Multicenter evaluation of signal enhancement algorithms for hearing aids[a)]

Heleen Luts[b)], Koen Eneman[c)], Jan Wouters

ExpORL – Department of Neurosciences, Katholieke Universiteit Leuven, Herestraat 49 bus 721, B-3000 Leuven, Belgium

Michael Schulte, Matthias Vormann

Hörzentrum Oldenburg GmbH, Oldenburg, Germany

Michael Büchler, Norbert Dillier

Department of Otorhinolaryngology, University Hospital Zürich, Switzerland

Rolph Houben, Wouter A. Dreschler

AMC, KNO-Audiologie, Amsterdam, the Netherlands

Matthias Froehlich, Henning Puder

Siemens Audiologische Technik GmbH, Erlangen, Germany

[a)] Portions of this work were presented in "Signal processing in hearing aids: results of the HEARCOM project," Abstract of the Acoustics'08 Conference, Paris, France, June 2008 and in "Evaluation of signal enhancement strategies for hearing aids: a multicenter study," Abstract of the International Hearing Aid Research Conference (IHCON), Lake Tahoe, California, August 2008 and in "Evaluation of signal enhancement algorithms for hearing instruments," Proceedings of the 16th European Signal Processing Conference (EUSIPCO), Lausanne, Switzerland, September 2008.
[b)] Electronic mail: heleen.luts@med.kuleuven.be
[c)] Presently at Group T-International University College Leuven, Belgium

Giso Grimm, Volker Hohmann

Medical Physics Section, Carl von Ossietzky-Universität Oldenburg, Germany

Arne Leijon

Sound and Image Processing Lab, Royal Institute of Technology (KTH), Stockholm, Sweden

Anthony Lombard

Multimedia Communications and Signal Processing, Universität Erlangen-Nürnberg, Germany

Dirk Mauler

Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum, Germany

Ann Spriet[d]

ESAT/SISTA, Katholieke Universiteit Leuven, Belgium

Submitted to the Journal of the Acoustical Society of America on February 20, 2009.

Running title: Multicenter evaluation of signal enhancement algorithms

[d] Postdoctoral researcher funded by the F.W.O.-Vlaanderen

ABSTRACT

In the frame of the European HearCom project, promising signal enhancement algorithms were developed and validated for future use in hearing instrument devices. To assess the algorithm performance, five of the algorithms were selected and implemented on a common real-time hardware/software platform. A multicenter study was set up across four different test centers in Belgium, the Netherlands, Germany and Switzerland to perceptually evaluate the selected signal enhancement approaches. Listening tests were performed with large numbers of normal-hearing and hearing-impaired subjects. Three perceptual measures were used in these evaluations: speech reception threshold tests (SRT), listening effort scaling and preference rating. Tests were carried out in two types of test rooms and in different listening conditions. In a pseudo-diffuse noise scenario, only one algorithm provided an SRT improvement relative to the unprocessed condition. Despite the lack of improvement in SRT, increased preference was measured for a number of algorithms compared to the unprocessed condition at all tested signal-to-noise ratios. Additionally, an improvement in listening effort was observed at 0 dB SNR. These effects were found across different subject groups and test sites.

PACS numbers: 43.66.Ts, 43.71.Ky, 43.71.Gv, 43.60.Fg

INTRODUCTION

The main complaints of hearing aid users are speech understanding in noisy listening environments and spatial awareness (Noble and Gatehouse, 2006; Kochkin, 2005). By the advent of fully digital hearing aids around 1995 there was a general believe that different new signal processing strategies would be developed and (older) schemes would be integrated as software modules in the new digital engines. Several speech-in-noise enhancement strategies, including directional microphone systems, can be considered as a pre-processing or front-end to the core processing of the hearing aid or cochlear implant. Whereas a lot of research has been carried out on speech enhancement approaches in challenging noise environments, until now only a limited number of algorithms have found a breakthrough in actual commercial devices (Spriet *et al.*, 2007; Hu and Loizou, 2008; Blamey, 2005; Dillon, 2001). In fact, the customization of a signal processing scheme towards the implementation in a hearing aid or cochlear implant device makes strong demands in terms of computational complexity and processing delay, and requires a profound performance assessment through physical and perceptual validation tests.

The reason for this limited implementation and integration of digital signal enhancement techniques in commercial hearing instruments is twofold. First, a lot of smartly developed signal processing schemes have not been developed and sufficiently evaluated with the real-world application in mind. Secondly, only a limited number of these signal processing schemes have actually demonstrated real benefits for the hearing aid user both in laboratory environments and under real world daily life listening conditions. Moreover, the test results are hard to compare and extrapolate across different developers and test sites. Based on published results in the engineering literature it is sometimes difficult to

repeat the results because of the incomplete knowledge of details. In addition, the physical evaluations and objective measures are not always relevant for the perception of speech in noise. For instance, not only the signal-to-noise ratio (SNR) is important, but also intelligibility-weighted measures like the intelligibility-weighted SNR (Greenberg *et al.*, 1993) and the speech intelligibility index (SII) (ANSI S3.5, 1997). The evaluation should not be limited to experiments with artificial noise (e.g. white noise), which is important for first algorithm evaluations in-lab, but should include profound testing with real-life signals as well. Apart from that, perceptual evaluation experiments should be included, not only with normal-hearing, but also with hearing-impaired subjects.

One of the subprojects within the framework of the European research project HearCom (Hearing in the Communication Society, www.hearcom.eu) has focused on the development and evaluation of signal enhancement techniques for improved speech understanding in hearing aids. This research was carried out by a number of hearing aid signal processing research groups (Katholieke Universiteit Leuven, Universität Oldenburg, Friedrich-Alexander-Universität Erlangen-Nürnberg, Ruhr-Universität Bochum, Kungl Tekniska Hogskolan Stockholm). Based on physical performance measures (Eneman *et al.*, 2008a; 2008b) five signal enhancement techniques (single-channel as well as multi-channel) were selected from a large set of state-of-the-art algorithms and implemented on a common low-delay real-time test platform (Grimm *et al.*, 2006). The study presented in this paper comprises the perceptual evaluation of these five algorithms at four different test sites (Leuven, Zürich, Amsterdam, Oldenburg) with large numbers of normal-hearing listeners and hearing aid users of two different auditory profiles.

Adaptive speech reception threshold (SRT) tests were performed to evaluate the effect of the signal enhancement algorithms on speech intelligibility. Several previous studies on the perceptual benefits of single-channel noise reduction algorithms described increased comfort or ease of listening, while finding no evidence for improvement in speech intelligibility (Bentler *et al*., 2008; Marzinzik and Kollmeier, 1999; Ricketts and Hornsby, 2005; Walden *et al.*, 2000). Therefore, in the current study listening effort and overall preference were assessed besides the SRT measurements. The listening effort scaling and preference rating measurements were carried out at a number of fixed SNRs up to +10 dB, which are representative for many daily-life listening conditions. In this way, the effect of SNR on the algorithm performance can be monitored.

The following research questions are addressed and discussed in this paper: Do signal processing strategies lead to similar outcomes across different test sites, taking into account the different test environments/rooms, test materials and evaluation test platforms? Do the enhancement strategies yield different outcomes for different auditory profiles? How do the results on speech reception threshold, listening effort scaling and preference rating assessment vary across the signal processing algorithms?

I. SIGNAL ENHANCEMENT TECHNIQUES

This section describes the five signal enhancement algorithms that have been selected and evaluated.

A. Single-channel noise suppression based on perceptually optimized spectral subtraction (SC1)

Spectral subtraction is a well-known computationally efficient noise reduction technique. However, a major drawback of this approach is the presence of musical noise artifacts. By carefully selecting the amount of under- or oversubtraction, the enhanced signal can be perceptually optimized to eliminate the musical noise. To control the trade-off between speech distortion and noise suppression, the subtraction is made adjustable by incorporating a frequency dependent parameter $\alpha(k)$ that is a function of the noisy-signal to noise ratio. The algorithm variant that has been selected for the HearCom project is a low-delay version of the original perceptually tuned spectral subtraction method (PSS) (Samuelsson, 2006), which was obtained by reducing the time shift between subsequent frames. If the noise reduction filter would be allowed to change freely for every frame-shift block, the rapid filter variations would be perceptually unacceptable. Therefore, the parameters controlling the adaptation speed of the algorithm had to be properly tuned (Samuelsson, 2006). The noise spectrum is estimated from the modulation pattern of the noisy input signal, using the Minimum Statistics algorithm (Martin, 2001). This algorithm was also slightly modified to use data across the same time-span as in the original implementation, although the time-delay is reduced.

B. Wiener-filter-based single-channel noise suppression (SC2)

A second single-channel noise suppression algorithm that is considered in the HearCom project is a Wiener filter-based technique, which minimizes the mean squared error between the (unknown) desired speech signal and a filtered version of the observed noisy speech (Martin, 2001; Mauler and Martin, 2006). Since speech is stationary only on short time intervals, statistical expectation operations have to be replaced by short term averages. Therefore, instead of using the actual a-priori SNR, estimated a-priori SNR

values are computed following Ephraim and Malah (1984). The samples of the observed noisy speech signal are partitioned into overlapping frames, weighted with an analysis window and then transformed to the DFT domain. The enhanced speech spectral coefficients S(m,n), with frame index m and DFT bin n, are obtained as $S(m,n)=H(m,n)\cdot Y(m,n)$, where Y(m,n) are the noisy DFT coefficients and H(m,n) is a time and frequency dependent gain. As for the Wiener filter approach H(m,n) is real valued, only the amplitude of the noisy DFT coefficient is changed and the phase is left unchanged. After weighting with a synthesis window, the time domain signal is reconstructed via overlap-add operations. The algorithm variant used in the HearCom project is a low-complexity and delay-optimized solution. The frame length and frame shift have been reduced resulting in a larger frame overlap and hence an increased correlation of the spectral data. As a consequence, the noise power is typically underestimated. To overcome this problem an improved Minimum Statistics noise power estimator has been developed (Mauler and Martin, 2006). The estimation of the noise power spectral density via Minimum Statistics (Martin, 2001) rests on the observation that the power of the noisy speech signal frequently decays to the level of the noise. An estimate of the noise power can hence be obtained by tracking minima of the spectral power. Then, the bias between minimum and mean is compensated for. Due to the minimum principle, noise power estimation via Minimum Statistics does not require an explicit voice activity detection.

C. Broadband blind source separation based on second-order statistics (BSS)

The blind source separation (BSS) algorithm considered in the HearCom project is based on work published in Aichner *et al.* (2006), Aichner *et al.* (2007), Buchner *et al*. (2004),

Buchner *et al*. (2005b), and in Buchner *et al.* (2005a), where a class of broadband time-domain and frequency-domain BSS algorithms were derived that are based on second-order statistics. These broadband BSS approaches simultaneously take advantage of nonwhiteness and nonstationarity, and inherently avoid the permutation problem as well as circular convolution effects. Hence, no accurate geometric information about the placement of the sensors is needed. The algorithm selected for evaluation in the HearCom project is a low-cost, low-delay algorithm variant using frequency-domain-based fast convolution techniques. The algorithm is applied to bilateral hearing aids, using one microphone signal from each hearing aid as its inputs. This two-microphone implementation allows the separation of two point sources and offers therefore two output signals. Thus the output containing the desired signal has to be selected and presented to the hearing aid user. Additional diffuse sound sources have only a limited influence on the algorithm performance (Aichner *et al.*, 2006). Based on the approach in Buchner *et al.* (2005b) the time-difference-of-arrival (TDOA) of the sound waves originating from the separated sources can be determined from the demixing filters of the BSS algorithm, without any prior knowledge on sensor positions. Note that as the microphone spacing is not accurately known and as also head shadowing effects influence the TDOA estimate, accurate direction-of-arrival (DOA) cannot be calculated for each separated output source. However, it is assumed that the desired source is located approximately in front of the hearing aid user, which is a standard assumption in current state-of-the-art hearing aids. The TDOA estimates are then sufficient for identifying the source which is the closest to the position in front of the hearing aid user

(i.e., the source with the smallest TDOA). The output channel containing the desired source is selected based on this information.

D. Spatially preprocessed speech-distortion-weighted multichannel Wiener filtering (MWF)

The MWF is an adaptive noise suppression technique that is based on work described in Doclo *et al*. (2005), Doclo *et al*. (2007), Spriet *et al*. (2004), and Spriet *et al.* (2005). It consists of a fixed spatial preprocessor, i.e. a fixed beamformer and blocking matrix, and an adaptive stage. As a consequence, the MWF can be viewed as a variant of the well-known generalized sidelobe canceller (GSC) structure. Whereas in the case of the GSC the filter weights converge to a solution that merely reduces the residual noise, the cost function of the MWF approach minimizes a weighted sum of the residual noise energy and the speech distortion energy. In this way, a trade-off is provided between noise reduction and speech distortion. If the trade-off parameter in the cost function is set to infinity speech distortion is completely ignored and the algorithm reduces to a GSC structure. The MWF algorithm can therefore be considered as an extension of the GSC. As the MWF approach makes a trade-off between noise suppression and speech distortion, the algorithm is more robust against speech leakage than the standard GSC (Spriet *et al.*, 2004). Several algorithm variants have been developed, leading to cheaper implementation and/or improved performance *(*Spriet *et al.*, 2004; Spriet *et al.*, 2005; Doclo *et al.*, 2005; Doclo *et al.*, 2007). For the evaluation in the HearCom project a three-microphone version of the algorithm is considered that relies on a frequency-domain variant of the cost function and that uses efficient correlation matrix updating.

E. Binaural coherence dereverberation filter (COH)

Dereverberation algorithms are designed to increase listening comfort and speech intelligibility in reverberating environments and diffuse background noise (e.g. babble noise). The dereverberation technique studied in the HearCom project is a binaural coherence filtering based approach that builds on work described in Wittkop and Hohmann (2003). It estimates the coherence, i.e. the signal similarity, between the signals captured at the left and the right ear. The estimate is computed in different frequency bands using an FFT-based filter bank with a non-linear frequency mapping that approximates the Bark scale. As a coherence estimate, a time average of the interaural phase difference is computed. If the signals are coherent in a specific frequency band, the sound is expected to be directional, hence the gain in the frequency band is set to a high value. If on the other hand the coherence is low, a diffuse sound field is present, and accordingly, the frequency band is attenuated. The frequency-dependent gains are derived from the phase difference vector strength by applying an exponent (between 0.5 and 2) to the coherence estimate. High values for the exponent provide efficient filtering, but lead to more audible artifacts. Because of the head geometry, the coherence is always high at low frequencies, independently of the type of signal. At medium and high frequencies, on the other hand, the coherence is low for reverberated signal components (late reflections) and for diffuse babble noise, while it is high for the direct-path contribution of the signal of interest. Hence, by applying appropriate gains reverberated signal components and diffuse noise can be suppressed with respect to direct-path signal components.

II. Common evaluation platform

The signal enhancement algorithms have been implemented on a common real-time hardware/software platform, called Personal Hearing System (PHS). The hardware

platform consists of a (laptop) PC running a real-time low-latency Linux operating system. The PC is equipped with a multi-channel RME sound card, which is connected to a pair of hearing aids via a pre-amplifier box. The devices used in this study are Siemens Acuris behind-the-ear hearing aids with 3 microphones and a single receiver. There was no processor included in the hearing aid devices themselves. All signal processing (sampling rate 16 kHz) was done externally on the real-time Linux PC. All algorithm developers incorporated a C/C++ implementation of their algorithm into the Master Hearing Aid (MHA). This software environment simulates the processing performed by a real-life hearing aid, and is controlled by the PHS. Hence, apart from passing signals to and from the algorithms, the MHA software is also responsible for applying basic hearing aid processing to the signals, such as frequency dependent gain setting according to the audiogram of the subject (to compensate for reduced audibility in hearing-impaired listeners) and a calibration of microphones and loudspeakers. More information about the MHA and PHS can be found in Grimm *et al.* (2006) and Grimm *et al.* (2009).

The single-channel noise suppression algorithms SC1 and SC2 only use the front microphone signal and both hearing aids run the same algorithm with identical parameters independently of each other (double monaural system). The BSS and COH approaches are truly binaural algorithms using the front microphone of the left and the right hearing aid as their inputs. The MWF beamformer, on the other hand, is a multi-channel noise reduction algorithm that processes all 3 microphone signals of the hearing aid, and hence disposes of more degrees of freedom to cancel background noise. Similar to SC1 and SC2, the left and right hearing aid run the same algorithm independently of each other.

For a possible future integration in a commercial hearing aid device, the computational complexity and signal delay of the algorithms have to be monitored (see Table I). The computational complexity measurements were performed on a Dell Latitude D610 with Intel Pentium M 1.6 GHz processor running a Linux operating system. The baseline processing by the PHS-MHA system when all signal enhancement algorithms are switched off requires 10.3% of CPU time. The primary objective of the HearCom project was to prove the validity of the different algorithmic approaches in a hearing aid context. There is still some room left to further reduce the computational load of the algorithms. The total input/output delay from the signal sent into the AD converter to the signal that appears at the DA converter output was measured on a Dell Latitude D620 with Intel Core Duo 1.83 GHz processor running a Linux operating system. Table I shows the total input/output delay, which includes the combined delay of the PHS-MHA system and of the selected signal enhancement algorithm. With all signal enhancement algorithms switched off an input/output delay of 10.6 ms was measured. Both the PHS-MHA system and the signal enhancement algorithms operate in the frequency domain and thus require an analysis and synthesis filterbank. The algorithms SC1 and COH use the analysis and synthesis filterbank of the PHS-MHA system. The current implementation of SC2, BSS and MWF, however, uses a separate filterbank which causes an additional delay. In an optimized implementation with a shared filterbank, this delay could be reduced. There usually is a trade-off between processing delay and computational complexity. In this respect, the complexity numbers that are shown in Table I can typically be reduced at the expense of a larger processing delay.

III. MATERIALS AND METHODS

All five signal enhancement algorithms that were discussed in section II have been validated through listening tests in Dutch and German across four different test sites in Belgium (*ExpORL, Dept. Neurosciences, K.U.Leuven,* 'BE'), the Netherlands (*AMC, KNO-Audiologie, Amsterdam,* 'NL'), Germany (*Hörzentrum Oldenburg GmbH,* 'DE') and Switzerland (*Dept. Otorhinolaryngology, University Hospital Zürich,* 'CH'). Three types of tests have been performed in different noise scenarios and test rooms: an adaptive speech reception threshold (SRT) sentence test in noise, a listening effort scaling (LES) and a preference rating test (PR).

A. Subjects

In total 109 subjects participated in the study. The groups were defined based on audiogram information only. A first group consisted of 38 normal-hearing subjects (NH) with average hearing thresholds better than or equal to 20 dBHL for octave frequencies between 250 and 8000 Hz. The other 71 subjects had a moderate sensorineural hearing loss and were experienced bilateral hearing aid users (at least 6 months of experience). Bilateral inputs were used for all tests. Therefore, only a limited amount of asymmetry between both ears could be tolerated. To this end, the average absolute difference between left and right hearing thresholds at the octave frequencies 500 to 4000 Hz (referred to as symmetry) was aimed to be below 10 dB (range 1 to 19 dB, but only in 4 subjects more than 10 dB). The hearing-impaired subjects were divided in two groups based on the slope of their hearing loss. Slope was defined as the difference between the maximum and minimum hearing threshold for octave frequencies between 500 and 4000 Hz. The group of hearing-impaired subjects with a flat hearing loss (HI-F) had a slope

(averaged for both ears) of no more than 25 dB. The group with a sloping hearing loss (HI-S) had a slope of more than 25 dB. More details on the subject groups can be found in Table II. Degree, slope and symmetry are calculated for octave frequencies of 500 to 4000 Hz. The number of subjects tested at each test site was 30, 30, 28 and 21 for BE, NL, DE, and CH, respectively.

B. Fitting

Bilateral fittings were based on the ISO-389 headphone audiogram (octave frequencies between 250 and 8000 Hz). The hearing aids were fitted with the NAL-RP prescription rule (Byrne *et al.*, 1991). No compression was included, only a limiter set at 100 dBSPL. The gain was adjusted in eight frequency bands. Fine-tuning was limited to only two situations, namely wrong overall gain or the occurrence of feedback. Instead of earmolds, disposable foam earplugs with tubing were used for all ears. This is a more systematic approach across subjects compared to using the subjects' own earmolds. In normal-hearing subjects the amplification settings only compensated (depending on frequency) for the occlusion of the ears, but no additional gain was provided.

C. Room characteristics

To investigate the effect of reverberation time on the performance of the speech enhancement algorithms, evaluation tests have been conducted in two types of listening rooms: office-like rooms, representative for many everyday listening conditions, and highly reverberating rooms, which have been included to validate the algorithms under more challenging acoustic conditions. At all test sites the evaluation tests were conducted in an office-like room, with a target reverberation time for frequencies between 300 and 8000 Hz (measured according to the ISO 3382-1997-standard) between 300 and 600 ms.

The critical distance for each of these rooms was 128, 102, 186 and 145 cm for BE, CH, DE and NL respectively. Additionally, at two test sites (BE and DE) the algorithms were assessed in a highly reverberation room with a reverberation time larger than 1 second. The critical distance of these rooms was 37 and 119 cm for BE and DE respectively. The maximum permissible background noise level in the rooms was set to 35 dB(A). For both types of rooms no additional specifications (room dimensions, room organization, position of test subject etc.) have been defined. At DE the measurements in office-like as well as reverberant conditions were carried out in a room called the "Communications Acoustic Simulator", which allows to systematically change the room acoustics.

D. Environmental conditions

During the perceptual evaluation tests the listener is given a hearing aid pair controlled by the PHS-MHA. The subject is seated in the test room amidst 4 loudspeakers that are positioned at 0°, 90°, 180° and 270°, at 1 meter distance from the center of the listener's head. Only in the highly reverberating room of partner DE the distance between loudspeakers and listener was increased to 2 meters. All loudspeakers were directed towards the listener. The loudspeakers had a flat frequency response and provided no audible distortion up to at least 80 dB(A). Speech was always presented through the front loudspeaker. Two different noise configurations were used. A first noise configuration consisted of 3 uncorrelated noise sources at 90°, 180° and 270° (S0N90/180/270). All algorithms were evaluated in this noise scenario, in the office-like room as well as in the highly reverberating room. In the highly reverberating room this leads to a diffuse noise scenario as the loudspeakers are positioned outside the critical distance of the room. In the office-like room, however, this is not the case and the noise field can be called

pseudo-diffuse. A second noise configuration, which was only used in the office-like room, consisted of one interfering noise source at 90° (S0N90). Only two algorithms (BSS and MWF) were also evaluated in this single point-source scenario.

A multitalker babble noise from the CD *Auditory Tests (Revised)* (Auditec of St. Louis) was used. The noise sources were calibrated to produce a combined sound level of 65 dB(A) at the center of the listener's head. The speech level was then adjusted to obtain the desired signal-to-noise ratio. The noise started 5 seconds before the first speech sound was presented, to allow the algorithms to initialize properly.

E. Evaluation measures and speech materials

All subjects performed three types of listening tests: adaptive speech reception threshold tests, listening effort scaling and preference rating. SRT, LES and PR were measured for all algorithms (and for the unprocessed condition) at all four test sites in an office-like room with three noise sources of multitalker babble with a fixed combined noise level of 65 dB(A). These results can be used to determine the effect of test site, subject group and test-retest reliability. LES and PR were carried out after the SRT test so the subjects were already familiar with the listening situation. For SRT, a number of additional measurements were performed in other test conditions in a subset of test sites.

1. Adaptive speech reception threshold test (SRT)

At the Dutch-speaking test sites the VU-sentences (male speaker) were used *(*Versfeld *et al.*, 2000). This is an open-set sentence test. Binary sentence scoring was applied. This speech material consists of 39 lists of 13 sentences. An adaptive test procedure was used. The SRT was defined as the average of the ten last speech presentation levels (including the 14th imaginary level). The noise was presented at a fixed level and the level of the

sentences was adapted in 2-dB steps. At BE, the Apex software was used for the SRT testing (Francart *et al.*, 2008). At NL own software was used.

At the German-speaking test sites the OLSA sentence test was utilized (Wagener *et al.*, 1999). This is a closed-set sentence test. This speech material (male speaker) consists of 10 test lists of 10 sentences with a fixed structure (5 words), combined to lists of 20 sentences. Binary sentence scoring was used. The adaptive procedure and the fit are explained in Brand and Kollmeier (2002). The OLSA sentence test was used as incorporated in the Oldenburg Measurement Applications software package (www.Hoertech.de). Two training lists were assessed in quiet (20 sentences each).

SRT results of the algorithms were always compared to the SRT in the unprocessed condition, i.e. when all signal enhancement algorithms were switched off and only the basic processing of the PHS-MHA system was activated. In this way, the improvement of speech understanding in noise was assessed. Moreover, absolute differences in SRT between the subjects were taken into account. An overview of all test conditions is shown in Table III. All test conditions were randomized and conducted twice (test and retest).

2. Listening effort scaling (LES)

The goal of the listening effort scaling was to evaluate the individual listening effort with distinct hearing aid algorithms. Each subject had to rate the effort for each algorithm and for the unprocessed condition at 5 different SNRs: -10, -5, 0, +5 and +10 dB SNR. In this way, the algorithm performance can be evaluated over a broad range of SNRs. The subjective scaling was performed using a 13 point scaling (7 subcategories, and always one empty button in between). The subcategories ranged from “extreme effort” (score 6) to “no effort” (score 0). LES was performed in the office-like room, with three interfering

noise sources (S0N90/180/270). The same speech material as for the SRT test was used. All conditions were tested and retested, resulting in a total of 60 trials to be rated.

3. Preference rating (PR)

The perceptual differences between the 5 speech enhancement algorithms and the unprocessed condition were also investigated with a preference rating test. As the perceptual differences between the algorithms are expected to be small, a paired comparison test was used. The algorithms were pairwisely compared to the unprocessed condition. Preference rating was done under office-like room conditions only, with three interfering noise sources (S0N90/180/270). Each pair of algorithms was presented at 3 different SNR ratios (0, +5 and +10 dB SNR) and every presentation was conducted twice (test and retest). Per test person, this leads to a total of 30 trials to be rated.

During the preference rating test the subject could listen to the algorithms as long as needed and could toggle between the algorithms as often as wanted. After having indicated a preference, the difference between the algorithms had to be graded, i.e. the subject had to rate how much better the preferred algorithm was compared to the other one. The outcome of the test is the amount of preference of an algorithm over the unprocessed condition. This preference score can vary between ‘very slightly worse’ to ‘very much worse’ (scored as -1 to -5) and ’very slightly better’ and ‘very much better’ (scored as +1 to +5). Since the preference rating test uses a forced-choice paradigm, equal preference (that would be represented by score 0) was considered not to be a valid response.

The preference-rating data was evaluated according to Dahlquist and Leijon (2003), where it is assumed that each algorithm has a specific neural representation in the sensory

system. This representation can be modeled as the outcome of two random variables which are supposed to be normally (Gaussian) distributed: firstly the preference (algorithm-under-test better than the unprocessed condition or vice versa) between the algorithms and secondly the "grading of the preference" which can be interpreted as the confidence of the subjective preference judgments. The Linear Gaussian Model (LGM) places the algorithms on an interval scale and estimates the (unknown) parameters, leading to a nonlinear and multidimensional maximization problem. The aim of rescaling the data with the LGM is to obtain a perceptually correct scale. The LGM provides the limits of the estimated confidence intervals. These limits can be used to evaluate the differences between two specific algorithms in terms of the 5 categories that were used by the subjects to grade the algorithms. As the unprocessed condition is used as a reference, all results are shifted so that the reference is placed at scale value zero. The LGM is applied to each set of data for every subject. Afterwards the results for the individuals are pooled together.

F. Statistical analyses

Statistical analyses were carried out with the SPSS software. To test whether the distribution of the variables is normal, Kolmogorov-Smirnov tests were used. To analyze the SRT data, repeated-measures analyses of variance (ANOVA) were carried out. For tests of within-subjects effects, the most conservative Lower Bound was taken. Post hoc tests consisted of pairwise comparisons. To keep the Type I error rate across all comparisons at 0.05 the Bonferroni correction was applied.

For LES and PR data, non-parametric statistics were applied. To compare several related samples, Friedman's ANOVA is used. Two related conditions are compared with the

Wilcoxon signed rank tests. To compare differences between several independent groups, the Kruskal-Wallis test was applied. Post hoc tests were done with the Mann-Whitney test. For preference rating the binomial test was used to check whether the algorithms were significantly more often preferred than the unprocessed condition. When several conditions had to be compared and consequently several tests were required, the significance level was always adjusted using the Bonferroni correction (unless mentioned differently).

IV. RESULTS

A. Test–retest reliability

Test and retest scores for SRT, LES and PR were compared for results obtained in the office-like room, with three interfering noise sources. Test and retest scores for all algorithms were taken together. In this way, for SRT only one paired comparison was carried out (using a paired-samples *t*-test). For LES and PR, data were compared for each tested SNR with a non-parametric Wilcoxon signed ranks test.

Test-retest scores for SRT were significantly different ($p \leq 0.001$). The mean difference was 0.8 dB with the retest scores being lower than the test scores. This might be due to a learning effect. For LES there was a significant difference between test and retest at -10, -5 and 0 dB SNR ($p \leq 0.001$) with the retest scores being, respectively, 0.1, 0.2 and 0.3 higher than the test scores. For PR, test and retest were not significantly different at any SNR (always $p > 0.5$).

Test-retest reliability was assessed by calculating the variability of the scores, measured as the within-subjects standard deviation of the scores (σ_w) with the formula:

$$\sigma_w = \sqrt{\frac{1}{2}\sum_{i=1}^{n}\frac{(x_{i1} - x_{i2})^2}{n}}$$

where x_{i1} is the i^{th} test score, x_{i2} is the i^{th} retest score and n is the total number of scores to compare. For SRT, the within-subjects standard deviation of the absolute scores was 1.8 dB. For LES, which is scored on a scale from 0 to 6, σ_w was 0.6, 0.9, 1.2, 1.0 and 0.7 at -10, -5, 0, 5 and 10 dB SNR respectively. The reliability is lower for the middle signal-to-noise ratios. At the lowest and the highest signal-to-noise ratio, subjects tend to give the maximum and minimum score respectively. At the extremes the variability between subjects as well as within subjects is thus small. For PR, which is scored on a scale from -5 to +5 the within-subjects standard deviation was 1.9, 2.0 and 2.0 for 0, +5 and +10 dB SNR, respectively.

For further analyses of the SRT results, repeated-measures analyses of variance (ANOVA) were carried out and test-retest was included in the ANOVAs as a within-subjects factor. For further analyses of LES and PR test and retest scores were averaged prior to analysis.

B. Speech reception threshold tests

1. Office-like room and three interfering noise sources

To compare the results of the different subject groups and test sites a repeated-measures ANOVA was carried out including the SRT scores measured in the office-like room with three noise sources, because these measurements were performed at all test sites. The analysis included two within-subject factors (test-retest and algorithms) and two between-subjects factors (subject group and test site). Tests of within-subjects effects showed a main effect for the factor test-retest ($p \leq 0.001$) and the factor algorithm ($p \leq 0.001$). The

mean difference between test and retest data was 0.8 dB with a standard error of 0.1 dB with lower scores for the retest. There were no interaction effects (always $p>0.05$). The effects of the different algorithms are thus equal for the test and the retest scores.

Both between-subject factors were highly significant (test site $p\leq0.001$ and subject group $p\leq0.001$). Pairwise comparisons of the test sites showed that the average SRT (averaged over all subject groups and all algorithms) for DE (-4.7 +/- 0.4 dB SNR) and CH (-3.4 +/- 0.5 dB SNR) were not significantly different ($p=0.290$). Similarly the average SRT for BE (0.2 +/- 0.4 dB SNR) and NL (-0.7 +/- 0.4 dB SNR) were not significantly different ($p=0.659$). In contrast, differences between German-speaking and Dutch-speaking test sites were highly significant (always $p\leq0.001$). As expected, pairwise comparisons of the subject groups (averaged over all test sites) showed that the normal-hearing subjects (mean SRT -5.2 +/- 0.4 dB SNR) have a significantly lower SRT than the hearing-impaired subject groups (always $p\leq0.001$). The difference between both hearing-impaired subject groups, however, was not significant ($p=1.000$). The group with flat hearing losses had an average SRT of -0.5 +/- 0.4 dB SNR, the group with sloping hearing losses -0.8 +/- 0.4 dB SNR.

As there are differences in absolute SRT scores between the subject groups and the test sites, another ANOVA was carried out to compare the SRT improvements of the different signal enhancement algorithms relative to the unprocessed condition. The analysis was comparable to the previous ANOVA, including two within-subject factors (test-retest and algorithm) and two between-subjects factors (subject group and test site). The main difference between both analyses is the dependent variable which is SRT scores in the first analysis and SRT improvement in the second analysis. The factor algorithm contains

six levels in the first analysis (SRTs for 5 algorithms and the unprocessed condition) and 5 levels in the second analysis (SRT improvement of the 5 algorithms relative to the unprocessed condition). Tests of within-subjects effects showed no main effect for the factor test-retest (p=0.435). There was a main effect of the factor algorithm ($p \leq 0.001$). There were no interaction effects (always $p>0.05$). Looking at the improvements in SRT relative to the unprocessed condition, both between-subject factors were not significant (test site p=0.942 and subject group p=0.953).

In what follows, relative scores will be used as there are no significant differences between test and retest, between subject groups and between test sites. Pairwise comparisons of the SRT improvements of the five algorithms showed that only BSS and MWF differ significantly from the unprocessed condition. The SRT score with MWF improved with 6.7 dB (Std. Error 0.2 dB) compared to the unprocessed condition ($p \leq 0.001$). With the BSS the SRT deteriorated with 1.9 dB (Std. Error 0.3 dB) compared to the unprocessed condition ($p \leq 0.001$). Figure 1 and Figure 2 show the mean SRT improvements (averaged over all subjects and test/retest) and standard deviations that were obtained relative to the unprocessed condition. Positive numbers in the figures indicate an improvement in speech understanding with respect to the unprocessed condition. In Figure 1 the results for the different test sites can be compared. Figure 2 shows the results for the three subject groups.

2. Evaluation of BSS and MWF in two different noise scenarios

At three test sites (BE, NL and CH) additional measurements were carried out in the office-like room with one noise source for the BSS and MWF. An ANOVA on the SRT improvements with three factors (test-retest, noise scenario and algorithm) showed a

main effect of noise scenario and algorithm and an interaction between both (p always ≤0.001). As shown in Figure 3, the performance of the MWF improves significantly from 5.4 (+/- 2.2) to 6.9 (+/- 2.1) dB when adding more noise sources. The BSS performs well with one interfering noise source (6.4 +/- 2.2 dB improvement), but breaks down for three uncorrelated noise sources (-2.0 +/- 2.8 dB relative to the unprocessed condition).

3. Comparison of test rooms

At two test sites (BE and DE) the algorithms were also evaluated in a highly reverberating room and three interfering noise sources. This enables investigation of the effect of room type on the absolute SRT results and on the SRT improvements. The absolute SRT in the highly reverberant room was on average 6.7 dB higher at BE and 5.2 dB higher at DE compared to the SRTs of the same subjects in an office-like room. This means that speech intelligibility is more difficult in the reverberating room compared to the office-like room. Again, the SRT improvements of the algorithms relative to the unprocessed condition are most interesting. A 3-factor repeated-measures ANOVA of the SRT improvements was carried out on the factors test-retest (2 levels), room (2 levels) and algorithm (5 levels). Subject group and test site were between-subject factors. The between-subject factors subject group and test site were not significant (p>0.05). The main effect of room was not significant (p=0.697), but there was a significant interaction between room and algorithm performance (p=0.009). The interaction effect is caused by the SRT improvement of the MWF which was lower in the reverberant room (5.4 +/- 2.2 dB) compared to the office-like room (6.7 +/- 2.0 dB). The effects of the other algorithms were comparable between both rooms.

4. Additional analyses of subject groups

In all previous analyses the effect of subject group was not significant. Subject groups were divided based on audiogram information only and the reduced audibility in hearing-impaired subjects was compensated. The main difference between both hearing-impaired groups was the slope of the hearing loss so there was a large degree of overlap in the audiograms. Therefore, additional analyses were carried out to investigate the effect of subject group more in depth. Pearson correlations were calculated between the SRT improvements (relative to the unprocessed condition) and three subject variables that can be related to the auditory profile: the average degree of hearing loss, the average slope of hearing loss (both for octave frequencies from 500 to 4000 Hz) and the SRT in noise for the unprocessed condition (average of test and retest) which represents the capability to understand speech in noise. Correlation analyses were done for each of the subject variables in relation to the SRT improvements for the 5 algorithms. Therefore the significance level is adjusted to 0.01. Degree and slope were not significantly correlated with any of the 5 algorithms. The subject variable SRT in noise, however, was positively correlated with the SRT improvement of SC1, MWF and COH. The correlation coefficients were 0.272 (p=0.004), 0.332 ($p\leq$0.001) and 0.466 ($p\leq$0.001) for SC1, MWF and COH respectively.

In addition to these correlation analyses, the same repeated-measures ANOVA as described above was carried out on the SRT improvements. Subjects were divided in four groups based on degree, slope and SRT in noise (see Table IV). The within-subjects factors were test-retest and algorithm and the between-subjects factor was group. Similar results were obtained as in the correlation analyses. For degree and slope the between-

subjects factor was not significant (p=0.326 and p=0.681, respectively). For SRT in noise, however, there was a significant effect of group (p=0.011) and an interaction effect of group and algorithm (Lower bound, p=0.045). This effect is depicted in Figure 4. The positive correlation that was found is clear for SC1, MWF and COH. For these algorithms a higher SRT in noise for the unprocessed condition is linked with a larger improvement. The improvement increased over the four subject groups with 0.9 dB, 1.7 dB and 1.6 dB for SC1, MWF and COH, respectively. For SC2 differences between the groups were small. The interaction effect between algorithm and subject group is caused by the results of the BSS, which show the same trend for the first three groups, but the group with the highest SRT in noise has the worst performance.

C. Listening effort scaling

Unlike the SRT scores, the LES scores are not normally distributed. Moreover, the distribution is very different for the different signal-to-noise ratios. Therefore, non-parametric tests were used. For illustrative purposes, (parametric) repeated-measures ANOVAs were also carried out, as non-parametric tests do not allow multifactorial comparisons. The results of the ANOVAs are used to select the most interesting effects to investigate with non-parametric statistics. However, the results of these ANOVAs need to be interpreted with care.

1. LES for the unprocessed condition

A repeated-measures ANOVA was carried out on the LES scores for the unprocessed condition (average of test and retest scores) mainly to detect baseline differences in absolute listening effort scores between subject groups, test sites and signal-to-noise ratios. There was a significant effect of both between-subjects factors ($p \leq 0.002$) and of

the factor signal-to-noise ratio ($p \leq 0.001$). There was no interaction between test site and subject group ($p=0.227$).

The LES scores for the unprocessed condition are depicted in Figure 5 as a function of signal-to-noise ratio for the four test sites (left figure) and three subject groups (right figure). At the five signal-to-noise ratios the effect of test site and subject group was evaluated with a non-parametric Kruskal-Wallis test. The effect of test site was significant at all signal-to-noise ratios except at -10 dB SNR (always $p \leq 0.005$). The differences between test sites clearly depend on signal-to-noise ratio. This effect is similar to the effect observed in the SRT scores. The SRTs for the Dutch-speaking subjects were higher than for the German-speaking subjects. Similarly, the Dutch-speaking subjects require more effort than the German-speaking subjects at the same SNR.

The effect of subject group was significant at -5, +5 and +10 dB SNR (always $p \leq 0.005$). At these signal-to-noise ratios Mann-Whitney tests were carried out to investigate the differences between the groups. At all three signal-to-noise ratios the hearing-impaired subject groups did not differ significantly (always $p > 0.4$), but the normal-hearing group differed significantly from both hearing-impaired subject groups (always $p \leq 0.011$). Again, LES scores are in parallel with the SRT scores for the unprocessed condition: compared to normal-hearing subjects, hearing-impaired subjects require a higher SNR to reach 50% intelligibility. Accordingly, at the same SNR, hearing-impaired subjects require more effort compared to the normal-hearing subjects. No difference was observed between both hearing-impaired subject groups for SRT as well as LES scores.

2. LES scores relative to the unprocessed condition

Similar as for SRT, we will focus on scores relative to the unprocessed condition for LES. A positive relative score means that less effort was required relative to the unprocessed condition. Again a repeated-measurements ANOVA was carried out on the relative LES scores to get an idea about main effects and interaction effects. The within-subjects factors were algorithm (5 levels) and signal-to-noise ratio (5 levels). Between-subjects factors test site and subject group were added. Because there was a significant effect of test site ($p \leq 0.001$) and several factors interact with test site, a similar analysis was done for each test site separately. In none of these analyses the factor subject group was significant (always $p > 0.142$). Therefore, the results shown in Figure 6 are averaged over the subject groups. In all four analyses there was a significant interaction effect between algorithm and signal-to-noise ratio. Analogous to Figure 1, LES differences relative to the unprocessed condition are shown for the different test sites, and overall for the complete data set. Because the results are not normally distributed, medians and quartiles are shown. Results are only depicted for 0 dB SNR as the largest effects are obtained at this SNR. To know if the effort scores for the algorithms were different compared to the unprocessed condition, Wilcoxon signed ranks tests were carried out. As 25 comparisons were needed the significance level was decreased to 0.002 (p=0.05/25). Algorithm scores that are significantly different from the unprocessed condition are indicated with an asterisk. The figure shows the LES scores relative to the unprocessed condition. The absolute effort scores for the algorithms can be estimated by adding the relative scores to the average effort scores for the unprocessed condition as shown in Figure 5. As becomes clear from the figure, results are different across test sites.

However, general trends are similar. It is the size of the effects that differs. Overall, we can conclude that the MWF requires less effort than the unprocessed condition. This is especially the case at the lower signal-to-noise ratios, so in the more difficult conditions. BSS is the only algorithm which requires significantly more effort than the unprocessed condition, and this is the case at signal-to-noise ratios ranging from -5 to +10 dB SNR. This effect, however, is not present at all test sites. The other algorithms do not seem to have a large effect on the effort scores. Only at 0 dB SNR, SC1, SC2 and COH have an overall positive effect on the effort scores.

D. Preference rating

1. Win count

A first way of analyzing the preference rating results is by looking at the number of wins, which is the number of times the algorithm was preferred over the unprocessed condition. If the subjects would have no preference the number of wins would be 50%. Figure 7 shows the percentage of wins of the 5 algorithms for the three tested SNR levels. SC1, SC2, MWF and COH are significantly more preferred than the unprocessed condition. In case of the BSS, subjects prefer the unprocessed condition more often (binomial test, always $p \leq 0.001$). The general trends are the same for the three SNRs.

To compare the percentage of wins between the different algorithms McNemars χ^2 test was applied. To limit the number of comparisons, the total percentage of wins for all three SNRs was compared. Except for the difference between SC1 and SC2, and SC2 and COH, all differences were significant (always $p \leq 0.001$, with a significance level of $p=0.005$ after Bonferonni correction).

Unlike with the SRT results, SC1, SC2 and COH do show improvements compared to the unprocessed condition as shown by the number of wins. The MWF has the highest percentage of wins. However, the difference with the other algorithms is very small. Similar to the SRT results for three noise sources, the percentage of wins is lower for BSS than for the unprocessed condition.

2. Degree of preference

The raw PR scores were not normally distributed, and thus parametric statistics were not appropriate. Similar as for LES, an exploratory repeated-measures ANOVA was carried out to search for main effects in the data. These effects were further investigated with non-parametric techniques. The results of the ANOVA again need to be interpreted with care. The main purpose of the ANOVA is to make a selection of effects to investigate, to avoid too many statistical comparisons which would increase the number of statistical errors. The ANOVA was carried out on the PR scores (raw data), averaged over test and retest. The within-subjects factors were algorithm and SNR. The between-subjects factors were test site and subject group. There was a significant main effect of algorithm and SNR ($p \leq 0.001$), and an interaction effect of algorithm and subject group ($p=0.007$). Other effects were not significant. This means that results can be generalized over the four test sites. The significant effects were further investigated by non-parametric techniques.

Figure 8 shows the effect of algorithm and SNR, averaged over the test sites and subject groups. The results are shown on a LGM Scale. For a noise scenario with three noise sources, MWF shows the highest preference score. SC1, SC2, and COH also show a positive preference over the unprocessed condition. Less favorable results were obtained for BSS than for the other algorithms. For each of the algorithms the effect of SNR was

significant (Friedman's ANOVA, always $p \leq 0.001$). The degree of preference (positive or negative) is in general lower at 0 dB SNR than at the higher SNRs. A Kruskal-Wallis test on the rescaled data showed a significant effect of subject group only for COH ($p=0.001$). COH is preferred more often than the unprocessed condition by the HI subjects, but not by the NH subjects.

E. Correlation SRT – LES – PR

SRT, LES and PR scores were obtained in all subjects and in the same test conditions: a pseudo-diffuse noise scenario with three sources of multitalker babble noise in an office-like room. Results of the three tests can be correlated.

1. Relative scores

For PR only scores relative to the unprocessed condition are available. Consequently, correlations for PR, SRT and LES are calculated for the relative scores (for S0N90/180/270). As not all variables are normally distributed, Spearman correlations are calculated. Doing this for each algorithm separately does not result in any significant correlation. This is not so surprising, as the variability between subjects (and thus the data range) is relatively small for the relative scores, compared to the absolute scores. Several algorithms do not show any improvement for SRT and LES scores, and thus no significant correlations can be observed.

To investigate if the improvement in SRT score is related to the improvement in LES or PR score, correlations are calculated for all algorithms together. The SRT scores are significantly correlated (always $p \leq 0.001$) with LES and PR scores at all tested SNRs (-10, -5, 0, 5 and 10 dB SNR for LES and 0, 5 and 10 dB SNR for PR). For LES the correlations range from 0.192 at 10 dB SNR to 0.418 at -5 dB SNR. For PR the

correlations range from 0.316 at 10 dB SNR to 0.379 at 0 dB SNR. The average SRT over all algorithms is between 0 and -5 dB SNR for the different subject groups, so it is not surprising that the highest correlations are obtained at these (low) SNRs.

PR and LES are also significantly correlated (always $p \leq 0.001$). The Spearman correlations are 0.237, 0.179 and 0.200 at 0, 5 and 10 dB SNR respectively.

2. Absolute scores

LES and SRT can also be compared using the absolute scores. Overall correlations (for all algorithms and the unprocessed condition together) range from 0.439 to 0.547 (always $p \leq 0.001$). The highest correlation is found at -5 dB SNR, although differences are small. Correlations for each algorithm separately are also all significant at p=0.05 (without correction for multiple comparisons). At -5 dB SNR, Spearman correlations are 0.454, 0.348, 0.442, 0.499, 0.277, 0.437 for the unprocessed, SC1, SC2, BSS, MWF and COH condition, respectively (always $p \leq 0.004$).

V. DISCUSSION

A. Effect of test site

One of the research questions in this study was to evaluate how comparable the results are when measured at different test sites. To ensure comparability of the test results across the five test sites, special attention had to be paid to the test protocol. On the one hand, the protocol had to be flexible to allow inherent variations e.g. on the specification of the test rooms and on the auditory profile of the test subjects. On the other hand, the protocol needed to be unambiguous and sufficiently strict to minimize the impact of unknown external factors on the test results. The effect of test site and subject group was

evaluated using the dataset measured in the office-like room with three interfering noise sources, as these test conditions were common for all test sites and subject groups. This gives an indication of how important specifications are on room characteristics, speech material, test procedure, instruction, etc. Comparing data from different test sites points out how reproducible the results are.

For the absolute SRT scores, a significant effect of test site was observed. There was a difference between the German-speaking and Dutch-speaking test sites. This is not surprising, as each speech material has its own reference psychometric curve with an average SRT and slope. The SRT of the German OLSA sentences in stationary speech-weighted noise for normal-hearing subjects is lower than the SRT in the same conditions for the Dutch VU sentences. In the current study, the overall SRT for the OLSA sentences in multitalker babble noise was also lower than the SRT of the VU sentences (about 3.6 dB overall). However, this absolute shift in SRT scores can be taken into account by looking at the relative scores or improvements relative to the unprocessed condition. Indeed, for the relative scores no difference between test sites was observed. In this study, the test procedure, calibration and fitting were precisely defined and a common evaluation platform was used. Other parameters such as the specific type of loudspeaker, sound card, speech material and exact room characteristics like size and reverberation time differed between test sites because of practical limitations. However, test results indicate that variations in these parameters (controlled to a certain extent) are less important and do not significantly influence the SRT improvements of the algorithms.

For PR, there were no significant differences between the test sites. For LES, the absolute scores for the unprocessed condition were different across test sites, similar to the SRT scores. Apparently, differences in SRT are revealed by the LES scores. Looking at the relative LES scores, slightly different results were observed between the test sites. However, general trends were the same across test sites. It was mainly the size of the effects that differed. It is not clear what the cause of these differences was. Test instructions were uniform across test sites.

B. Effect of subject group

Similar as for the effect of test site the absolute SRT and LES scores showed differences between subject groups, but the relative scores didn't. As expected, hearing-impaired subjects had higher (worse) absolute SRT scores than normal-hearing subjects. Nevertheless, the SRT improvements were not significantly different between the subject groups. For LES (relative scores) and PR, results were also very similar between subject groups.

The results of the physical evaluation of the algorithms (not reported in this paper) did show differences in algorithm performance for different auditory profiles (Eneman *et al.*, 2008a; 2008b). According to the physical evaluation NH subjects were expected to benefit more from the algorithms compared to the HI subjects. The SII-improvement and the segmental SII-improvement were the largest for the normal-hearing auditory profile. Additionally, a lower benefit was expected for the flat hearing losses compared to the sloping-type of hearing losses. The perceptual results may seem contradictory with the physical evaluation, but there are several possible explanations for this discrepancy. First, it is reasonable to assume that the noise suppression algorithms succeed in improving the

SNR by some dB, at least for positive SNRs. As they operate on the noisy speech signals before they are delivered to the NH or HI subject, similar improvements in SRT are expected for all profiles. Second, the segmental SII may reflect the expected algorithm improvement on intelligibility at a fixed sound-field SNR, whereas the SRT improvements reflect the increase in sound-field noise levels that the algorithm can compensate for. Third, the SRT (defined as the SNR for 50% intelligibility) and SRT improvement are determined at very different sound-field SNR levels for the NH and HI subjects. On the one hand, it was postulated that the algorithms would perform worse in terms of SRT improvement at low (negative) SNRs, e.g. the speech-noise detector of MWF will perform worse at low SNRs. The algorithms may decrease the noise level at these low SNRs, but this may be accompanied by a decrease of the speech level as well, as it is more difficult to separate the speech and the noise. This might have decreased the algorithm benefit measured in NH subjects compared to the benefit measured in the HI groups that have their SRT at higher SNRs. On the other hand, more noise can be suppressed when the SNR is low. Although speech will get distorted more at lower input SNRs, the noise reduction may still be selective, i.e. it will affect mainly those points in the time-frequency plane and the spatial plane where the noise dominates the speech. Therefore, although speech components might become attenuated, noise attenuation may still dominate so that an SNR improvement can be measured also for signals with low input SNR. Fourth, in the perceptual evaluation study the groups were defined on audiogram information only. The major difference between the HI-S and HI-F groups is the slope of the audiogram between 500 and 4000 Hz. Other factors of the auditory profile might have more impact on the perceptual scores. The additional analysis of

subject group already showed a greater algorithm improvement for participants with worse SRTs for the unprocessed condition. However, as each SRT measurement includes a random error, any random apparent worsening of the unprocessed SRT will automatically show up also as a corresponding apparent increase of the improvement. This may have contributed to the correlation. An extension of this study will relate the perceptual results to the auditory profile.

C. Algorithm performance: SRT results

In the pseudo-diffuse noise scenario, MWF was the only algorithm that provided a significant SRT improvement relative to the unprocessed condition. Overall, an improvement of 6.7 dB was obtained with the MWF. The single-channel noise reduction methods SC1 and SC2 on average did not impact speech intelligibility. This is remarkable, since in single-channel noise reduction the competing signal cannot be removed without impairing the desired signal. Single-channel noise reduction is a challenging task in particular at low (negative) input SNRs at which the SRT is measured. The BSS had a negative effect and deteriorated speech intelligibility relative to the unprocessed condition with on average 1.9 dB. This is not really surprising as the two-channel BSS algorithm considered in this project was designed to separate two distinct point sources only. At three test sites, BSS and MWF were also evaluated in a single point-source condition. The multi-microphone enhancement techniques such as BSS and MWF can take advantage of the spatial diversity of the setup and therefore succeed in improving speech intelligibility in this noise scenario. Whereas the MWF performs well in both noise scenarios (with 1 and 3 noise sources), the BSS algorithm seems optimally suited for the single point-source scenario only.

An explanation for the rather poor performance of the COH algorithm in the presence of diffuse-like noise might be that the noise field was created by playing back time-shifted multitalker babble noise files through three distinct loudspeakers. This may not be truly realistic and may offer a higher binaural coherence than a natural multitalker babble noise, especially if the loudspeakers are located within the critical distance (such as in the office-like room). This observation in fact reveals the trade-off that had to be made between a real-life diffuse sound reproduction that is as realistic as possible for the application we have in mind, on the one hand, and compliance with standardized SRT testing procedures and preserving test precision, reproducibility and compatibility of test results across test sites, on the other hand.

The additional measurements in the reverberating room show that speech understanding in a reverberating room is much more challenging than in an office-like room. The absolute SRT values dropped significantly, e.g. by 6.5 dB for the unprocessed condition at BE and 5.2 dB at DE, when going from the reverberating to the office-like room. However, looking at the SRT improvements very similar results were observed in both rooms. Only for MWF, the SRT improvement was significantly different in both rooms. In the reverberating room a decrease of SRT improvement of 1.3 dB was observed. Nevertheless, even in these very demanding listening circumstances, an improvement of 5.4 dB was obtained.

D. Algorithm performance: comparison of SRT, LES and PR

As a consequence of the adaptive procedure, SRTs (expressed as SNR required for 50% intelligibility) and SRT improvements were measured at different sound-field SNRs depending on the subjects' hearing status. Additionally, in many subjects the SRT is

obtained at negative SNR levels. This is a disadvantage of the speech intelligibility measurements, as it is known that several signal enhancement algorithms perform differently at low and high SNRs. Nonetheless, speech intelligibility improvements are most required in difficult listening situations (e.g. at 50% intelligibility) which occur at different SNR levels depending on the hearing loss. Therefore, SRT-measurements are considered to be highly relevant in the perceptual evaluation of speech enhancement algorithms. In addition to the SRT measurements, LES and PR can evaluate the algorithm improvements at high input SNRs, as the measurements are performed at different fixed SNR levels up to +10 dB SNR.

The LES scores were in line with the SRT results. In contrast with SRT, LES provides results at different SNR levels. Looking at the relative LES scores, differences between test sites were noticed in the size of the effects. However, some general observations could be made. In the pseudo-diffuse noise scenario, the MWF required less effort than the unprocessed condition, especially at the lower SNRs or the more difficult listening conditions. At the higher SNRs, speech was easy to understand in the unprocessed condition, and thus the MWF cannot decrease the listening effort that much. At some test sites, the BSS shows a negative effect on the listening effort, mainly at the higher SNRs, although variability is large. This is not truly unexpected since the reported LES scores (and PR scores) were obtained in a scenario with three uncorrelated noise sources, whereas the two-channel BSS algorithm considered in this project was designed to separate two distinct point sources.

The LES revealed improvements for algorithms which were not detectable with the SRT procedure, analogous to the results of Marzinzik and Kollmeier (1999) and Walden *et al.*

(2000). They showed a significant positive effect of noise reduction algorithms by rating the listening effort, although no improvement in speech intelligibility could be shown by SRT tests. Similarly, in the current study an improvement in listening effort for SC1, SC2 and COH was observed at 0 dB SNR, although no SRT improvement was measured for these algorithms. At the other signal-to-noise ratios LES could not reveal improved listening effort for these three algorithms.

In the PR results, we see higher preference compared to the unprocessed condition for MWF, and also for SC1, SC2 and COH, at all three tested SNR levels. So unlike the lack of improvement in SRT, PR does measure differences for a number of algorithms. Moreover, the effect of SNR that was observed in the LES results cannot be observed in the PR results. In contrast, the preference for the algorithms is higher at the high SNRs (+5 and +10 dB SNR) compared to 0 dB SNR. The increasing PR score for SC1, SC2, and COH for increasing SNR are most probably due to less (audible) speech and noise distortion at higher input SNR. Analogous results were obtained in a study of Ricketts and Hornsby (2005). They used a similar paired-comparison approach to determine preference. Although the single-channel noise reduction of a commercial hearing aid did not affect speech perception, their results indicated strong preference for the noise reduction algorithm in both low-level and high-level noise. Natarjan *et al.* (2005) also reported increased speech quality ratings, while no measurable changes in intelligibility occurred. A study of Alcantara *et al*. (2003), however, found no preference for a commercial noise reduction algorithm.

The discrepancy between LES and PR has different possible causes. Firstly, different dimensions are rated. LES rates the effort to understand speech in a noisy situation and

PR rates the preference for an algorithm in a noisy situation, which is much broader and includes the general quality of an algorithm. Secondly, the psychophysical method of LES and PR is different. LES makes use of a categorical magnitude rating, whereas PR is based on paired-comparison. In addition to this methodological difference (and as a consequence), the LES scores show clear ceiling effect at the lowest and highest SNR, as subjects tend to give the minimum and maximum score, respectively. Because of these ceiling effects it might be more difficult to notice differences between the algorithms and the unprocessed condition. This is avoided in the paired comparison procedure used for the PR test.

VI. CONCLUSIONS

In this multi-center study, five different state-of-the-art signal enhancement approaches were evaluated through listening tests with large numbers of normal-hearing listeners and hearing aid users with flat and sloping-type moderate hearing losses. The SRT improvements and PR scores for the different signal processing strategies were comparable across the different test sites. This indicates that the results are not influenced by e.g. exact room characteristics or specific type of loudspeaker, sound card and speech material. For LES, slightly different results were obtained across the test sites. Additional analyses are necessary to further interpret these differences. For the SRT improvements, LES improvements and PR scores, outcomes were very similar for the different subject groups that were tested in this study. In a pseudo-diffuse noise scenario, MWF was the only algorithm that provided a significant SRT improvement relative to the unprocessed condition. The BSS algorithm seemed optimally suited for the single point-source

scenario only. Unlike the lack of improvement in SRT for SC1, SC2 and COH, the PR did measure increased preference for these algorithms compared to the unprocessed condition at all tested SNR levels and the LES showed decreased listening effort at 0 dB SNR. This study shows that perceptual evaluations of signal enhancement algorithms can certainly be obtained in a precise, robust, reproducible and comparable manner across different research groups and subject groups when an appropriate test protocol is used.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Sofie Jansen, Niklas Harlander, Elja Schaaf, Andrea Kegel, Immo Ress, Melanie Zokoll, Müge Kaya, and Jessica Beeken for their help in collecting the data. This work was supported by grants from the European Union FP6, Project 004171 HEARCOM. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

REFERENCES

Aichner, R., Buchner, H., and Kellermann, W. (**2007**). "Exploiting narrowband efficiency for broadband convolutive blind source separation," EURASIP J. Adv. Signal Process. Article ID 16381, 9 pages. doi:10.1155/2007/16381.

Aichner, R., Buchner, H., Yan, F., and Kellermann, W. (**2006**). "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," Signal Process. **86**, 1260-1277.

Alcantara, J. I., Moore, B. C. J., Kühnel, V. and Launer, S. (**2003**). "Evaluation of the noise reduction system in a commercial digital hearing aid," Int. J. Audiol. **42**, 34-42.

ANSI S3.5-1997. (**1997**). "American National Standard: Methods for calculation of the Speech Intelligibility Index" (American National Standards Institute, New York).

Bentler, R., Wu, Y. -H., Kettel, J., and Hurtig, R. (**2008**). "Digital noise reduction: outcomes from laboratory and field studies," Int. J. Audiol. **47**, 447-460.

Blamey, P. J. (**2005**). "Adaptive dynamic range optimization (ADRO): a digital amplification strategy for hearing aids and cochlear implants," Trends Amplif. **9**, 77-98.

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**, 2801-2810.

Buchner, H., Aichner, R., and Kellermann, W. (**2004**). "Blind Source Separation for Convolutive Mixtures: A Unified Treatment," in *Audio Signal Processing for Next-Generation Multimedia Communication*, edited by Y. Huang and J. Benesty (Kluwer Academic Publishers), pp. 255-293.

Buchner, H., Aichner, R., and Kellermann, W. (**2005a**). "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," IEEE Trans. Speech Audio Process. **13**, 120-134.

Buchner, H., Aichner, R., Stenglein, J., Teutsch, H., and Kellermann, W. (**2005b**). "Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, Pennsylvania.

Byrne, D., Parkinson, A., and Newall, P. (**1991**). "Modified Hearing Aid Selection Procedures for Severe/profound Hearing Losses," in *The Vanderbilt Hearing Aid Report II*, edited by G. Studebaker, F. Bess, and L. Beck (York Press, Parkton), pp. 295-300.

Dahlquist, M. and Leijon, A. (**2003**). "Paired-comparison rating of sound quality using MAP parameter estimation for data analysis," in Proceedings of Tutorial and Research Workshop on Auditory Quality of Systems (AQS), Mont Cenis, Germany, pp.79-84.

Dillon, H. (**2001**). *Hearing Aids* (Boomerang Press, Sydney).

Doclo, S., Spriet, A., Moonen, M., and Wouters, J. (**2007**). "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," Speech Commun. **49**, 636-656.

Doclo, S., Spriet, A., Wouters, J., and Moonen, M. (**2005**). "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech enhancement*, edited by J. Benesty, S. Makino, and J. Chen (Springer), pp.199-228.

Eneman, K., Leijon, A., Doclo, S., Spriet, A., Moonen, M., and Wouters, J. (**2008a**). "Auditory-profile-based Physical Evaluation of Multimicrophone Noise Reduction Techniques in Hearing Instruments," in *Advances in Digital Speech Transmission*, edited by R. Martin, U. Heute, and C. Antweiler (John Wiley & Sons), pp. 431–458.

Eneman, K., Luts, H., Wouters, J., Büchler, M., Dillier, N., Dreschler, W., Froehlich, M., Grimm, G., Hohmann, V., Houden, R., Leijon, A., Lombard, A., Mauler, D., Moonen, M., Puder, H., Schulte, M., Spriet, A., and Vormann, M. (**2008b**). "Evaluation of Signal Enhancement Algorithms for Hearing Instruments," in Proceedings of the European Signal Processing Conference (EUSIPCO), Lausanne, Switzerland, pp. 1-5.

Ephraim, Y., and Malah, D. (**1984**). "Speech Enhancement Using A Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," IEEE Trans. Acoust. Speech Signal Process. **32**, 1109-1121.

Francart, T., van Wieringen, A., and Wouters, J. (**2008**). "APEX 3: a multi-purpose test platform for auditory psychophysical experiments," J. Neurosci. Meth. **172**, 283-293.

Greenberg J. E., Peterson, P. M., and Zurek, P. M. (**1993**). "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," J. Acoust. Soc. Am. **94**, 3009-3010.

Grimm, G., Guilmin, G., Poppen, F., Vlaming, M. S. M. G., and Hohmann, V. (**2009**). "The Personal Hearing System - a Software Hearing Aid for a Personal Communication System," Submitted to EURASIP J. Adv. Signal Process.

Grimm, G., Herzke, T., Berg, D., and Hohmann, V. (**2006**). "The Master Hearing Aid: A PC-based platform for algorithm development and evaluation," Acta Acust. United with Acust. **92**, 618-628.

Hu, Y., and Loizou, P. C. (**2008**). "A new sound coding strategy for suppressing noise in cochlear implants," J. Acoust. Soc. Am. **124**, 498-509.

Kochkin, S. (**2005**). "MarkeTrak VII: Customer satisfaction with hearing instruments in the digital age," Hear. J. **58**, 30-43.

Martin, R. (**2001**). "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process. **9**, 504-512.

Marzinzik, M. and Kollmeier, B. (**1999**). "Development and Evaluation of Single-Microphone Noise Reduction Algorithms for Digital Hearing Aids," in *Psychophysics,*

Physiology and Models of Hearing, edited by T. Dau, V. Hohmann, and B. Kollmeier (World Scientific, Singapore), pp. 279-282.

Mauler, D., and Martin, R. (**2006**). "Noise power spectral density estimation on highly correlated data," in Proceedings International Workshop on Acoustic Echo and Noise Control (IWAENC), Paris, France, pp. 1-4.

Natarajan, A., Hansen, J. H. L., Arehart, K. H., and Rossi-Katz, J. (**2005**). "An auditory-masking-threshold-based noise suppression algorithm GMMSE-AMT[ERB] for listeners with sensorineural hearing loss," EURASIP J. Appl. Sign. Proc. **18**, 2938-2953.

Noble, W., and Gatehouse, S. (**2006**). "Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the Speech, Spatial, and Qualities of Hearing Scale (SSQ)," Int. J. Audiol. **45**, 172-181.

Ricketts, T. A., and Hornsby, B. W. Y. (**2005**). "Sound quality measures for speech in noise through a commercial hearing aid implementing digital noise reduction," J. Am. Acad. Audiol. **16**, 270-277.

Samuelsson, J. (**2006**). "Single-channel noise reduction with short delay," Master's thesis, KTH Sound and Image Processing, Stockholm.

Spriet, A., Moonen, M., and Wouters, J. (**2004**). "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," Signal Process. **84**, 2367-2387.

Spriet, A., Moonen, M., and Wouters, J. (**2005**). "Stochastic gradient-based implementation of Spatially Preprocessed Speech Distortion Weighted Multichannel

Wiener filtering for noise reduction in hearing aids," IEEE Trans. Signal Process. **53**, 911-925.

Spriet, A., Van Deun, L., Eftaxiadis, K., Laneau, J., Moonen, M., van Dijk, B., van Wieringen, A., and Wouters, J. (**2007**). "Speech understanding in background noise with the two-microphone adaptive beamformer BEAM (TM) in the nucleus Freedom (TM) cochlear implant system," Ear Hear. **28**, 62-72.

Versfeld, N., Daalder, L., Festen, J. M., and Houtgast, T. (**2000**). "Extension of sentence materials for the measurement of the speech reception threshold," J. Acoust. Soc. Am. **107**, 1671-1684.

Walden, B., Surr, R, Cord, M., Edwards, B., and Olson, L. (**2000**). "Comparison of benefits provided by different hearing aid technologies," J. Am. Acad. Audiol. **11**, 540-560.

Wagener, K., Brand, T., and Kollmeier, B. (**1999**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache. Teil III: Evaluation des Oldenburger Satztests" ("Development and evaluation of a sentence test for the German language. Part 3: Evaluation of the Oldenburg sentence test"), Z. Audiol. **39**, 86-95.

Wittkop, T., and Hohmann, V. (**2003**). "Strategy-selective noise reduction for binaural digital hearing aids," Speech Commun. **39**, 111-138.

TABLE I. Performance evaluation of the different signal enhancement approaches when implemented on the PHS system.

Algorithm	Number of microphones used	% CPU time required	Input/output delay
SC1	2 x 1 double monaural	8.3%	10.8 ms
SC2	2 x 1 double monaural	4.2%	16.8 ms
BSS	2 binaural	59.9%	10.8 ms
MWF	2 x 3 double monaural	4.3%	13.2 ms
COH	2 binaural	1.2%	10.6 ms

TABLE II. Description of the subject groups.

Group	**N**	**Age (years)**		**Degree (dBHL)**		**Slope (dB)**		**Symmetry (dB)**	
		Average	Range	Average	Range	Average	Range	Average	Range
NH	38	30	16-52	4	-3-12	9	0-18	4	0-9
HI-F	34	62	22-79	49	34-65	17	5-25	5	1-11
HI-S	37	68	51-80	45	34-61	44	28-70	6	1-19

TABLE III. Overview table indicating which partners performed which SRT test conditions.

Test Condition			**Algorithm**					
Room	**Class**	**Masker Azimuth**	**Unpr**	**SC1**	**SC2**	**BSS**	**MWF**	**COH**
Office-like room	Pseudo-diffuse	90/180/270°	All	All	All	All	All	All
Office-like room	Point-source	90°	BE CH NL	-	-	BE CH NL	BE CH NL	-
Reverberating room	Diffuse	90/180/270°	BE DE	BE DE	BE DE	BE DE	BE DE	BE DE

TABLE IV. Three grouping definitions to divide the subjects in four groups based on 1) their degree of hearing loss, 2) the slope of their audiogram and 3) their SRT in noise for the unprocessed condition.

Group	**Degree**	**Slope**	**SRT in noise**
1	NH, $x \leq 20$ dBHL n=38	NH subjects n=38	$x \leq -4$ dB SNR n=28
2	$20 < x \leq 40$ dBHL n=13	$x \leq 20$ dB n=25	$-4 < x \leq -1$ dB SNR n=39
3	$40 < x \leq 50$ dBHL n=33	$20 < x \leq 40$ dB n=24	$-1 < x \leq 2$ dB SNR n=21
4	$50 < x \leq 65$ dBHL n=25	$x > 40$ dB n=22	$x > 2$ dB SNR n=21

Figure captions

Figure 1: Mean SRT improvements (and standard deviations) relative to the unprocessed condition in an office-like room with three noise sources. Results are shown for the different test sites averaged over all subject groups.

Figure 2: Mean SRT improvements (and standard deviations) relative to the unprocessed condition in an office-like room with three noise sources. Results are shown for the three subject groups averaged over all test sites.

Figure 3: Comparison of results in two noise scenarios. Mean SRT improvements (and standard deviations) relative to the unprocessed condition in an office-like room. Results are averaged over three subject groups and three test sites.

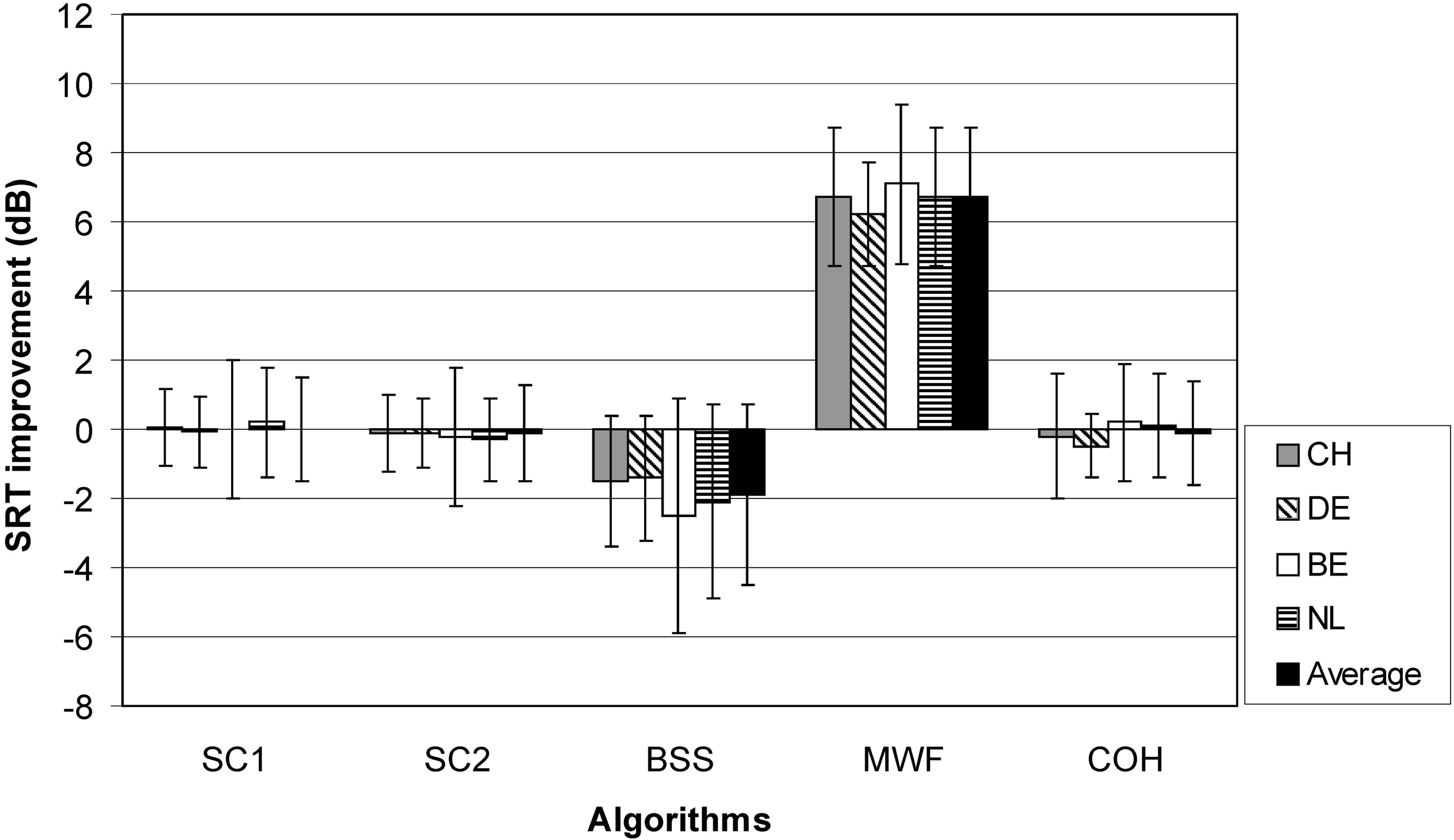
Figure 4: Mean SRT improvements (and standard deviations) relative to the unprocessed condition in an office-like room with three noise sources. Results are shown for subject groups, divided based on the SRT in noise (see Table IV).

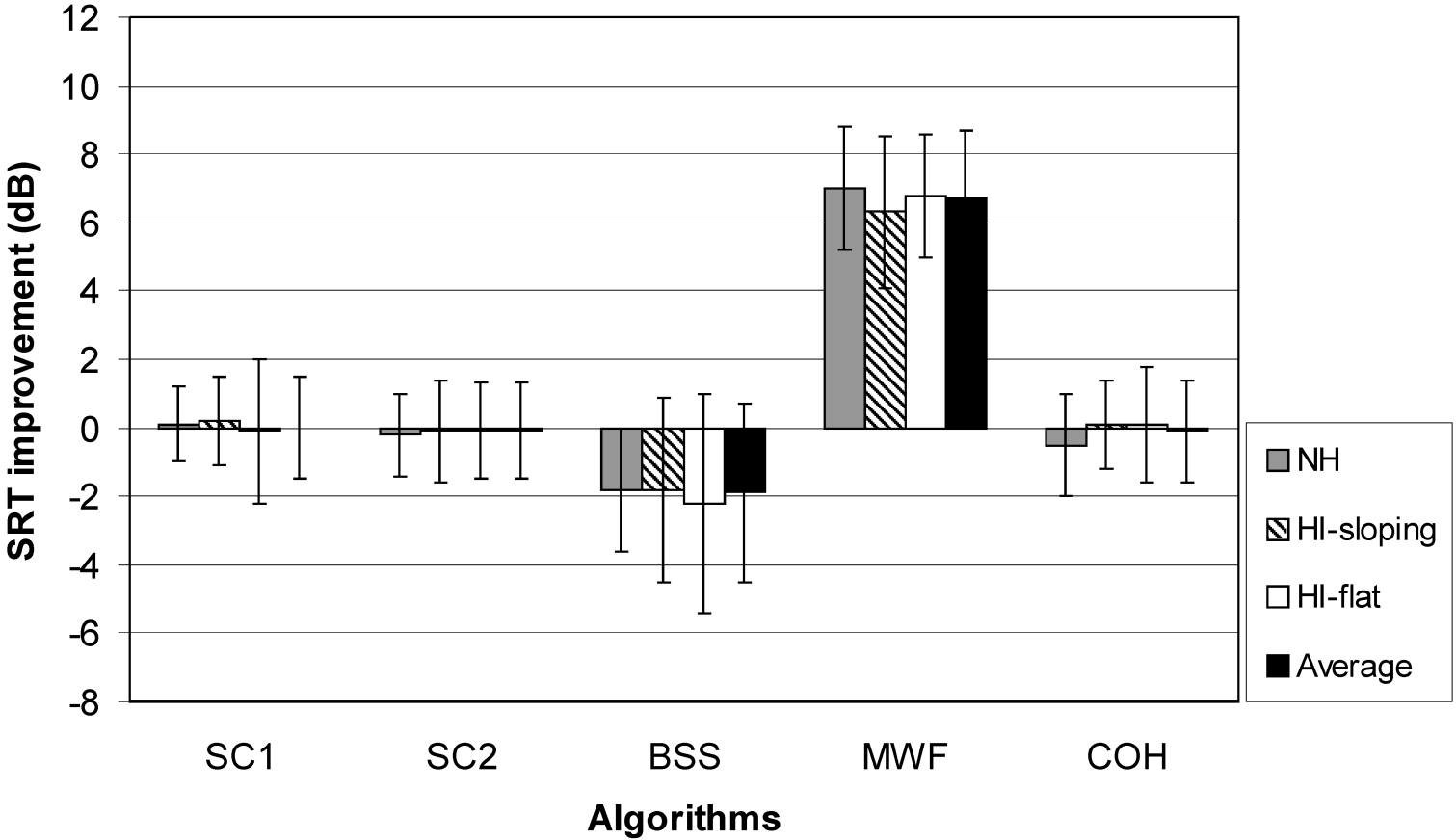
Figure 5: Median LES for the unprocessed condition at different signal-to-noise ratios. The effect of test site is depicted in the left panel and subject group in the right panel. Error bars represent first and third quartile.

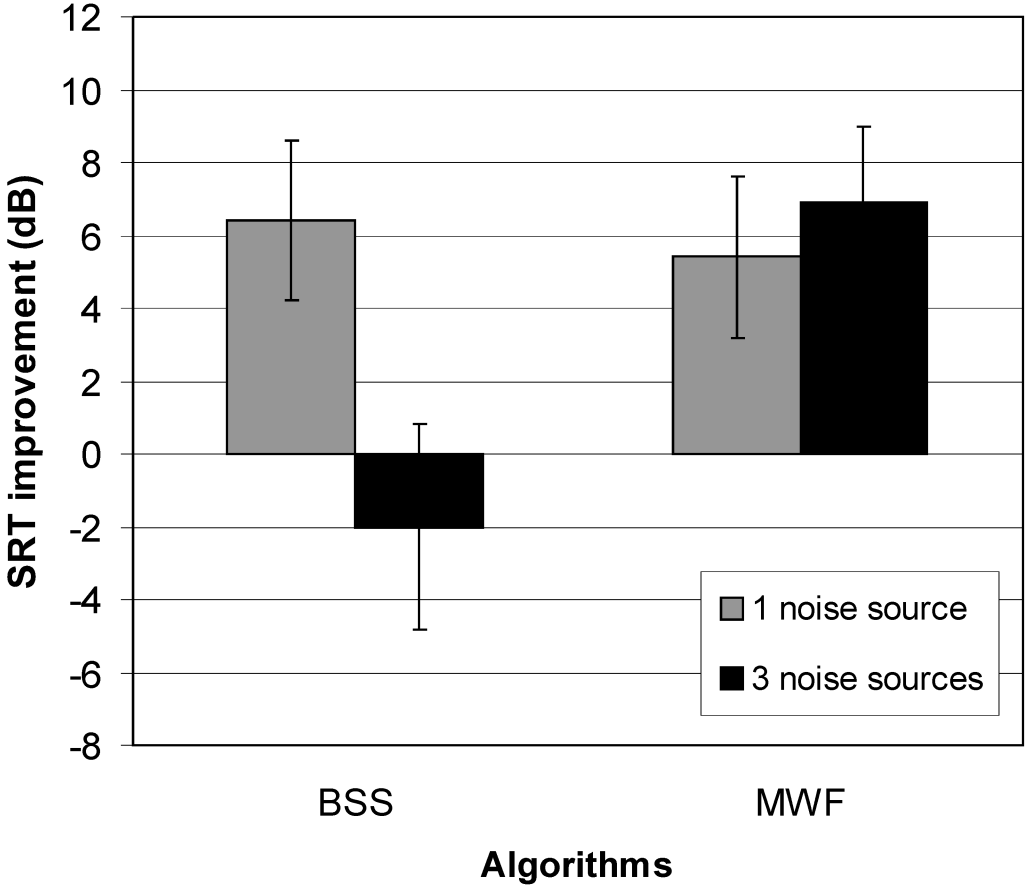
Figure 6: LES relative to the unprocessed condition for 0 dB SNR. The error bars show the first and third quartile. Algorithm scores that are significantly different from the unprocessed condition are indicated with an asterisk.

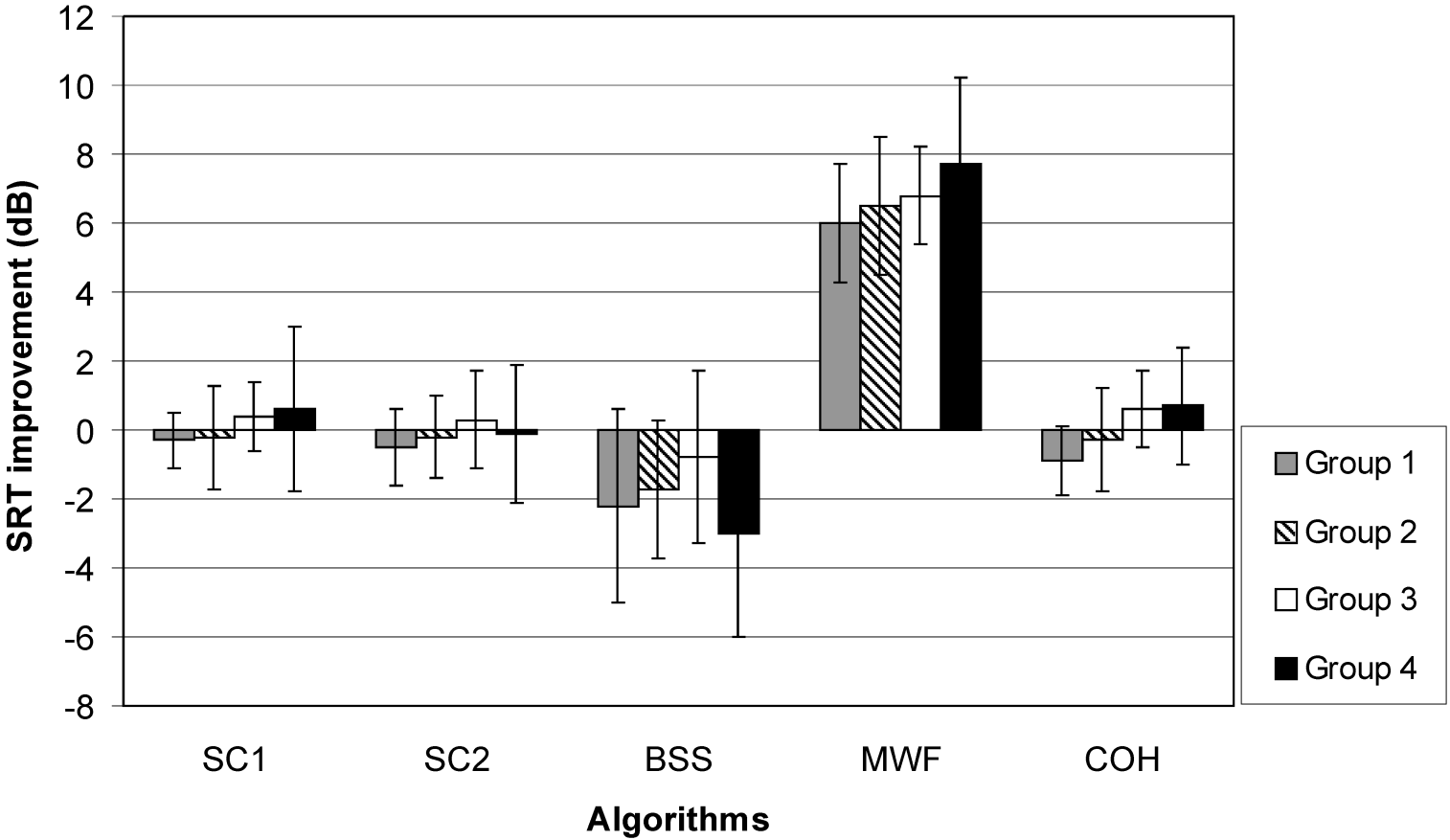
Figure 7: Preference rating results expressed as the percentage of wins for the algorithms compared to the unprocessed condition at three signal-to-noise ratios.

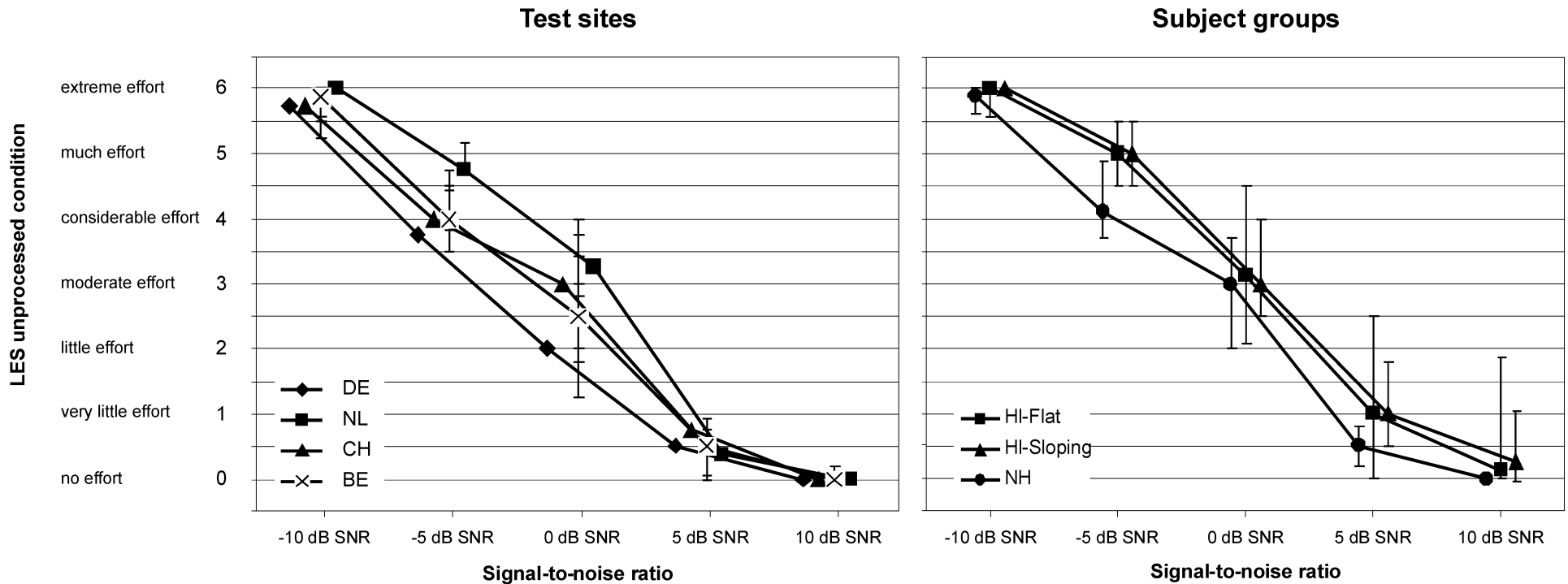
Figure 8: PR scores for the different SNR levels averaged over the test sites and subject groups. The medians (for the rating of the algorithms and the estimated confidence intervals) of the individual (subject-)data, transformed with a linear Gaussian model are plotted with bars. The offset along the LGM-scale is adjusted to zero for the unprocessed condition for every subject individually. Additionally, the categories are indicated.

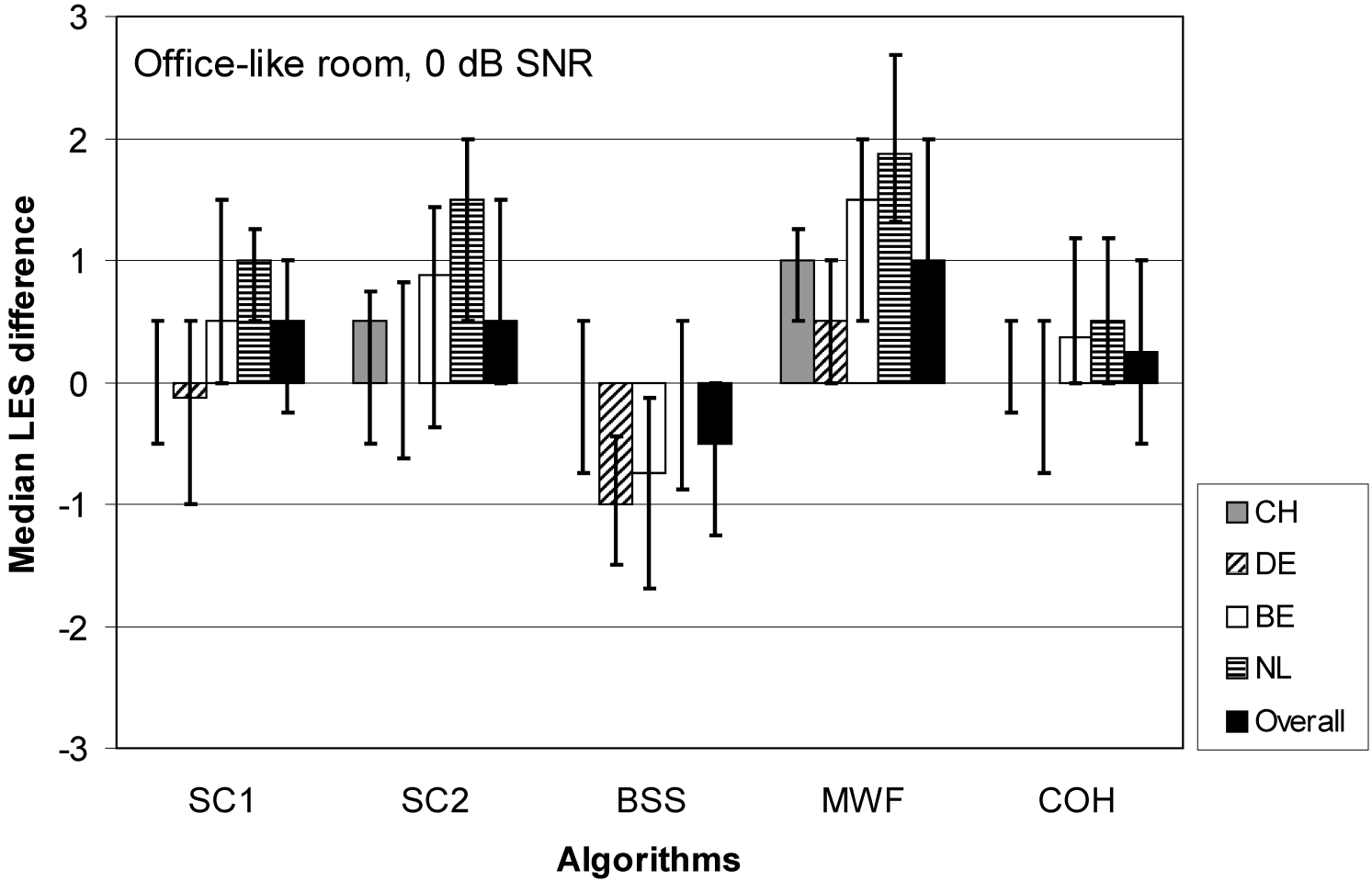

12
10
8
6
4
2
0
-2
-4
-6
-8
SRT improvement (dB)
SC1
SC2
BSS
MWF
COH
Algorithms
CH
DE
BE
NL
Average

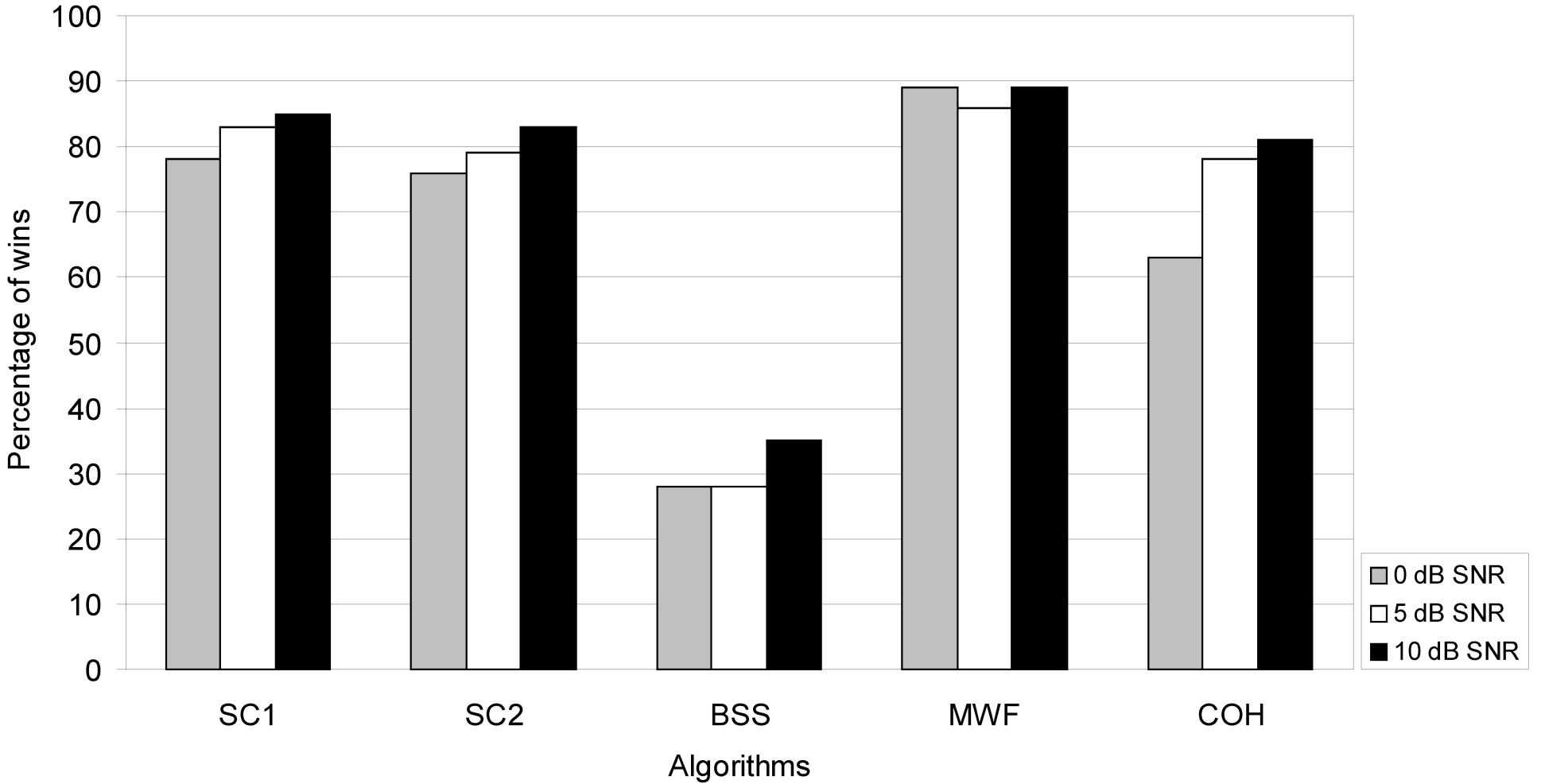

12
10
8
6
4
2
0
-2
-4
-6
-8
SRT improvement (dB)
SC1
SC2
BSS
MWF
COH
Algorithms
NH
HI-sloping
HI-flat
Average

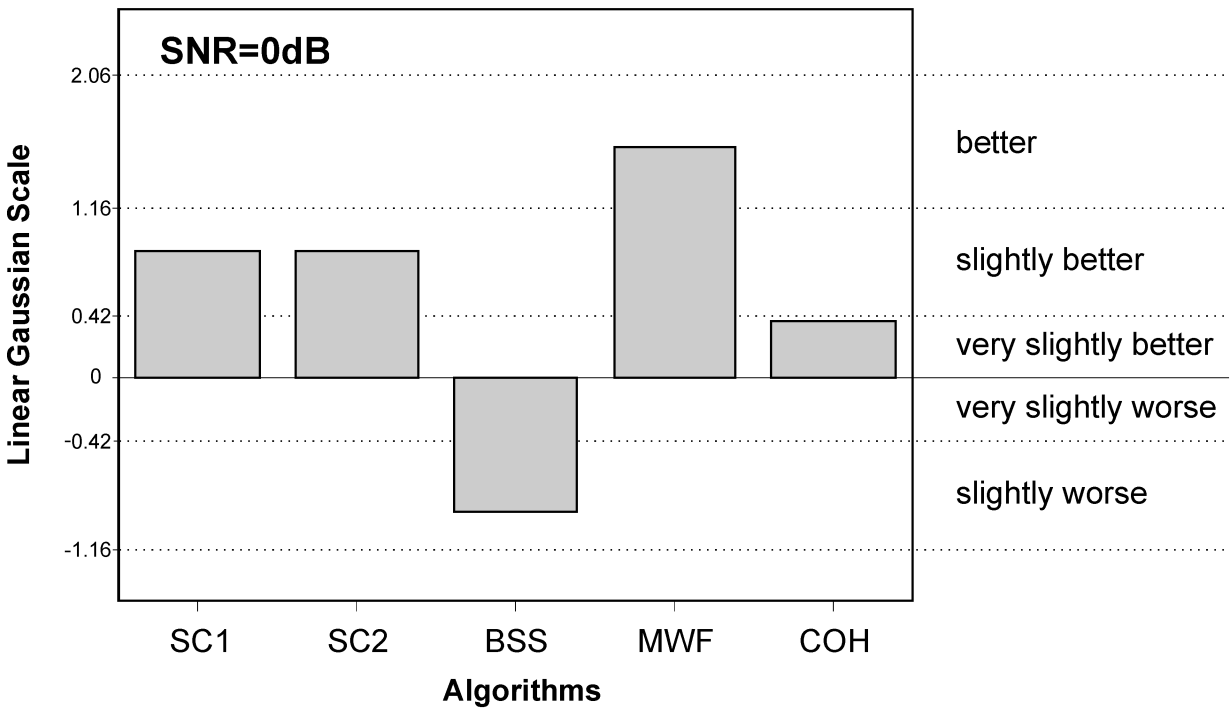

12
10
8
6
4
2
0
-2
-4
-6
-8
SRT improvement (dB)
1 noise source
3 noise sources
BSS
MWF
Algorithms

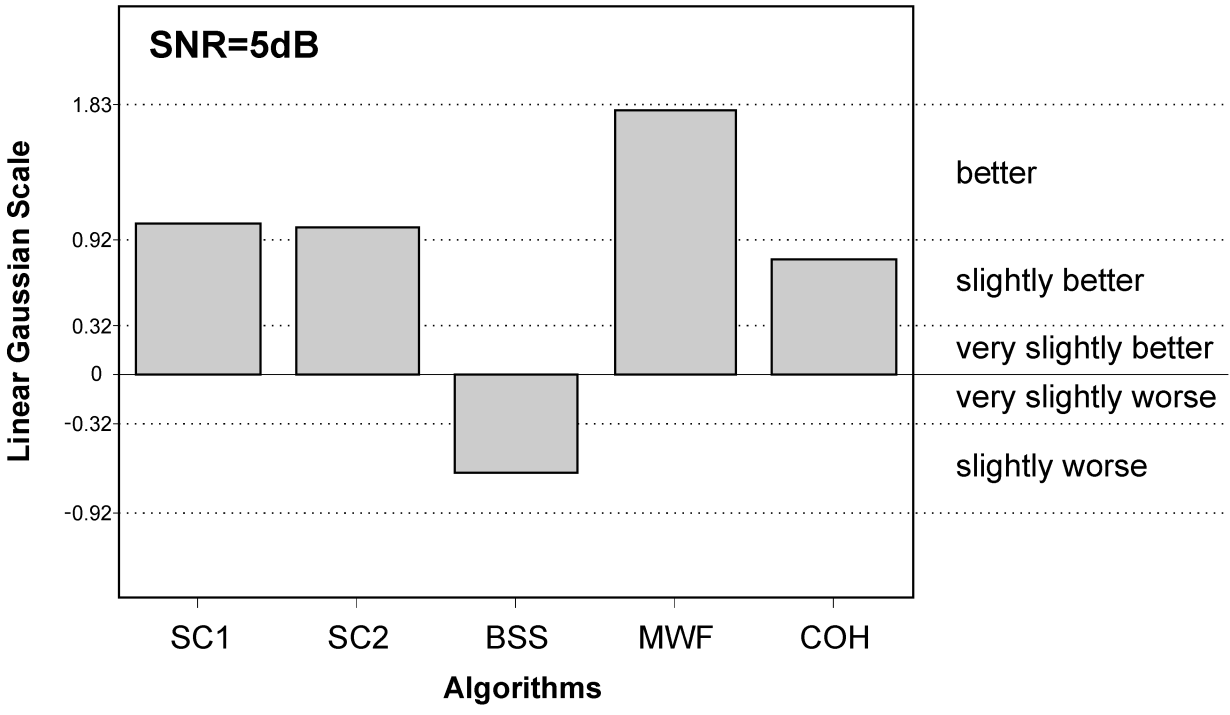

12
10
8
6
4
2
0
-2
-4
-6
-8
SRT improvement (dB)
SC1
SC2
BSS
MWF
COH
Algorithms
Group 1
Group 2
Group 3
Group 4

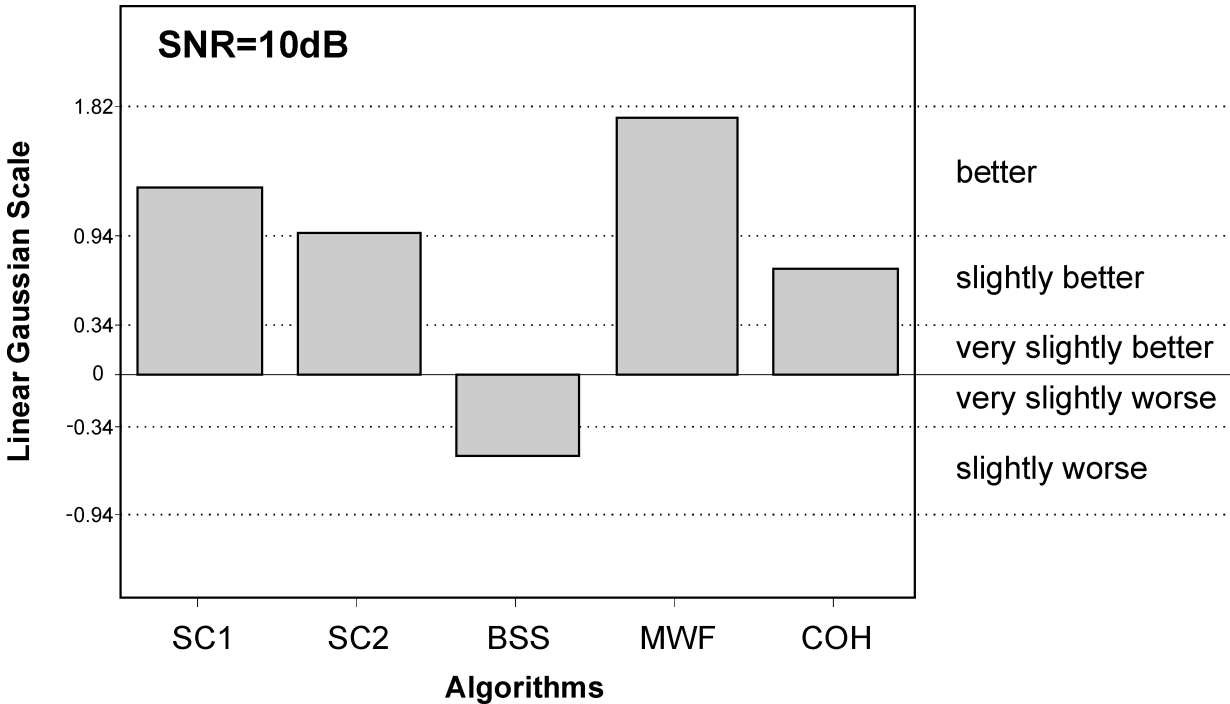

Test sites
Subject groups
LES unprocessed condition
extreme effort
much effort
considerable effort
moderate effort
little effort
very little effort
no effort
6
5
4
3
2
1
0
DE
NL
CH
BE
HI-Flat
HI-Sloping
NH
-10 dB SNR
-5 dB SNR
0 dB SNR
5 dB SNR
10 dB SNR
Signal-to-noise ratio


Office-like room, 0 dB SNR
Median LES difference
3
2
1
0
-1
-2
-3
SC1
SC2
BSS
MWF
COH
Algorithms
CH
DE
BE
NL
Overall


100
90
80
70
60
50
40
30
20
10
0
Percentage of wins
SC1
SC2
BSS
MWF
COH
Algorithms
0 dB SNR
5 dB SNR
10 dB SNR


SNR=0dB
2.06
1.16
0.42
0
-0.42
-1.16
Linear Gaussian Scale
better
slightly better
very slightly better
very slightly worse
slightly worse
SC1
SC2
BSS
MWF
COH
Algorithms


SNR=5dB
1.83
0.92
0.32
0
-0.32
-0.92
Linear Gaussian Scale
better
slightly better
very slightly better
very slightly worse
slightly worse
SC1
SC2
BSS
MWF
COH
Algorithms


SNR=10dB
1.82
0.94
0.34
0
-0.34
-0.94
Linear Gaussian Scale
better
slightly better
very slightly better
very slightly worse
slightly worse
SC1
SC2
BSS
MWF
COH
Algorithms