of Figure 3.7 show that several subjects seem to be inconsistent in whether they place more weight on speech naturalness or noise annoyance. The majority of these subjects (subjects 1, 4, 5, and 10) placed more weight on noise annoyance at +4 dB SNR than they did at -4 dB SNR. As shown in Figure 3.3, differences in noise annoyance were larger at +4 dB SNR than at -4 dB SNR, whereas differences in speech naturalness were smaller at +4 dB than at -4 dB SNR. If subjects do not perceive a reduction in speech naturalness, this aspect will no longer play a role in their choice. Therefore, it is not surprising that noise annoyance gets more weighting at +4 dB SNR. However, for half the subjects speech naturalness still played a role in their preference at +4 dB SNR. Thus, it seems that speech naturalness is not completely unaffected at +4 dB SNR. For several subjects (subjects 2, 3, 6, 8, and 10), the small changes in speech naturalness were still too disturbing to fully benefit from the reduction in noise annoyance.

With respect to the effect of SNR, it is remarkable that Hu and Loizou (2007b) found that their normal-hearing listeners placed more emphasis on speech distortion rather than on the noise when judging the overall quality of stimuli at +5 and +10 dB SNR. These authors evaluated different state-of-the-art noise-reduction algorithms, which are not yet implemented in hearing aids. The implementations used in that study were given in Loizou (2007). We applied these algorithms to our noisy speech and listened to these stimuli in comparison to our own hearing aid recordings. It seemed that Loizou’s algorithms removed more noise and that they affected the speech quality more than our selection of hearing aid noise-reduction algorithms. It seems that our hearing aid noise reduction was fine-tuned to preserve speech quality, whereas Loizou’s algorithms were primarily aimed at reducing the background noise.

Because noise annoyance and speech naturalness were measured using the same procedure as was used to measure overall preference, they were more likely to be correlated to preference than intelligibility and listening effort. For each pair of processing conditions, subjects answered the three questions about noise annoyance, speech naturalness, and overall preference successively, so that the judgment for overall preference
could correlate strongly with their previous answers on the speech and noise criteria. Furthermore, all three answers were with regard to the same sentence, whereas noise reduction might act differently on other sentences. To test whether the correlations we found were caused by the experimental design, we shuffled the answers across all the repeats. Thus, we considered whether the values for noise annoyance from the first run combined with that for speech naturalness from the second run could predict the overall preference from the third run, and so on. At the group level, the correlations between overall preference and both other judgment criteria were reduced but still significant. At -4 dB SNR, the group regression coefficient for speech naturalness was slightly higher than the coefficient for noise annoyance, whereas at +4 dB SNR, the noise annoyance was clearly dominating. At an individual level, there were more subjects for whom none of the factors were included in the model (six subjects at -4 dB SNR and two subjects at +4 dB SNR). For the remaining subjects, the effect of SNR was still visible, with more emphasis on speech naturalness at -4 dB SNR and on noise annoyance at +4 dB SNR. From these results, we conclude that the succession of the three questions indeed enhanced the relationship between the overall preference and the noise and speech criteria. However, the effect of SNR on the weighting of the judgment criteria and the differences among subjects in their preferred weighting remained consistent when the successive answers were separated.

3.4.6 Limitations

This study was the first exploration into a largely uninvestigated area, and the conclusions only apply for the limited conditions that we measured.

First, our study population differs from the target population of hearing aid users. The results obtained with our normal-hearing subjects might be representative of listeners with a conductive hearing loss. For listeners with a sensorineural hearing loss, the evaluation becomes more complicated because hearing aids for this type of hearing loss apply dynamic-range compression. The interactions of noise reduction with compression have been studied only occasionally thus far (Chung 2007; Anderson et al. 2009). The results obtained emphasize that the configuration of noise reduction and compression strongly influences the processing. The interactions between noise reduction and compression demand more exhaustive investigations, especially into the complex systems that are currently available for hearing aid users.

In addition, we evaluated just one type of noise and speech combined at a limited number of SNRs. To obtain a more complete impression of the effect of noise reduction, one should investigate a much broader range of speech, noise types, and SNRs (Houben 2011).
Last, we made measurements in a laboratory setting and presented the stimuli to the listeners via headphones. Field studies should reveal whether our results hold in real-life listening situations.

3.5 Conclusions

We conclude that noise reduction differs among hearing aids in the degree that they reduce the noise annoyance and the speech naturalness perceived by normal-hearing listeners. These differences among noise-reduction systems may explain the divergent results of previous studies on the effects of noise reduction on preference and listening effort. Differences in intelligibility were small, as shown in previous noise-reduction studies. Our results imply that it may be useful to give hearing aid users the possibility to compare different noise-reduction systems.

In addition, we conclude that individuals differ in their preferred weighting of noise annoyance and speech naturalness. This finding suggests that listeners may benefit from individualization of noise reduction in hearing aids, and supports the earlier statement that hearing aid users should have the possibility of comparing noise-reduction systems.

Clearly, the next step toward developing guidelines for clinicians to fit noise reduction in hearing aids should include listeners with sensorineural hearing loss, and a set-up to determine the effects of the interaction between noise reduction and compression. If such research would finally lead to fitting rules that help clinicians to actively select the best noise-reduction system and settings for individual listeners, hearing aid satisfaction may increase.