OBJECTIVE MEASURES TO QUANTIFY THE PERCEPTUAL EFFECTS OF NOISE REDUCTION IN HEARING AIDS

KAROLINA SMEDS[1], FLORIAN WOLTERS[1,2], ANDERS NILSSON[1,3], SARA BÅSJÖ[1], SOFIA HERTZMAN[1], AND ARNE LEIJON[3]

[1] *Widex A/S ORCA Europe, Stockholm, Sweden*
karolina.smeds@orca-eu.info
[2] *University of Applied Sciences, Oldenburg, Germany*
[3] *KTH, Stockholm, Sweden*
leijon@kth.se

Twenty listeners with hearing impairment evaluated three noise-reduction algorithms using paired comparisons of speech clarity, noise loudness, and preference. The subjective test produces results in terms of physical signal-to-noise ratios that correspond to equal subjective performance with and without the noise-reduction algorithms. This facilitates a direct test of how well a number of objective performance measures correspond with the subjective test results.

INTRODUCTION

Noise reduction (NR) is commonly used in modern hearing aids. Various rationales can be used when designing such NR algorithms. Most hearing impairments reduce the ability to understand speech in background noise. Therefore, an appropriate NR design goal could be to increase speech intelligibility in noise. Another design goal could be to increase listening comfort or ease of listening, an important aspect since hearing aids usually are used all day. Previous measurements [1] have shown that hearing aid NR algorithms function in very different ways. The effects of these NR algorithms are usually evaluated in listening tests with participants with or without hearing impairment. It would, however, be of great value if some objective measures could be used to indicate the effect of various NR algorithms prior to laboratory or field testing with listeners.
The aim of the current study was to explore a number of physical, objective measures to see to what extent they have the potential to quantify the effect of noise reduction for hearing-impaired listeners. In the study, both speech intelligibility and sound quality was evaluated. The focus in this paper will be on the sound quality results.

1 METHOD

Twenty listeners with hearing impairment participated in a laboratory study, where they listened to noisy sound files that were pre-processed using three software-based noise reduction algorithms. Two of the NR algorithms were general speech enhancement algorithms, whereas the third NR algorithm was fine-tuned for hearing aid use. All tests were conducted binaurally in a sound-proof booth with linearly and individually fitted hearing aids to compensate for each participant's hearing loss. Paired Comparison Ratings (PCR) of Preference, Speech Clarity, and Noise Loudness were performed.

1.1 Participants

Twenty listeners, eleven women and nine men, with symmetrical, sensorineural, mild-to-moderate hearing loss were recruited from a research database at ORCA Europe. Their ages ranged from 62 to 82 years with a mean age of 71.5 years. In Fig. 1, audiogram information is presented. The participants were not paid for their participation, but they received a symbolic gift (value: € 10) at the last visit.

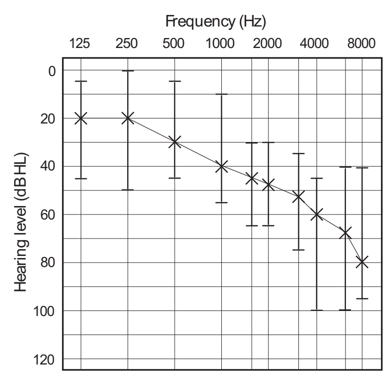

Figure 1: Median thresholds and the range of hearing losses for the forty ears in the study.

1.2 Hearing aids

Binaural high-quality hearing aids (Inteo 9, Widex A/S) without brand labels were linearly programmed according to the NAL-R prescription [2] reduced by 6 dB across the frequency range. The reduction was motivated by the binaural fittings and a presentation level that was higher than the normal speech assumed in the prescription. All advanced signal processing, such as feedback suppression and directional microphones, was switched off. The hearing aids were used with tight earmoulds. The hearing aid fittings were verified using real ear insertion gain measurements, and the linearity was confirmed with coupler gain measurements in a test chamber using a large range of input levels.

1.3 Sound Files

A listening situation where speech masks speech at a realistic SNR of +4 dB was used for the laboratory test at an overall presentation level of 72 dB SPL. An unintelligible artificial babble noise was derived using the International Speech Test Signal (ISTS) [3]. The ISTS signal consists of six different female speakers who read out the same story in their mother tongue. Randomly selected segments of these six recordings are attached to each other to form an audio file, where the language is changed from segment to segment. To create a babble noise, the ISTS signal was superimposed eight times with randomly varying starting points and with varying levels (pair-wise decreased by 2 dB from the first pair). The babble noise was then filtered to the long-term average spectrum of the speech that was used for the measurements.

The speech material was derived from recordings of a female talker telling stories. A set of running speech samples were selected based on overall levels and level fluctuations to create six 40-s long samples with similar properties. The overall RMS level was adjusted to be the same for all samples. These speech files were mixed with the artificial babble noise at two SNRs, +4 and +9 dB.

1.4 Noise Reduction Algorithms

Three software-implemented noise reduction (NR) algorithms were used in the study. These were selected, based on informal listening tests, to create sound files that really differed after NR processing. Two of the algorithms (WEDM and Wiener) were based on general concepts of speech enhancement, while a third algorithm (PSSLP) was optimized for use in hearing aids. Since evaluation of the NR algorithms was not the focus in the study, the algorithms are described very briefly below.

The WEDM (a Bayesian noise estimator based on the weighted Euclidean distortion measure) and the Wiener (Wiener filtering based on a priori SNR estimation) are described in a textbook by Loizou [4] and the Matlab codes provided in the textbook were used in the experiment. These general speech enhancements algorithms produced audible distortion that could be described as “musical noise”, something that is often associated with NR processing. To the normal-hearing listeners who participated in the informal listening test prior to the experiment, the WEDM seemed to distort the speech signal more than the Wiener did.

In contrast to these general speech enhancement algorithms, the third NR algorithm, the PSSLP (the Perceptually tuned Spectral Subtraction algorithm with Low-Pass filtered spectral filter coefficients, [5]), was fine-tuned for hearing aid use. This perceptually motivated optimization of the NR parameters is meant to achieve a well balanced trade-off between speech distortion and noise reduction with a low amount of musical noise artefacts. To the normal-hearing listeners who participated in the informal listening, the PSSLP provided good sound quality without audible musical noise for the SNRs used.

By comparing the SNR for the unprocessed and the processed noise files, SNR improvements as a result of the various NRs were calculated. The SNR after NR processing was calculated according to a method by Hagerman and Olofsson [6], where speech and noise after the processing can be calculated separately by using two input signals with speech and noise, where one noise signal has a 180° phase shift compared to the other. By adding and subtracting these signals, the levels of the speech and the noise after processing can be calculated separately.

The long-term mean SNR value after the processing of all running speech recordings was calculated to get one single value for the SNR improvement achieved by each NR algorithm. Note that these values are not subject specific, but only indicate the physical effect of the various NR algorithms. The results show the highest SNR improvement for the WEDM algorithm with 4.6 dB, followed by the Wiener algorithm with a value of 2.7 dB, and the PSSLP, which improves the SNR on average by 1.3 dB. These calculations showed that the NR algorithms all produced a physical improvement in SNR.

1.5 Instrumentation

The listening tests were performed in a sound-proof booth (3.2×3.05 m) under sound field conditions using one loudspeaker (Jamo D400) placed one meter in front of the listener. The frequency response from the loudspeaker to the listening position was flat within ±3 dB except for two peaks (exceeding the 3-dB limit with less than 1 dB) around 0.2 and1 kHz. However, the measured frequency response was included in all theoretical calculations. The program material was stored in a computer and played back with an external 24-bit RME Fireface 880 sound card. Steinberg Cubase SE3 software was used to operate the paired comparison test through a remote control (Frontier Tranzport). The processed and the unprocessed sounds were running in parallel in two audio channels. The remote control was modified so that only the desired functions were operable and visible to the participants, and the participants could change as many times as they wanted between the two sound files under comparison.

1.6 PCR

Sound quality was measured in three dimensions: Preference, Speech Clarity, and Noise Loudness. Paired comparisons were combined with ratings of the perceived differences between the processing schemes in a procedure called Paired Comparison Rating (PCR), developed for the evaluation of NR algorithms [7]. One advantage with the PCR method is that it produces results in terms of physical SNRs that correspond to equal subjective performance with and without the NR activated. The obtained SNR values can then be expressed as an SNR change achieved by the NR algorithm.

The basic idea of the PCR is that the unprocessed and the processed sound files are compared at the same SNR, in this study at +4 dB. In addition to this comparison, the processed sound files are compared to a version of the unprocessed files with an increased SNR, in our case at +9 dB. The listener selects one condition in the paired comparison, but the listener also rates how different the two conditions are on a visual analogue scale (Fig. 2).

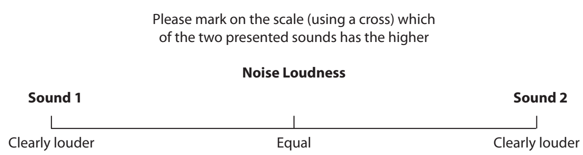

Figure 2: The visual analogue scale used in the PCR, here illustrated by the scale for the Noise Loudness.

After the ratings are completed, the marks on the scale are measured and the scale is manually normalized to a ±10 unit scale. With this method, the point of subjective equality is determined, i.e., the noise reduction has the same effect as increasing or decreasing the SNR of the unprocessed material by the amount of SNR change. The procedure is described in Fig. 3.

The listeners participated in two PCR measurement sessions, the second approximately one week after the first. At each session, each paired comparison was repeated twice; once the processed file was presented as Sound 1 and the unprocessed file was presented as Sound 2, and once the order was reversed. When analyzing the results, any SNR change exceeding ±7 dB was set to these limits. The three rating criteria, i.e., Preference, Speech Clarity and Noise Loudness, were rated one at the time. One third of the listeners started with each of the rating criteria. Within each rating criterion, the presentation order (NR algorithm and SNR for the unprocessed signal) was randomly selected.

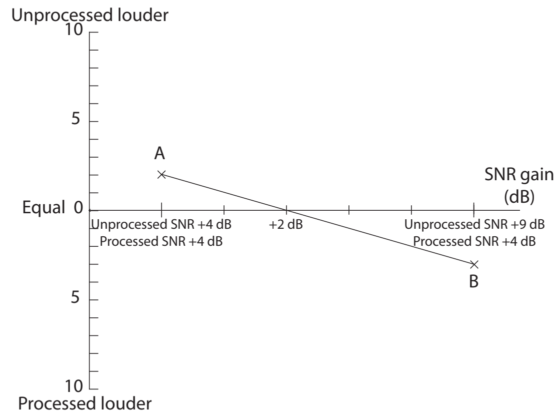

Figure 3: An example of the general PCR procedure for the Noise Loudness criterion. When the processed and the unprocessed sound files are compared at the same SNR, the hypothetical listener in this example rated that the unprocessed sound had the highest loudness, a rating of 2 units (A). When the unprocessed sound had an SNR of +9 dB, the person rated the processed sound to have the highest loudness, a rating of 3 units (B), i.e., the NR helped decrease the noise loudness. A line is drawn between the two rating points, and the point of subjective equality is determined, in this example corresponding to 2 dB lower noise loudness with the NR than without.

1.7 Calculated Physical SNR Change

The physical SNR improvement calculations used in the initial evaluation of the NR algorithms (described above) were used as the most basic objective measure.

1.8 Calculated Noise Loudness

The partial loudness of speech and noise was calculated for each participant. The calculations were made using the sound files produced by each NR algorithm at the +4 dB SNR used in the PCR measurements. Since the PCR measurements resulted in an SNR change in dB, the calculations were also made for the unprocessed sound files at the individual SNRs judged as giving equal noise loudness in the PCR test. The ratio between these two measures should be 1 if the Calculated Noise Loudness is an appropriate predictor of PCR Noise Loudness. The calculation required input speech and noise signals to be stored in separate channels with relative amplitudes defined by the mixed files' SNRs. For the WEDM and the Wiener algorithms, the speech and noise components were separated using the method of Hagerman and Olofsson [6]. This was not needed for the PSSLP algorithm, because it was implemented for testing purposes to process speech and noise components separately, just as if they were mixed together (shadows filtering).

The loudness calculation method was similar to that of Moore et al. [8], Glasberg and Moore [9], and Moore and Glasberg [10], including the following auditory transformations: (1) a fixed linear filter for the transmission from the sound field to the eardrum, (2) a fixed linear filter representing middle-ear transmission, and (3) linear and non-linear filtering at the outer hair cell level to mimic auditory frequency resolution. The implementation deviated slightly from Moore et al. (1997) in the following details: The auditory filter shape was modelled by one “tail” and one “peak” filter in parallel, as suggested by Baker and Rosen [11]. The tail filter had a fixed shallow slope. The peak filter had a fixed symmetric shape with a normal auditory equivalent rectangular bandwidth (ERB), and a variable maximum gain controlled by the output of the peak filter, in order to model the outer-hair-cell compression. The absolute pure-tone thresholds were defined at a fixed detectability index (d'=1) instead of a fixed loudness value. For calculation of the partial loudness of noise in the presence of speech, and vice versa, the signal components were treated symmetrically. It was verified that the implemented calculation method agrees with empirical loudness-balance data, at least as well as the version implemented and validated by Moore et al. [8].

The loudness calculations assumed that individual hearing threshold losses were caused mainly by a loss of the natural outer hair cell gain, up to a hearing loss of 50 dB HL. The amount of hearing threshold loss exceeding the maximum outer hair cell gain was represented by a level-independent attenuation at the inner hair cells. Loudness was calculated separately for the left and right ears, and the binaural loudness was taken as the sum of the monaural loudness values. Only the partial loudness of the noise is reported here.

1.9 PESQ

The Perceptual Evaluation of Speech Quality (PESQ) is registered as recommendation P.862 of the International Telecommunication Union (ITU) for objective evaluations of end-to-end speech quality of narrow-band telephone networks and speech codecs [12]. The output from the method is a score which is supposed to be a proper predictor of the quality assessment made by listeners, and is based on a comparison of an original/clean input signal and the corresponding output signal which has been fed through the system under test. The PESQ employs a perceptual model to account for psychophysical aspects of the human auditory system. The PESQ software version available in the textbook by Loizou [4] was used.

For our particular application, there were three main difficulties with the PESQ: 1. It has not been validated for artefacts from NR algorithms. 2. The reference condition, the unprocessed file, contained babble noise and was not a clean speech signal. 3. The PESQ does not take hearing loss into account. Despite these drawbacks, the measure was used. In order to, at least partly, compensate for the hearing losses of the listeners, an individual threshold-shaped noise and the individually measured real ear insertion gain values were added to the sound files.

1.10 Hu and Loizou’s Composite Measure

In a Composite Measure developed by Hu and Loizou [described in 4], the authors examined different combinations of existing objective measures to form three basic predictors for overall quality (c_{ovl}), signal distortion (c_{sig}) and background noise distortion (c_{bak}). To find the optimal weighting factors for the different measures within one predictor, linear and nonlinear regression analysis procedures were used. The measures that were used to form the three composite measures include a segmental SNR measure, a frequency-based segmental SNR, the PESQ, a weighted spectral slope measure, the log-likelihood ratio, and a cepstrum distance measure in various combinations.

The Compisite Measure software version available in the textbook by Loizou [4] was used. The problems described with the PESQ for the type of testing performed in the current study also apply to the Composite Measure.

1.11 Signal to Distortion Ratio

The Signal to Distortion Ratio (SDR) analyzes the amount of distortion caused by nonlinear processing by comparing two output signals derived from two input signals where one is the Hilbert transform of the other. After nonlinear processing, the signals will not be a perfect Hilbert pair, and the amount of mismatch indicates the degree of nonlinearity [13]. This mismatch is quantified as the A-weighted signal-to-distortion ratio in dB; the higher the SDR, the lower the distortion.

This method has shown a high correlation between subjective and objective data on distortion produced by the nonlinear behaviour of compression systems [13]. However, the authors state that the requirement that the two output signals of a system form a Hilbert pair (when the respective input signals form a Hilbert pair) is a necessary but probably not sufficient condition for the system to be free from perceptible distortion.

SDR calculations were performed for each running speech sample processed by each of the included NR algorithms as well as for the unprocessed files. An individual hearing-threshold-shaped noise and individual real ear insertion gain values were added to the files prior to the analysis to account for the hearing losses of the listeners.

2 RESULTS

2.1 PCR

For the Paired Comparison Rating data (PCR), the median SNR change (in dB) over the four replications were calculated for each participant, and the data for all participants are presented in Fig. 4. Unfortunately, an error occurred in the processing of one of the six running speech samples. When this was discovered, the faulty speech sample was replaced. However, this was not done until after some of the tests had already been completed, and therefore some of the data are missing for around half of the participants.

Possible deviations from 0 dB were analysed statistically using the Wilcoxon signed-rank test (tested on a 5% significance level with a Bonferroni correction, $p<0.017$, to compensate for the multiple comparisons). It was found that both the WEDM and the Wiener algorithms managed to subjectively reduce the Noise Loudness (Fig. 4, top panel). For the WEDM algorithm this was achieved at the expense of decreases in both Speech Clarity (Fig. 4, middle panel) and Preference (Fig. 4, bottom panel).

2.2 Calculated Physical SNR Change

The basic physical SNR change calculations that were made for each NR algorithm showed that the PSSLP improved the SNR with 1.3 dB, the WEDM with 4.6 dB, and the Wiener with 2.7 dB (see above). For the PSSLP and the WEDM these values correspond to the upper quartile values for the PCR Noise Loudness data, whereas for the Wiener the objective value is close to the median for the subjective PCR measure (Fig. 4, top panel).

2.3 Calculated Noise Loudness

For the individual Calculated Noise Loudness, each participant's PCR Noise Loudness result was entered into the calculations. The ratio between the noise loudness for the processed and the unprocessed conditions was calculated individually. If the Calculated Noise Loudness is a good predictor of the PCR Noise Loudness, these ratios should be 1. The results are presented in Fig. 5. A two-sided Wilcoxon signed-rank test showed that the result did not deviate from 1 for any of the three NR algorithms (tested on a 5% significance level with a Bonferroni correction, $p<0.017$, to compensate for the multiple comparisons).

2.4 PESQ, Composite, and SDR Measures

The calculated results for the PESQ, Composite Measure, and the SDR are presented in Table 1. It can be seen that both the PESQ and the Composite Measure led to low values for all NR algorithms, indicating a poor sound quality.

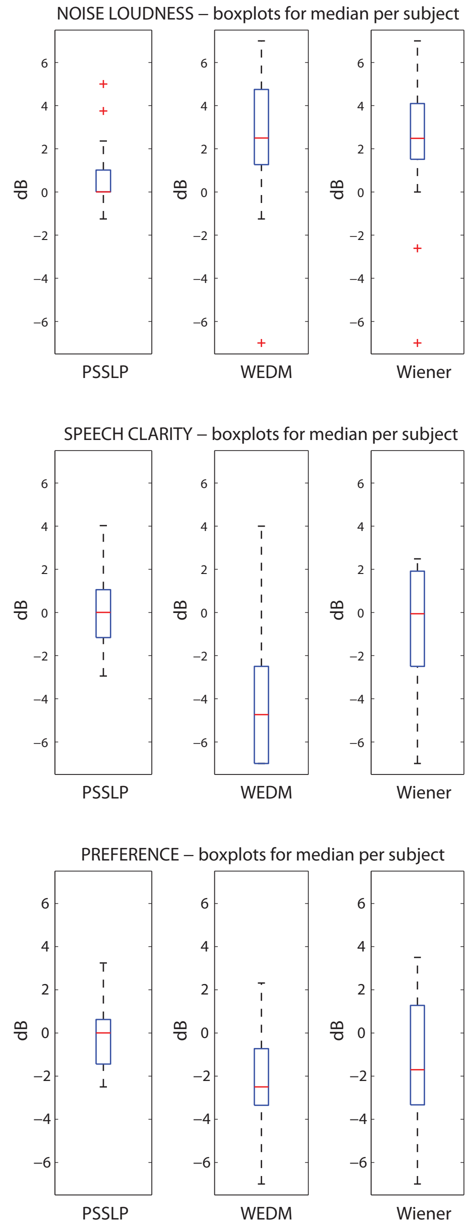

Figure 4: The results of the PCR measurements. On each vertical axis the SNR change (in dB) is shown. The top panel shows the results for Noise Loudness (where positive values indicate that the Noise Loudness was decreased by the NR), the middle panel the results for Speech Clarity (where negative values indicate that the Speech Clarity was decreased by the NR), and the bottom panel the results for Preference (where negative values indicate that the NR decreased the Preference). Inter-quartile values are shown by the box, the median is the line within the box, the outliers (marked by +) are defined as values outside 1.5 times the box, and the whiskers extend to the highest and lowest values when the outliers are excluded.

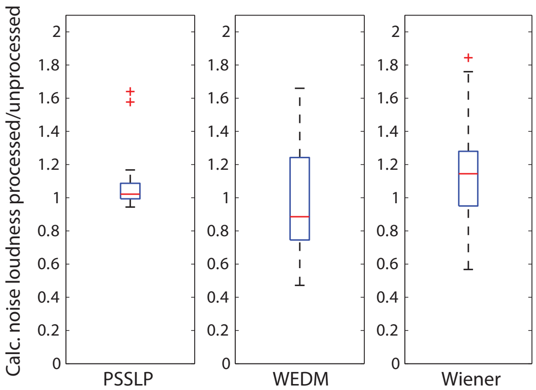

Figure 5: Calculated Noise Loudness results for the three NR algorithms. Inter-quartile values are shown by the box, the median is the line within the box, the outliers (marked by +) are defined as values outside 1.5 times the box, and the whiskers extend to the highest and lowest values when the outliers are excluded.

	PSSLP	WEDM	Wiener
PESQ	1.6	1.3	1.3
c_{sig}	2.8	1.9	2.1
c_{bak}	2.1	1.8	1.8
c_{ovl}	2.1	1.5	1.6
SDR (dB(A))	60.9	57.2	59.2

Table 1: The mean across-subject results for each physical measure (PESQ, the three dimensions in the Composite Measure, and the SDR) and each NR algorithm.

The WEDM and the Wiener algorithms are based on different noise reduction strategies, but they led to similar sound quality. Both algorithms substantially increased the physical SNR, a result that corresponds well with the PCR Noise Loudness (Fig. 4 top panel) and the Calculated Noise Loudness results (Fig. 5). However, this SNR improvement was achieved at the expense of distortion characterized as “musical noise”, which affected the results for the PCR Speech Clarity (Fig. 4, middle panel) and Preference (Fig. 4, bottom panel) negatively. Also the results from the PESQ and the Composite Method were negatively affected (Table 1), but the reduced sound quality was not captured by the SDR measure (Table 1), which produced high scores for these methods. The SDR does not seem to be sensitive to this type of distortion.

The PSSLP algorithm, on the other hand, produced the best sounding speech and introduced the lowest amount of musical noise. It did not increase the physical SNR much, a result that corresponds well with the PCR Noise Loudness (Fig. 4 top panel) and the Calculated Noise Loudness results (Fig. 5), but it gave higher values on the PESQ and the Composite Method than the WEDM and Wiener (Table 1). It lead to similar SDR results as the WEDM and the Wiener (Table 1).

In Fig. 6, the individual PCR Speech Clarity results for the three NR algorithms are plotted against the individual results of the Composite dimension Signal Distortion, c_{sig}. It can be seen that the c_{sig} has a limited variability across participants for each NR algorithm, and for this particular comparison, the three NR algorithms form three fairly distinct clusters. This was not the case for most other comparisons.

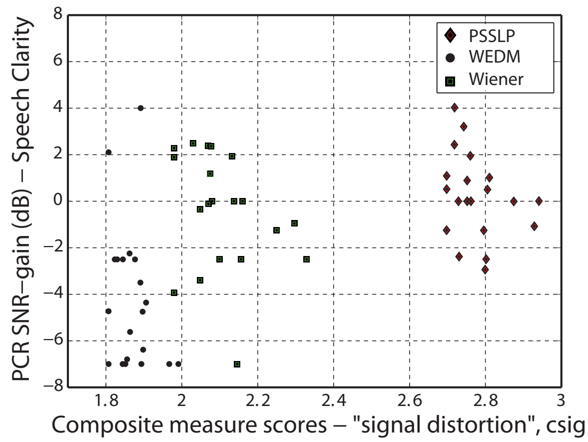

Figure 6: Individual PCR Speech Clarity results vs. the results of the Composite c_{sig} calculations for the three NR algorithms used.

3 DISCUSSION

Twenty listeners with hearing impairment participated in a laboratory study where they performed paired comparison ratings (PCRs) of Noise Loudness, Speech Clarity, and Preference. The results were then compared to the results of a number of physical measures.

The PCR procedure worked fairly well. An error in the NR processing of one of the running speech samples, made the data set incomplete. The results are presented as medians over four replications, but for half of the participants, these medians were calculated using fewer data points.

Compared to the results of Dahlquist et al. [7], the current study showed a larger number of strange or unexpected responses. It is difficult to determine why this happened, but the actual NR algorithms used might have caused the problem. The musical noise introduced by the WEDM and the Wiener might have resulted in difficulties judging some of the PCR criteria. The very small audible changes that were introduced by the PSSLP might have made it difficult to judge that method in the PCR as well.

Both the PESQ and the Composite Method were used outside their intended scope (see above). The absolute values for these measures are difficult to interpret, but the measures did a good job of ranking the three NR algorithms, where the PSSLP achieved higher scores than the other algorithms on these calculated measures and also higher scores on PCR Speech Clarity and Preference.

A general problem with the physical methods used in this study is that only the Calculated Noise Loudness method takes individual hearing thresholds into account. For the PESQ, the Composite Method, and the SDR, a rough compensation for the individual hearing thresholds was included by introducing the hearing-threshold shaped noise and the individual real ear insertion gains.

4 CONCLUSION

The individually Calculated Noise Loudness correlated well with the PCR Noise Loudness data. On a group level, the very basic measure of physical SNR improvement achieved by the NR algorithms also worked better than expected.

For the PCR Speech Clarity and Preference measures, the PESQ and the Composite Method produced values that gave the correct ranking order between NR algorithms, but the absolute values were difficult to interpret. The SDR measure did not correlate with the perceived sound quality.

The Calculated Noise Loudness is the only of the physical measures used in this study that includes the individual hearing thresholds. The study emphasizes the importance of theoretical measures that are developed to predict the results of subjective tests with listeners with hearing loss. For the evaluation of noise reduction algorithms this is crucial. It is difficult to find another application where noise reduction algorithms are used a large number of hours every day.

ACKNOWLEDGMENTS

The authors thank Martin Dahlquist, Åke Olofsson, Josefina Larsson, and Martin Hansen for valuable discussions, and the participants in the experiment for their work.

REFERENCES

[1] K. Smeds, N. Bergman, S. Hertzman, and T. Nyman. "Noise reduction in modern hearing aids – Long-term average gain measurements using speech." in *International Symposium on Auditory and Audiological Research (ISAAR).* Helsingør, Denmark (2009).

[2] D. Byrne and H. Dillon, "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid". *Ear Hear*. vol.4, pp. 257-65 (1986).

[3] I. Holube, "Short description of the International Speech Test Signal (ISTS)". Center of Competence HörTech and Institute of Hearing Technology and Audiology:Oldenburg, Germany (2007).

[4] P.C. Loizou, *Speech enhancement theory and practice.* Boca Raton. FL, USA: Taylor & Francis Group (2007).

[5] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Büchler, N. Dillier, R. Houben, W. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, M. Moonen, and A. Spriet, "Multicenter evaluation of signal enhancement algorithms for hearing aids". *J Acoust Soc Am*. vol. (2010, in press).

[6] B. Hagerman and A. Olofsson, "A method to measure the effect of noise reduction algorithms using simultaneous speech and noise". *Acta Acustica united with Acustica*. vol.2, pp. 356-361 (2004).

[7] M. Dahlquist, M.E. Lutman, S. Wood, and A. Leijon, "Methodology for quantifying perceptual effects from noise suppression systems". *Int J Audiol*. vol.12, pp. 721-32 (2005).

[8] B.C.J. Moore, B.R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness". *J Audio Eng Soc*. vol.4, pp. 224-240 (1997).

[9] B.R. Glasberg and B.C.J. Moore, "A model of loudness applicable to time-varying sounds". *J Audio Eng Soc*. vol.5, pp. 331-342 (2002).

[10] B.C.J. Moore and B.R. Glasberg, "A revised model of loudness perception applied to cochlear hearing loss". *Hear Res*. vol.1-2, pp. 70-88 (2004).

[11] R.J. Baker and S. Rosen, "Auditory filter nonlinearity across frequency using simultaneous notched-noise masking". *J Acoust Soc Am*. vol.1, pp. 454-462 (2006).

[12] ITU-P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", (2001).

[13] A. Olofsson and M. Hansen, "Objectively measured and subjectively perceived distortion in nonlinear systems". *J Acoust Soc Am*. vol.6, pp. 3759-69 (2006).