SCIENTIFIC ARTICLE

OPEN ACCESS

Using the phase inversion method and loudness comparisons for the evaluation of noise reduction algorithms in hearing aids

Hendrik Husstedt[1,*] , Jannine Kreyenhagen[2], Laura Langhof[3], Steffen Kreikemeier[3], Florian Denk[1], Simone Wollermann[1], and Marlitt Frenz[1]

[1] German Institute of Hearing Aids, Anschützstr. 1, 23562 Lübeck, Germany
[2] Academy of Hearing Acoustics, Bessemerstraße 3, 23562 Lübeck, Germany
[3] Aalen University, Beethovenstraße 1, 73430 Aalen, Germany

Abstract – The phase inversion method, a technical measurement procedure, is often used to evaluate the performance of noise reduction algorithms in hearing aids. However, a detailed comparison of these technical measurements with the perceived loudness is missing. Therefore, commercially available hearing aids of six different manufacturers were evaluated technically and in a study with 18 normal-hearing listeners. First, the output signals of the hearing aids with and without activated noise reduction were recorded in a test box. Then, the test subjects evaluated the perceived loudness of these recordings within multiple two alternative forced choice (2-AFC) tasks. During one task, the test subjects had to focus either on the speech or noise signal and were asked to select the louder of two signals, which both contained a mixture of speech and noise. These results provide not only the perceived SNR but also the perceived speech and noise levels. Comparing the results of the 2-AFC tasks and the phase inversion method basically shows good agreement. Nevertheless, a simple computation of the sound pressure level can lead to significant deviations. Therefore, another possibility for the analysis of the results of the phase inversion method to better match the perceived loudness is presented.

Keywords: Hearing aid evaluation, Single-channel noise reduction, Perceived loudness, Phase inversion method

1 Introduction

Modern digital hearing aids provide a complex signal processing with several features, e.g., multi-channel wide dynamic range compression (WDRC), feedback reduction, directional microphones, and noise reduction [1–3]. Since many of these terms are commonly used without any unique definition, the standardization project IEC 60118-16 ED1 "Electroacoustics – Hearing aids – Part 16: Definition and verification of hearing aid features" has been started by the International Electrotechnical Commission (IEC) [4]. Here, a noise reduction (NR) is defined as "feature of the signal processing of a hearing aid intended to reduce noise with respect to the absolute level or relative to the level of a target signal". In addition, a noise reduction for speech enhancement (NRSE) is defined as "feature of the signal processing of a hearing aid intended to increase the signal-to-noise ratio even if speech and noise are presented simultaneously from the same direction and have the same long-term average spectrum". In literature, this type of noise reduction is also denoted as single-microphone or single-channel noise reduction, since only one microphone input signal is required

[3]. For devices on the market, it is often not obvious how this feature is exactly realized [5], but commonly Wiener filters or spectral subtraction are used for the realization of a NRSE in hearing aids [1–3, 6].

Various studies with test subjects or technical measurements for the evaluation of a NRSE can be found in literature [7] which show that a NRSE mostly does not increase speech intelligibility (e.g., [8–12]) but can reduce listening effort. To determine the impact on listening effort, various approaches have been reported, e.g., measuring pupil dilation [13, 14], electroencephalography (EEG) [15, 16], electrodermal activity [17], response time [18, 19], using a dual task paradigm [20–24], or asking for subjective ratings, e.g., using the acceptable noise level (ANL) [25, 26], questionnaires [27], or adaptive scaling procedures [28]. For the technical evaluation of a NRSE in hearing aids, the prevalent procedure is the phase inversion method [29], whereas other procedures have also been reported, e.g., computing the modulation transfer function [30], using the percentile analysis [31], or performing a signal separation in the frequency domain [32].

One main advantage of the phase inversion method is that it enables the separation of speech and noise at the output of the hearing aid. Consequently, the individual

*Corresponding author: h.husstedt@dhi-online.de

levels of the processed speech and noise signals and thus the resulting SNR can be analyzed, e.g., as an average value or over frequency and time. Moreover, an extension of the phase inversion method can be used to analyze complex acoustic environments with multiple signals from different directions [33]. Thus, the phase inversion method is a powerful objective tool to evaluate noise reduction algorithms in hearing aids. Furthermore, it does not depend on test subjects, which greatly reduces effort, expenses, and increases repeatability. For an audiological interpretation, it is important that the results of this technical procedure reflect the experiences of human test subjects. Some studies reported that the changes in SNR as calculated with the phase inversion method were related to changes in acceptable noise level (ANL) [25, 26]. However, it is not clear if the output of the phase inversion method, the individual sound pressure levels of speech and noise, reflect the loudness perception of the individual signals in the mixture of speech and noise. Thus, a detailed comparison between the results of the phase inversion method and the perception of human test subjects is missing. For a direct comparison, experiments with human test subjects should yield comparable results to those gathered with the phase inversion method, i.e., the SNR as well as the individual speech and noise levels. Therefore, previous approaches such as assessing ANLs [34] are not applicable to reach the goals of this study where we want to (i) design an experiment with human test subjects which provides the SNR as well as the levels of speech and noise at equal perceived loudness within the mix; (ii) evaluate the NRSE of commercially available hearing aids of different manufactures with the phase inversion method and within the experiment with human test subjects; (iii) compare both results to each other; (iv) provide explanations and possible corrections if differences between both results will be observed.

For the first goal, our approach was using multiple two alternative forced choice (2-AFC) tasks with loudness comparisons where the level of speech and noise in a mixture were evaluated separately within multiple rounds. During the trials of one round, the test subjects had to focus either on the speech or noise signal and were asked to select the louder of two signals, which both contained a mixture of speech and noise. As reference for the comparison, individually recorded signals for each manufacturer with deactivated NRSE were used so that systematic errors, e.g., due to different transfer characteristics across the hearing aids, were compensated. Thus, the 2-AFC tasks of this work provide sound pressure levels of equal loudness individually for speech and noise by comparing recordings with activated and deactivated NRSE. For the second goal, we have evaluated the SNR improvement through the NRSE processing of commercially available hearing aids of six manufacturers with the phase inversion method, and compared these results with data from the 2-AFC tasks of 18 normal-hearing test subjects.

The individual evaluation of speech, noise, and SNR gives insights into the relation between the results of the phase inversion method and the perceived loudness of human test subjects and let us achieve our remaining goals.

The rest of the paper is organized as follows. After the introduction, the evaluation with the phase inversion method is presented including the test signals, measurement equipment, hearing aid models and settings, and the measurement results. Then, the 2-AFC loudness comparisons are explained including the test setup, test procedures, study design, and the measurement results. Finally, both results are discussed and compared with each other, and a conclusion is drawn.

2 Phase inversion method

2.1 Materials and methods

2.1.1 Test signals

As input signals, the international speech test signal (ISTS) [35] was presented together with the international female noise (IFnoise) [36]. Both speech and noise have the same long term average spectrum so that a linear frequency shaping does not affect the broadband SNR. The sound pressure levels of the speech signal $(s(t))$ was varied between 60 dB SPL and 75 dB SPL in steps of 2.5 dB, and the level of the noise signal $n(t)$ was chosen to cover an input SNR between -5 dB and 10 dB in steps of 2.5 dB. For each speech level and SNR, three mixed measurement signals:

$$x_1(t) = s(t) + n(t), \tag{1}$$

$$x_2(t) = s(t) - n(t), \tag{2}$$

$$x_3(t) = -s(t) - n(t), \tag{3}$$

each with a length of 60 s were created as suggested in the draft of the IEC 60118-16 ED1 [4]. After the processing of the hearing aid, the output signals $y_1(t)$, $y_2(t)$, and $y_3(t)$ were recorded, and the processed speech signal $s'(t)$, the processed noise signal $n'(t)$, and a verification signal $v'(t)$ were computed as,

$$s'(t) = 0.5(y_1(t) + y_2(t)), \tag{4}$$

$$n'(t) = 0.5(y_1(t) - y_2(t)), \tag{5}$$

$$v'(t) = y_1(t) + y_3(t). \tag{6}$$

The verification signal was used to check if the basic technical assumptions for the application of the phase inversion method were valid, but it does not provide any audiological verification of the phase inversion method. To neglect transient effects, the intervals between 30 s and 60 s of $s'(t)$, $n'(t)$, and $v'(t)$ were considered only.

2.1.2 Measurement equipment

All technical measurements were performed in an anechoic test box of type 4232 made by Brüel & Kjaer (Denmark). In a frequency range of 100 Hz to 10 kHz, the input sound pressure level was calibrated and equalized

within a tolerance of ± 2 dB. Moreover, the actual sound pressure was monitored with a reference microphone beside the sound inlets of the hearing aids, and the output signals were recorded with a 2 ccm coupler according to IEC 60318-5 [37] and an RME Fireface UC sound card. As reference microphone and for the 2 ccm coupler, two pressure-field microphones of type 4192 made by Brüel & Kjaer were used. The automation of the measurement and the data analysis were realized with a computer and Matlab R2018b.

2.1.3 Hearing aid models and settings

Commercially available behind-the-ear (BTE) hearing aids released between 2017 and 2019 of the six largest hearing aid manufacturers were considered (HA1–HA6). The programming of the devices should be similar to the reference test setting (RTS) as defined in IEC 60118-0 [38]. This means that all non-linear features are disabled and a linear gain is applied so that the output sound pressure level averaged at 1000 Hz, 1600 Hz, and 2500 Hz is 17 dB lower than the maximum output sound pressure level of the device averaged at the same frequencies. Most manufacturers provide a predefined setting for RTS, but in this predefined configuration no changes can be made as activating the NRSE. Therefore, we manually tried to program RTS to the devices. Nevertheless, differences to the predefined RTS provided by the manufacturer may occur, because not all features were accessible in the fitting software and could be disabled. Moreover, a linear gain of 10 dB below RTS was chosen to be sure that all hearing aids provide a linear gain for all input signals without reaching the output limit of the receiver. Since HA3 showed a strong nonlinear processing for higher speech levels, the gain of this device was reduced about 20 dB relative to RTS. In the following, "NRSE OFF" corresponds to the linear settings explained, and "NRSE ON" means that additionally the noise reduction for speech enhancement was activated at maximum level. For some hearing aids, there were multiple settings for the NRSE available. Here, different configurations were investigated and the setting with the highest SNR increase was used. However, not all possible configurations were tried so that other settings may lead to a higher SNR increase.

2.1.4 Evaluation of the separated noise signal within speech pauses

For the discussion of Section 4.3.2, we determined speech pauses of the ISTS, which are longer than 80 ms. To this end, we searched in the "clean" wave file of the ISTS for intervals, where the moving RMS with a window size of 20 ms is 15 dB below the RMS of the complete signal. Next, we considered the pauses within the analysis window from 30 s to 60 s only and computed the level of the noise signal separated with the phase inversion method at the output of the hearing aid over all pauses. In Figure 1, the relative noise level is exemplary plotted for HA1 and HA2 against time. Moreover, the amplitude of the ISTS of the wave file is plotted w.r.t. full scale and the speech pauses are depicted with a grey background color.
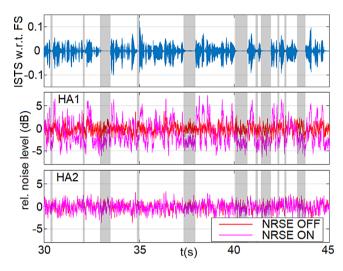

Figure 1. In the upper graph, the amplitude of the ISTS of the wave file is plotted w.r.t. full scale. In the middle and lower graph, the level of the noise signal separated with the phase inversion method at the output of the hearing aid is plotted exemplary for HA1 and HA2 against time. For this visualization, the moving RMS with a window size of 20 ms is computed. Moreover, the level of the complete time window from 30 s to 60 s is subtracted. In all three graphs, speech pauses are depicted with a grey background color.

2.1.5 Percentile sound pressure level of the separated speech signal

For the discussion of Section 4.3.3, the percentile sound pressure levels were computed for the separated speech signals according to IEC 60118-15 [39]. This means the sound pressure levels of the separated speech signals were computed within 125 ms time intervals with an overlap of 50%. Then, the sound pressure level below which a certain percentage of all sound pressure levels falls was determined. For instance, a 65th percentile sound pressure level of 45 dB SPL means that 65% of the sound pressure levels of all 125 ms time intervals are below 45 dB SPL. Furthermore, according to IEC 60118-15, the percentile sound pressure levels are computed separately within 1/3-octave bands with nominal center frequencies from 0.25 kHz to 6.3 kHz. However, we compute the percentile sound pressure levels for the broad band signals only. To this end, we considered the speech signals after the separation with the phase inversion method during the time window from 30 s to 60 s. Moreover, we normalized the sound pressure levels by subtracting the sound pressure level of the complete time window.

2.2 Results

2.2.1 Verification

All input signals were monitored with the reference microphone, and as verification, the phase inversion method was also performed for those recordings. This reveals that the absolute sound pressure levels of the separated speech and noise signals at the reference microphone for all

measurements are within a tolerance of ±0.3 dB, and that the SNRs are within a tolerance of ±0.15 dB. Moreover, the levels of the verification signals at the output of the hearing aids ($v'(t)$) were compared to the levels of the separated speech ($s'(t)$) and noise signals ($n'(t)$). As suggested in the draft of the IEC 60118-16 ED1 [4], the minimum of the levels of the separated speech and noise signals was always 10 dB above the level of the verification signal.

2.2.2 Broadband gain

First, the gain for NRSE OFF is evaluated for the separated speech and noise signals (not depicted in a figure). The gain is lowest for HA3 with 25 dB and highest for HA5 with 45 dB. Since a linear amplification is programmed, the gain should be equal for the separated speech and noise signal as well as for all input levels. Up to an input speech level of 70 dB SPL, the gain is constant over all input speech levels and the complete SNR range within a tolerance of ±1.5 dB for both the separated speech and noise signal. For speech levels of 72.5 and 75 dB SPL, the gain especially at lower SNRs is reduced for HA2, HA4, and HA5. The reduction for the speech signals is a little higher than for the noise signals. The highest reduction in gain of approx. 3.5 dB can be observed for the speech signal of HA2 at the highest input speech level of 75 dB SPL and the lowest SNR of −5 dB.

Next, the gain with NRSE ON for the separated speech and noise signals is considered (see Fig. 2). All hearing aids reduce the gain for both signals at low SNRs. The highest reduction in gain with decreasing SNR can be found for HA5 where the speech signal is reduced in a range of 8.6 dB and the noise signals in a range of 12.4 dB.

2.2.3 SNR change

To clearly see the effect of the NRSE, the difference between the input SNR (SNR_{in}) and the SNR at the hearing aid output (SNR_{HA}) is visualized in Figure 3 for NRSE ON. HA1, HA3, HA4, and HA5 provide a clearly noticeable SNR increase with a maximum in the range of 3.5–4.8 dB. For HA2 and HA6, the maximum SNR increase is 1 dB and 0.7 dB, respectively. Beside the maximum SNR increase also the corresponding input SNR is of interest for the user. HA5 shows the highest SNR increase for lower input SNRs between −5 and 0 dB, HA1 and HA3 have their maximum SNR increase around an input SNR of 2.5 dB, and HA4 shows the highest SNR increase for higher input SNRs between 5 and 10 dB.

3 2-AFC loudness comparisons

3.1 Materials and methods

3.1.1 Test setup

The 2-AFC measurements took place in an audiological test room with a ground area of $3.55 \times 2.52 \ m^2$ and a height of 2.5 m, which fulfills all requirements according to ISO 8253-2 [40]. The acoustic signals were presented with an
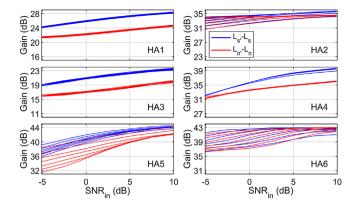

Figure 2. Gain with NRSE ON for all input speech levels for the separated speech ($L_{s'} - L_s$, blue curves) and noise signals ($L_{n'} - L_n$, red curves).
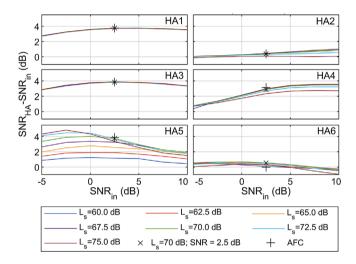

Figure 3. Difference between the input SNR (SNR_{in}) and the SNR at hearing aid output (SNR_{HA}) with NRSE ON. The SNR difference at an input speech level of 70 dB SPL and an input SNR of 2.5 dB is marked with "×", and the SNR difference of the sound files used for the 2-AFC loudness comparisons is depicted as "+" (see Sect. 3).

RME sound card of type "FIREFACE 802" and a Genelec loudspeaker of type "8020C" placed in 1 m distance to the test subjects. Furthermore, instructions were presented on an iiyama touch display of type "ProLite T1532MSC", and answers of the test subjects were received by typing on the screen. A computer with Matlab 2018b was used to automate the test procedures. Before the measurements, the levels of the signals were calibrated within a tolerance of ±1 dB SPL, and the frequency characteristic of the loudspeaker was equalized between 100 Hz and 10 kHz within a tolerance of ±2 dB both using Brüel & Kjaer's free field microphone of type "4190".

3.1.2 Test procedure

Signals to be evaluated

For the 2-AFC loudness comparisons, recordings of all six hearing aids with NRSE ON and NRSE OFF at an

input speech level of 70 dB SPL and an input SNR of 2.5 dB were considered which results in 12 signals to be evaluated. Input speech level and input SNR were selected such that most hearing aids provide almost maximum SNR increase according to the results of the phase inversion method (see "×" markers in Fig. 3). During the 2-AFC task, the signals to be evaluated were presented at a fixed sound pressure level of 65 dB SPL.

Due to technical reasons, the recordings of the signals to be evaluated during the 2-AFC tasks were not exactly the same as used for the technical evaluation at an input speech level of 70 dB SPL and an input SNR of 2.5 dB. To verify that the settings are comparable, the SNRs at the output of the hearing aids of the recordings used during the 2-AFC tasks are also depicted in Figure 3 with "+" markers. Beside small measurement tolerances, all SNRs are comparable except for HA6 where a difference of 0.55 dB can be noticed. The reason is that the characteristics of the noise reduction of HA6 depend on many different settings and after the study with the 2-AFC loudness comparisons, a configuration has been found leading to a higher SNR increase. Therefore, the results of HA6 with these settings are presented for the technical evaluation.

Comparison signals

For the second signal of the 2-AFC tasks, separate speech and noise signals were successively recorded for all six hearing aids with NRSE OFF using the same measurement equipment as explained in the previous section. Moreover, the same input levels (speech level of 70 dB SPL, input noise level of 67.5 dB SPL) were used as for the signals to be evaluated. During the 2-AFC task, the separate recorded speech and noise signals were adaptively mixed together and presented as comparison signals.

2-AFC task

For each of the 12 signals to be evaluated (also denoted as condition 1 (C1) to 12 (C12)), the test subjects had to separately compare the loudness of speech and noise in three different 2-AFC tasks which are denoted as 1. round, 2. round, and 3. round (see Figs. 4 and 5). During each round either the speech or noise level of the comparison signal was adaptively changed according to a 1-up/1-down rule. At the beginning, a step size of 3 dB, after the second reversal point a step size of 2 dB, and after the fourth reversal point a step size of 1 dB was used. One round was stopped, if either 12 reversal points or 40 trials were reached, and as result, the mean of the last four reversal points was computed.

1. round – comparison of noise

During the first round, the test subjects had to choose the signal where the noise is louder. Then, the level of the noise signal was adaptively changed for the comparison signal as described in the foregoing. The level of the speech signal was fixed at 63.1 dB SPL. As starting point,
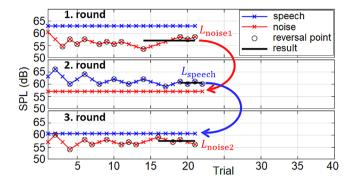

Figure 4. Visualization of the three different 2-AFC tasks, which are denoted as 1. round, 2. round, and 3. round.
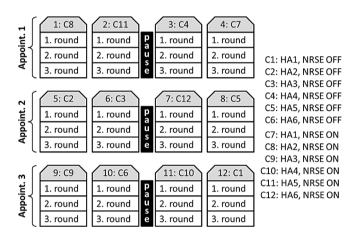

Figure 5. Flow chart of the test procedure exemplary for one test subject. Overall 36 2-AFC tasks were completed by the test subjects.

a noise level of 60.6 dB SPL was chosen so that the starting sum SPL was 65 dB and the starting SNR corresponds to the input SNR of 2.5 dB as used for the signals to be evaluated.

2. round – comparison of speech

During the 2. round, the subjects had to choose the signal where the speech level is louder. Here, a fixed noise level was chosen for the comparison signal equal to the result of the 1. round. The speech level started at the level of the 1. round (63.1 dB SPL) and was adaptively changed according to the 2-AFC procedure.

3. round – comparison of noise

During the 3. round, again the noise level was compared and adaptively changed for the comparison signal. As starting point, the result of the 1. round was chosen. The speech signal was set to a fixed level equal to the result of the 2. round. Thus, the speech level during this round corresponds to the experienced speech level of the signal to be evaluated. Comparing the results of this round with the 1. round should give insights, whether or not the speech level has an effect on the loudness rating of the noise.

Time windows

The recordings of the signals to be evaluated and the individual recordings of speech and noise used for the creation of the comparison signals have a duration of 60 s. Since this is too long to be used during a 2-AFC task, 40 different 2 s long pieces were cut out of the signals, and a fade-in and fade-out was applied. To neglect transient effects, the pieces were cut out only between 30 s and 60 s, and different pieces may overlap partly. Moreover, during one trial of the 2-AFC task, only paired pieces were compared, which were cut out at the same position in the ISTS.

3.1.3 Study design

Overall 18 normal hearing test subjects (11 female, 7 male) with an age between 23 and 35 years (mean 29) participated in the study. Before the actual test started, a pure tone audiometry was performed at frequencies of 0.125, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, 8 kHz and all test subjects fulfilled the definition of normal hearing according to ISO 8253-2 [41]. The order of the 12 conditions to be evaluated was randomized using a Latin square. This means for the first 12 test subjects each of the signals to be evaluated was once presented at first, second, . . . , and twelfth position. The completion of all three rounds for one condition took approx. 10 min. The evaluation of all 12 signals was split into three appointments so that at one appointment four signals were evaluated (see Fig. 5). Moreover, between the four signals of one appointment a break of approx. 5 min was made.

3.2 Results

3.2.1 General

All three appointments were completed by the test subjects within a maximum of 11 days, and on average within in 3.7 days. Different appointments were usually at different days except for one test subject where the first appointment was in the morning and the second appointment in the evening.

The 2-AFC tasks were stopped for all test subjects and all rounds after 12 reversal points, i.e., no test subject reached 40 trials. On average 20.6 trials were required to reach 12 reversal points.

3.2.2 Performance at different appointments

The randomization according to a Latin square is complete only for the first 12 test subjects, i.e., each condition was presented equally often at each position. Consequently, only the first 12 test subjects are considered to compare the performance of the test subjects at different appointments. To neglect variations due to the condition, the average over all test subjects per condition was computed and this average was subtracted from the results of
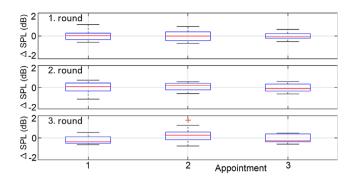

Figure 6. Performance of the first 12 test subjects at different appointments where the inter-individual variation and the variation due to the condition were eliminated.

each test subject separately for each condition. Next, to neglect also inter-individual variation, the average over all conditions per test subject was computed and this value was subtracted from the results of each test subject. Finally, the results of all four conditions at one appointment were averaged. These results are visualized in Figure 6 for all three rounds and all three appointments. Testing on normality using the Shapiro–Wilk test with Bonferroni correction indicates that all data are normally distributed. A two-way repeated measures ANOVA was performed using the factors "appointment" and "round", which shows no significant effects neither for the factor "appointment", $F(2,22) = 1.79$, $p = 0.19$, nor for the factor "round", $F(2,22) = 4.56\text{e-}29$, $p = 1$.

3.2.3 Performance over time during one appointment

At one appointment four conditions were evaluated by the test subjects and between the second and third condition a break of approx. 5 min was made (see Fig. 5) so that one appointment took approx. 1 h. To compare the performance over time during one appointment, the results of the first 12 test subjects were considered and the inter-individual variation and the variation due to the condition were eliminated as described in the foregoing section. Then, the results of each of the four conditions tested at one appointment were averaged over all three appointments (see Fig. 7) w.r.t. the order of the condition tested. Testing on normality using the Shapiro–Wilk test with Bonferroni correction indicates that all data are normally distributed. A two-way repeated measures ANOVA was performed using the factors "order/time" and "round", which shows no significant effects neither for the factor "order/time", $F(3,33) = 1.79$, $p = 0.07$, nor for the factor "round", $F(2,22) = 4.55\text{e-}29$, $p = 1$.

3.2.4 Comparison of the noise level evaluated during the 1. and 3. round

All noise levels evaluated during the 1. round (L_{noise1}) and 3. round (L_{noise2}) for all 18 test subjects were tested on normality using the Shapiro–Wilk test with Bonferroni
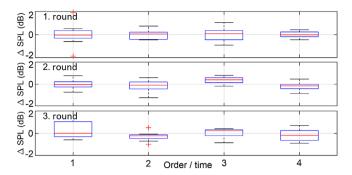
Figure 7. Performance of the first 12 test subjects over order/time during one appointment where the inter-individual variation and the variation due to the condition were eliminated.

correction, which indicates that 23 of 24 data sets are normally distributed. Thus, we decided to still use parametric tests. A two-way repeated measures ANOVA was performed using the factors "condition" and "1. or 3. round", which shows a significant effect for both the factor "condition", $F(11,187) = 173.8$, $p < 0.001$***, and the factor "1. or 3. round", $F(1,17) = 6.96$, $p = 0.02$*. The interaction between both factors was not significant, $F(11,187) = 1.19$, $p = 0.06$. (Throughout this work, significant effects are indicated with stars where "*" corresponds to $p < 0.05$, "**" to $p < 0.01$, and "***" to $p < 0.001$.) Since the differences between the 1. and 3. round is of interest, a pairwise comparison for all conditions between both rounds with a paired t-test and Bonferroni correction was performed, which shows no significant differences (see Fig. 8). Finally, the average over all conditions was computed for all test subjects and the difference between the noise level of the 3. round (L_{noise2}) and 1. round (L_{noise1}) was computed ($M = 0.28$ dB, $SD = 0.45$ dB), which is depicted on the right side of the boxplot of Figure 8. Testing on normality using the Shapiro–Wilk test indicates that this data is normally distributed. Performing a paired t-test shows that the noise level during the 3. round is significantly higher rated as during the 1. round $t(17) = 2.64$, $p = 0.02$*.

3.2.5 Noise level

In Figure 9, the noise level evaluated by the test subjects during the 3. round (L_{noise2}) for all 18 test subjects and the results of the phase inversion method are visualized together. For all conditions, Shapiro–Wilk tests with Bonferroni correction were performed, which indicate that 9 of 12 data sets are normally distributed. Hence, we still used t-tests to compare the results of the 2-AFC tasks with the results of the phase inversion method and applied a Bonferroni correction. For the conditions with NRSE OFF, there was no significant difference found between the results of the 2-AFC tasks and the phase inversion method. For the conditions with NRSE ON, there are high and highly significant differences for three of six hearing aids (HA1, HA3, HA5), where the loudness of the noise is rated softer during the 2-AFC task.
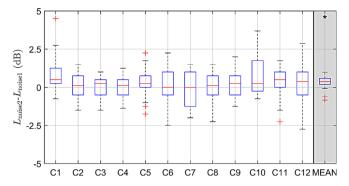

Figure 8. Difference between the noise level evaluated in the 3. round (L_{noise2}) and 1. round (L_{noise1}) for all 12 conditions and all 18 test subjects. For details about the conditions see Figure 5. Moreover, the average over all conditions is depicted on the right side of the boxplot.
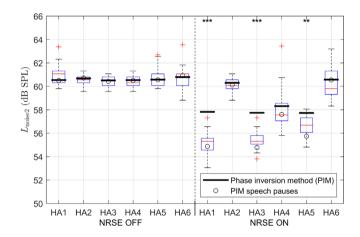

Figure 9. Sound pressure levels of noise as result of the 2-AFC tasks during the 3. round (L_{noise2}) for all 18 test subjects depicted as boxplots. In addition, the sound pressure levels of noise as result of the phase inversion method (PIM) are shown with black lines. Furthermore, the noise levels of the phase inversion method were evaluated in the speech pauses only, and these noise levels are indicated by circles.

3.2.6 Speech level

In Figure 10, the speech level evaluated by the test subjects during the 2. round (L_{speech}) for all 18 test subjects and the results of the phase inversion method are visualized together. For all conditions, Shapiro–Wilk tests with Bonferroni correction were performed, which indicate that 9 of 12 data sets are normally distributed. Thus, we still used t-tests to compare the results of the 2-AFC tasks with the results of the phase inversion method and applied a Bonferroni correction. For the conditions with NRSE OFF, there was no significant difference found between the results of the 2-AFC tasks and the phase inversion method. For the conditions with NRSE ON, there are highly significant differences for four of six hearing aids (HA1, HA3, HA4, HA5), where always the loudness of the speech is rated softer during the 2-AFC task.
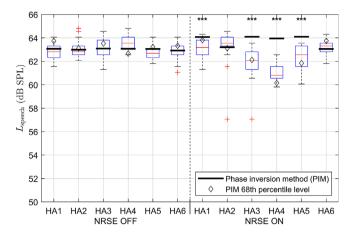
Figure 10. Sound pressure levels of speech as result of the 2-AFC tasks during the 2. round (L_{speech}) for all 18 test subjects depicted as boxplots. In addition, sound pressure levels of speech as result of the phase inversion method (PIM) are shown with black lines. Furthermore, the 68th percentile sound pressure levels were computed for the speech signals of the phase inversion method depicted with diamond markers.
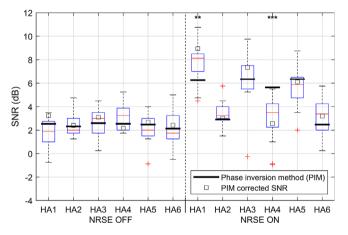

Figure 11. SNR computed as the difference between the evaluation of the speech level with the 2-AFC tasks during the 2. round and the noise level during the 3. round ($L_{speech} - L_{noise2}$) for all 18 test subjects depicted as boxplots. In addition, the results of the SNR of the phase inversion method (PIM) are shown with black lines, and the results of the phase inversion after a correction are indicated with square markers.

3.2.7 SNR

In Figure 11, the SNR computed as the difference between the evaluation of the speech level within the 2-AFC tasks during the 2. round and the noise level during the 3. round ($L_{speech} - L_{noise2}$) for all 18 test subjects and the results of the phase inversion method are visualized together. For all conditions, Shapiro–Wilk tests were performed with Bonferroni correction, which indicate that 10 of 12 data sets are normally distributed. Hence, we still used t-tests to compare the results of the 2-AFC tasks with

the results of the phase inversion method and applied a Bonferroni correction. For the conditions with NRSE OFF, there was no significant difference found between the results of the 2-AFC tasks and the phase inversion method. For the conditions with NRSE ON, there are significant differences for two of six hearing aids, where the results of the 2-AFC tasks indicate a higher SNR for HA1 and a lower SNR for HA4 compared to the phase inversion method.

4 Discussion

4.1 Phase inversion method

4.1.1 Gain reduction for lower input SNRs

The results of the phase inversion method clearly show that all hearing aids reduce the gain for lower input SNRs. This reduction is highest for HA5 and lowest for HA2. Moreover, for HA1, HA3, and HA4, the reduction of gain is almost independent of the speech level whereas HA2, HA5, and HA6 show a higher reduction for higher speech levels. Hence, we can conclude that the noise reduction of all hearing aids considered includes a gain reduction for lower input SNRs, which probably should increase the comfort in noisy environments.

4.1.2 SNR increase

As stated in the draft of the IEC 60118-16 ED1 [4], a NRSE shall provide an increase of at least 1 dB SNR. HA1, HA3, HA4, and HA5 clearly fulfilled this requirement whereas HA6 would have failed with a maximum SNR increase of approx. 0.7 dB. HA2 reached a maximum SNR increase of approx. 1 dB whereas a consideration of measurement uncertainties may also lead to a fail.

From the perspective of the user, not only the amount of SNR increase but also corresponding input SNR is of interest. An SNR increase seems to be more helpful in noisy situations, i.e., for lower input SNRs. Here, there are some differences between the hearing aids. HA5 shows the highest SNR increase for lower input SNRs between −5 and 0 dB, HA1 and HA3 have their maximum SNR increase around 2.5 dB, and HA4 shows the highest SNR increase for higher input SNRs between 5 and 10 dB. Consequently, a different working range of the NRSE of two hearing aids with the same maximum SNR increase can be the reason for different benefits experienced by the user. Another important factor for the benefit experienced by the user is the amount of non-linear distortion caused by the noise reduction which is not considered in this work (see Sect. 4.5.2).

4.2 2-AFC loudness comparison

4.2.1 Effects due to training and fatigue

Training effects and fatigue can change the performance of test subjects over time. One can expect that fatigue

averaged over all subjects changes the performance within one appointment, but not between appointments. By contrast, a training effect is assumed to take place within one appointment and also between appointments, since all three appointments were completed in a short period of time (on average within 3.7 days). Consequently, the comparison of the performances at different appointments should depict training effects (see Fig. 6). Nevertheless, there was no significant effect found for the factor appointment. The mean standard deviation over all appointments and rounds is 0.43 dB so that a difference of 0.28 dB should be noticeable considering a statistical power of 0.8. Thus, we conclude that there are no reasonable training effects.

As depicted in Figure 7, the performance of the test subjects does not significantly change over order/time. The mean standard deviation over all rounds and all four conditions tested within one appointment is 0.49 dB. If again a statistical power of 0.8 is considered, a difference of 0.34 dB should be noticeable so that we conclude that there are also no meaningful effects due to fatigue.

4.2.2 Influence of the speech level on the loudness comparison of the noise level

During the study, test subjects individually evaluated the loudness of speech and noise. A successive evaluation of speech and noise was chosen, because an interleaved task, where the evaluation of speech and noise alternates between the trials, is very difficult and exhausting for the test subjects. Although within one round only speech or noise was compared, speech and noise are always presented simultaneously. Therefore, the evaluation of the signal of interest can be impaired by the loudness of the second signal. To analyze this effect and to compensate for it, an iterative approach where the two signals are evaluated multiple times seems reasonable but also time consuming. We decided to start with the evaluation of the noise signal and included only two iterations for the noise signal due to the following reasons.

The signals to be evaluated consist of speech and noise, which are presented at a sum level of 65 dB SPL. Since a positive SNR was chosen and the activated NRSE should further increase the SNR, the speech level is presented for all signals within a small range of 63.1 dB SPL at 2.5 dB SNR and 65 dB SPL for an infinite positive SNR. On the contrary, if a constant or an increased SNR is assumed, the noise level could reach values of 60.6 dB SPL and below. Hence, we assumed that we have a better starting point for the speech level (63.1 dB SPL) than for the noise level so that we started with the evaluation of the noise level in the 1. round. Then, during the 2. round, the result of the 1. round was used as noise level (see Fig. 4), which should already be a good estimate. Therefore, we decided that for the evaluation of the speech level a second iteration was not required.

In Figure 8, the differences between the evaluation of the noise level during the 1. and 3. round are depicted. The mean standard deviation over all conditions is 1.1 dB
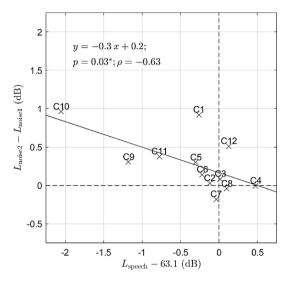

Figure 12. Mean difference between the noise level of the 1. and 3. round for all conditions plotted against the mean change of the speech level during the 1. and 3. round.

so that a difference of 0.79 dB should be noticeable considering a statistical power of 0.8. If a paired t-test is performed and a Bonferroni correction for 12 comparisons is considered, there are no significant differences. However, the average over all conditions showed a significant difference with a mean of 0.28 dB. Since we expect a stronger effect for those conditions where the speech level has changed more between the 1. and 3. round, the mean difference between the noise level of the 1. and 3. round for all conditions against the mean change of the speech level during the 1. and 3. round is plotted in Figure 12. Here, we found a significant negative correlation ($\rho = -0.63$; $p = 0.03^*$), and a linear fit with a slope of -0.3 dB/dB. These results suggest that the speech level during the evaluation of the noise has an impact on the result so that for all further considerations the results of the 3. round are considered only. In addition, the change of the mean noise level between the 1. and 3. round is in the range of 1 dB. Consequently, the noise level during the evaluation of the speech level in the 2. round is already very close to the level of convergence so that only one evaluation of the speech level seems reasonable. Furthermore, we can also conclude that the speech level during the 3. round is near to the level of convergence so that one iteration for the evaluation of the noise signal seems to be sufficient.

4.3 Comparison of the results of the phase inversion method and 2-AFC loudness comparison

4.3.1 Impact of additional noise due to non-linear distortions

Due to a non-linear signal processing within the NRSE, non-linear distortions may be added to the output signals which could be perceived as additional noise. This additional noise would not be canceled out by the phase inversion method. This means that we can consider this effect as additive noise term in all separated signals, i.e., the speech

signal (4), the noise signal (5), and the verification signal (6). Since we found that the minimum of the levels of the separated speech and noise signals was always 10 dB above the level of the verification signal (see Sect. 2.2.1), additive noise caused by non-linear distortions seems to have a minor impact and is neglected.

4.3.2 Noise level

For all devices with NRSE OFF and for those devices with a low reduction of the noise level, the results of the 2-AFC tasks and phase inversion method show no significant differences. By contrast, HA1, HA3, and HA5 show the highest reduction of the noise level and there were significant differences found. In [31], it is shown that noise reduction algorithms may decrease the noise level in speech pauses more than during the speech pulses. Hence, we determined speech pauses of the ISTS and computed the noise level within these intervals (see Sect. 2.1.4). In Figure 1, we can observe that the noise level for HA1 and HA2 with NRSE OFF is almost constant over time. However, there are fluctuations noticeable for HA1 with NRSE ON within a range of ±5 dB. Furthermore, the noise level within the speech pauses is also shown in Figure 9 with circles. Here, we can observe that the noise level within the speech pauses and the noise level of the complete analysis window differs more for conditions where the noise is stronger reduced. Moreover, we see that the differences to the median of the results of the 2-AFC tasks are smaller. Consequently, these results suggest that humans mainly rate the loudness of the noise within the speech pauses and not within the speech pulses. This is similar to the observation that listeners are able to glimpse pieces of the target speech occurring at different times and somehow patch them together to hear out the target speech which is also known as "listening in the gaps" [42, 43]. Compared to the evaluation of the noise signal in our experiment, one main difference is that the noise signal is the target and the speech signal is the time varying interferer.

4.3.3 Speech level

For all devices with NRSE OFF and for HA2 and HA6, the results of the 2-AFC tasks and phase inversion method show no significant differences for the speech level (see Fig. 10). On the other hand, with NRSE ON, the speech levels of HA1, HA3, HA4 and HA5 were rated significantly lower during the 2-AFC tasks than the speech levels determined with the phase inversion method. Using a similar approach as for the noise signal would be computing the speech level during the speech pulses only. However, this approach fails, because (i) the speech signal is rated softer and this approach would increase the sound pressure levels of the speech signal, and (ii) we found that this computation results in a constant offset of about 1.1 dB for all manufacturers with and without activated NRSE so that the individual differences depicted in Figure 10 cannot be explained.

In [31], it is reported that a NRSE may lead to a nonlinear amplification with expansion, i.e., louder speech parts
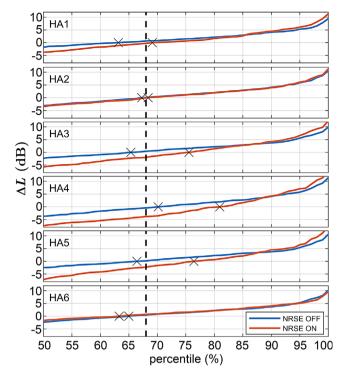

Figure 13. Broadband percentile levels for the speech signals after the separation with the phase inversion method. The percentile levels are depicted relative to the RMS level of the analysis window from 30 s to 60 s. In addition, the zero crossing points ($\Delta L = 0$) are marked with "×". Furthermore, the 68th percentile is marked with a dashed line, since the 68th percentile sound pressure levels best match the results of the 2-AFC tasks (see diamond markers in Fig. 10).

are higher amplified than softer speech parts. Although [44] showed that an expansion for clean speech has a minor effect on the loudness perception, we believe that this could be different, if noise is present. For the analysis of expansion, in [31], the percentile sound pressure levels were computed according to IEC 60118-15 [39]. In this standard, the percentile sound pressure levels are computed separately within 1/3-octave bands whereas we computed the percentile sound pressure levels for the broad band signals only (see Sect. 2.1.5). In Figure 13, the 50th–100th percentile sound pressure levels are depicted for all hearing aids with and without NRSE. Here, we see an expansive amplification for all conditions for which we found differences between the results of the 2-AFC tasks and the phase inversion method. For HA1, HA3, HA4, and HA5, below the 75th percentile the speech parts are lower and above the 90th percentile the speech parts are higher amplified with NRSE ON compared to NRSE OFF. In addition, the zero crossing points ($\Delta L = 0$) are marked with crosses in Figure 13. At these points the percentile sound pressure level is equal to the sound pressure level of the complete signal. For HA1, HA3, HA4, and HA5, this point is shifted to higher percentiles for NRSE ON, i.e., the louder parts of the speech signal contribute more to the sound pressure level of the complete signal. However, since we know that the loudest parts of speech signals are usually very short

and the loudness perception of short sounds is lower [45–47], the loudness perception w.r.t. the sound pressure level of the complete signal is different for HA1, HA3, HA4, and HA5 with NRSE ON due to the expansion of the signal. One approach could be to use a percentile sound pressure level instead of the sound pressure level of the complete signal. We have tested all percentile sound pressure levels and found that in our experiment with the ISTS the 68th percentile best matches the results of the 2-AFC task. Therefore, this percentile is marked in Figure 13 with a dashed line and the 68th percentile sound pressure levels are plotted with diamond markers in Figure 10.

4.3.4 SNR

For HA3 and HA4, the speech and noise levels are both rated softer during the 2-AFC tasks so that the SNR computed with the phase inversion method shows no significant differences. Looking at HA1 and HA4, we can notice highly significant differences whereas the sign is different. Compared to the phase inversion method, the results of the 2-AFC tasks suggest a higher SNR for HA1 and a lower SNR for HA4. For HA1, the noise level of the 2-AFC task is about 2.5 dB lower and the speech level is just about 0.9 dB lower than the result of the phase inversion method. By contrast, for HA4, the noise level of the 2-AFC task shows no significant difference and the median of the speech levels is about 3.1 dB lower than the result of the phase inversion method. If we now use the 68th percentile sound pressure level for the speech signal and consider the noise level in the speech pauses for the separated signals of the phase inversion method, the SNR values better match the results of the 2-AFC tasks (see squares in Fig. 11).

4.4 Generalization of the results

As input signals, the ISTS was presented together with the IFnoise. The ISTS is a standardized signal, which includes speech parts of six languages spoken by 21 different female speakers, and features as many of the most relevant properties of natural speech as possible. Therefore, we believe our results especially of Section 4.3.3 can be seen as general example, whereas the exact percentile sound pressure level which best meets the perceptual results may vary for other speech signals. The IFnoise was generated by overlapping multiple time-shifted instances of the ISTS [36]. It is an almost stationary noise, which has the same long term average spectrum as the ISTS. Hence, we believe our conclusions especially of Section 4.3.2 are also valid for other stationary noise signals, whereas for fluctuating maskers further studies are needed. Furthermore, hearing aids of the six largest hearing aid manufacturers were analyzed so that a variety of different signal processing strategies for a NRSE were covered. Thus, we believe our conclusions on the performance, characteristics, and analysis of a NRSE is valid for state-of-art realizations of a NRSE, whereas future developments may require further investigations.

4.5 Limitations

4.5.1 Interpretation of SNR increase

The phase inversion method and the 2-AFC loudness comparisons can be used to analyze different features of hearing aids in terms of SNR, e.g., the NRSE or directional microphones. A SNR increase caused by one of these features may also reduce listening effort or increases comfort whereas in most cases speech intelligibility is only increased by directional microphones. Thus, the SNR increase measured should not be translated into speech intelligibility. For this purpose, a direct measure of speech intelligibility or other means should be considered.

4.5.2 Sound quality

From the perspective of the user, not only the increase of SNR but also the sound quality is of importance. Usually, the strength of noise reduction algorithms can be adapted, and there is a trade-off between SNR increase and nonlinear distortions. Therefore, a complete evaluation of noise reduction algorithms in hearing aids requires also a measure of sound quality, which was not the focus of this study.

4.5.3 Hearing aid settings

Since the focus of this study is on the NRSE and other features of the signal processing may affect the measurements with the phase inversion method, all adaptive features of the hearing aids beside the NRSE were deactivated during the experiments. This is different to the configuration in real life where multiple features are usually active at the same time. Hence, the results of this study do not cover interactions between other features and the NRSE. Nevertheless, as long as features do not behave differently because of the phase inversion and do not change the phase in a non-linear way, the phase inversion method should also be applicable in combination with other features such as the dynamic range compression. Consequently, another application of the 2-AFC loudness comparisons described in this work could be the investigation of the reliability of the phase inversion method in combination with other hearing aid features.

4.5.4 Test subjects

In this study, 18 normal hearing test subjects evaluated the effect of the NRSE within multiple 2-AFC loudness comparisons. It is well known that the loudness perception of hearing impaired people is different to normal hearing listeners. Nevertheless, in [34], hearing impairment was not correlated to the just noticeable difference (JND) in SNR. Moreover, since the relative loudness and not the absolute loudness is evaluated during the 2-AFC tasks, it is assumed that the influence of hearing impairment is reduced. However, further experiments with hearing impaired listeners are required to analyze the influence of hearing ability on the results of the 2-AFC loudness comparisons.

5 Conclusion

The results of the phase inversion method for multiple speech levels and input SNRs reveal differences of the noise reduction for speech enhancement (NRSE) between hearing aids of six different manufacturers. Differences have been found for the maximum SNR increase and for the working range of the noise reduction. This again demonstrates the possibilities of the phase inversion method for the analysis of noise reduction algorithms in hearing aids.

Furthermore, a 2-AFC procedure for a separate evaluation of the loudness of speech and noise within a mixed signal was presented. The individual ratings for the loudness of speech and noise were determined by an iterative approach. A study with 18 adult normal-hearing listeners demonstrated that the 2-AFC task is robust against training effects and fatigue. Moreover, we found that two iterations for the noise and one iteration for the speech signal were sufficient for this study.

Both the technical evaluation with the phase inversion method and the 2-AFC loudness comparison were performed for the same hearing aids of six different manufacturers with NRSE OFF and ON. A comparison of both results shows significant differences, and an evaluation of speech, noise and SNR gives insights into possible reasons for those differences. For the speech signal, a NRSE may lead to an expansive amplification, which changes the loudness perception w.r.t. to the sound pressure level of the speech signal. For the noise signals, we found that a NRSE may reduce the noise stronger in speech pauses than during speech activity, and the perceived loudness of the noise better matches the noise level within the speech pauses. With those findings we developed a way to post process the results of the phase inversion method to better match the perceived loudness of human test subjects. We found that the 68th percentile broadband sound pressure level of the separated speech signal, and the sound pressure level of the noise signal only evaluated within the the speech pauses well match the individual loudness ratings for speech and noise. While the results of the present investigation hold for a stationary noise signal with the same spectrum as the speech, the applicability to other noise signals has to be evaluated in further studies. Overall, we showed that the phase inversion method basically can predict the perceived loudness of normal-hearing listeners, but a simple computation of the sound pressure level may lead to significant deviations. The authors declare that there is no conflict of interest.

Conflict of interest

Author declared no conflict of interests.

References

1. A. Schaub: Digital Hearing Aids. Thieme, 2008.
2. H. Dillon: Hearing Aids. Thieme Publishers Series, Boomerang Press, 2012.
3. M. Metz: Sandlin's Textbook of Hearing Aid Amplification: Technical and Clinical Considerations. Plural Publishing, 2014.
4. Draft of IEC 60118-16 ED1: Electroacoustics – Hearing aids – Part 16: Definition and verification of hearing aid features. Jan 2020.
5. I. Brons: Perceptual evaluation of noise reduction in hearing aids. Dissertation at the University of Amsterdam – Faculty of Medicine, 2013.
6. R.C. Hendriks, T. Gerkmann, J. Jensen: DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art. Synthesis Lectures on Speech and Audio Processing 9 (Jan 2013) 1–80.
7. F.Y. Chong, L.M. Jenstad: A critical review of hearing-aid single-microphone noise-reduction studies in adults and children. Disability and Rehabilitation: Assistive Technology 13 (Oct 2017) 600–608.
8. J.I. Alcántara, B.C. Moore, V. Kühnel, S. Launer: Evaluation of the noise reduction system in a commercial digital hearing aid. International Journal of Audiology 42 (Jan 2003) 34–42.
9. T.A. Ricketts, B.W.Y. Hornsby: Sound quality measures for speech in noise through a commercial hearing aid implementing digital noise reduction. Journal of the American Academy of Audiology 16, 5 (2005) 270–277.
10. S. Nordrum, S. Erler, D. Garstecki, S. Dhar: Comparison of performance on the hearing in noise test using directional microphones and digital noise reduction algorithms. American Journal of Audiology 15 (Jun 2006) 81–91.
11. R. Bentler, L.-K. Chiou: Digital noise reduction: An overview. Trends in Amplification 10 (Jun 2006) 67–82.
12. H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W.A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, A. Spriet: Multicenter evaluation of signal enhancement algorithms for hearing aids. The Journal of the Acoustical Society of America 127 (Mar 2010) 1491–1505.
13. D. Wendt, R.K. Hietkamp, T. Lunner: Impact of noise and noise reduction on processing effort. Ear and Hearing 38, 6 (2017) 690–700.
14. B. Ohlenforst, D. Wendt, S.E. Kramer, G. Naylor, A.A. Zekveld, T. Lunner: Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. Hearing Research 365 (Aug 2018) 90–99.
15. C. Bernarding, D.J. Strauss, R. Hannemann, H. Seidler, F.I. Corona-Strauss: Neurodynamic evaluation of hearing aid features using EEG correlates of listening effort. Cognitive Neurodynamics 11 (Feb 2017) 203–215.
16. K. Miles, C. McMahon, I. Boisvert, R. Ibrahim, P. de Lissa, P. Graham, B. Lyxell: Objective assessment of listening effort: Coregistration of pupillometry and EEG. Trends in Hearing 21 (Jul 2017). https://doi.org/10.1177/2331216517706396.
17. I. Holube, K. Haeder, C. Imbery, R. Weber: Subjective listening effort and electrodermal activity in listening situations with reverberation and noise. Trends in Hearing 20 (Sep 2016). https://doi.org/10.1177/2331216516667734.
18. S. Gatehouse, J. Gordon: Response times to speech stimuli as measures of benefit from amplification. British Journal of Audiology 24, 1 (1990) 63–68. PMID: 2317602.
19. R. Houben, M. van Doorn-Bierman, W.A. Dreschler: Using response time to speech as a measure for listening effort. International Journal of Audiology 52 (Sep 2013) 753–761.
20. A. Sarampalis, S. Kalluri, B. Edwards, E. Hafter: Objective measures of listening effort: Effects of background noise and noise reduction. Journal of Speech, Language, and Hearing Research 52 (Oct 2009) 1230–1240.

21. J.L. Desjardins, K.A. Doherty: The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. Ear and Hearing 35, 6 (2014) 600–610.

22. E.M. Picou, T.A. Ricketts: The effect of changing the secondary task in dual-task paradigms for measuring listening effort. Ear and Hearing 35, 6 (2014) 611–622.

23. J.-P. Gagné, J. Besser, U. Lemke: Behavioral assessment of listening effort using a dual-task paradigm. Trends in Hearing 21 (Jan 2017). https://doi.org/10.1177/2331216516687287.

24. C. Kwak, W. Han: Comparison of single-task versus dual-task for listening effort. Journal of Audiology and Otology 22 (Apr 2018) 69–74.

25. S. Fredelake, I. Holube, A. Schlueter, M. Hansen: Measurement and prediction of the acceptable noise level for single-microphone noise reduction algorithms. International Journal of Audiology 51, 4 (2012) 299–308.

26. Y.-H. Wu, E. Stangl: The effect of hearing aid signal-processing schemes on acceptable noise levels. Ear and Hearing 34, 3 (2013) 333–341.

27. M. Schulte, M. Krüger, M. Meis, K.C. Wagener: Subjective listening effort. The Journal of the Acoustical Society of America 137 (Apr 2015) 2236–2236.

28. M. Krueger, M. Schulte, T. Brand, I. Holube: Development of an adaptive scaling method for subjective listening effort. The Journal of the Acoustical Society of America 141 (Jun 2017) 4680–4693.

29. B. Hagerman, A. Olofsson: A method to measure the effect of noise reduction algorithms using simultaneous speech and noise. Acta Acustica United with Acustica 90, 2 (2004) 356–361.

30. I. Holube, S. Fredelake, M. Hansen: Subjective and objective evaluation methods of complex hearing aids, in 8th International Congress on Audiology, Heidelberg. Jun 2007.

31. H. Husstedt, M. Kuttner: Quantitative measurement of signal-to-noise ratio with the percentile analysis according to IEC 60118-15, in 44. Jahrestagung für Akustik (DAGA 2018), Munich. Mar 2018.

32. N. Kienitz, M. Frenz, H. Husstedt: Evaluation der Störgeräuschunterdrückung von Hörsystemen durch eine Signalseparation im Frequenzbereich [Evaluation of the noise reduction of hearing aids by a separation of signals in the frequency domain], in 22. Jahrestagung der Deutsche Gesellschaft für Audiologie (DGA). Mar 2019.

33. H. Husstedt, A. Mertins, M. Frenz: Evaluation of noise reduction algorithms in hearing aids for multiple signals from equal or different directions. Trends in Hearing Jan 22 (2018). https://doi.org/10.1177/2331216518803198.

34. D. McShefferty, W.M. Whitmer, M.A. Akeroyd: The just-noticeable difference in speech-to-noise ratio. Trends in Hearing 19 (Feb 2015) 233121651557231.

35. I. Holube, S. Fredelake, M. Vlaming, B. Kollmeier: Development and analysis of an international speech test signal (ISTS). International Journal of Audiology 49, 12 (2010) 891–903.

36. EHIMA: Description and terms of use of the IFFM and IFnoise signals, in European Hearing Instrument Manufacturers Association. 2016.

37. IEC 60318-5: Electroacoustics – Simulators of human head and ear – Part 5: 2 cm³ coupler for the measurement of hearing aids and earphones coupled to the ear by means of ear inserts. Aug 2006.

38. IEC 60118-0: Electroacoustics – Hearing aids – Part 0: Measurement of the performance characteristics of hearing aids. Jun 2015.

39. IEC 60118-15: Electroacoustics – Hearing aids – Part 15: Methods for characterising signal processing in hearing aids with a speech-like signal. Feb 2012.

40. ISO 8253-2: Acoustics – Audiometric test methods – Part 2: Sound field audiometry with pure-tone and narrow-band test signals. Dec 2009.

41. ISO 8253-3: Acoustics – Audiometric test methods – Part 3: Speech audiometry. Mar 2012.

42. G.A. Miller, J.C.R. Licklider: The intelligibility of interrupted speech. The Journal of the Acoustical Society of America 22 (Mar 1950) 167–173.

43. N. Li, P.C. Loizou: Factors influencing glimpsing of speech in noise. The Journal of the Acoustical Society of America 122 (Aug 2007) 1165–1172.

44. J. Rennies, I. Holube, J.L. Verhey: Loudness of speech and speech-like signals. Acta Acustica United with Acustica 99 (Mar 2013) 268–282.

45. B.R. Glasberg, B.C.J. Moore: A model of loudness applicable to time-varying sounds. The Journal of the Audio Engineering Society 50, 5 (2002) 331–342.

46. J. Chalupper, H. Fastl: Dynamic loudness model (DLM) for normal and hearing-impaired listeners. Acta Acustica United with Acustica 88 (May 2002) 378–386.

47. J. Rennies, J.L. Verhey, H. Fastl: Comparison of loudness models for time-varying sounds. Acta Acustica United with Acustica 96 (Mar 2010) 383–396.