Comparison of predictive measures of speech recognition after noise reduction processing

Karolina Smeds,[a] Arne Leijon,[b] Florian Wolters, Anders Hammarstedt, Sara Båsjö, and Sofia Hertzman
ORCA Europe, Widex A/S, Maria Bangata 4, SE-118 63 Stockholm, Sweden

(Received 7 December 2013; revised 12 July 2014; accepted 15 July 2014)

A number of measures were evaluated with regard to their ability to predict the speech-recognition benefit of single-channel noise reduction (NR) processing. Three NR algorithms and a reference condition were used in the evaluation. Twenty listeners with impaired hearing and ten listeners with normal hearing participated in a blinded laboratory study. An adaptive speech test was used. The speech test produces results in terms of signal-to-noise ratios that correspond to equal speech recognition performance (in this case 80% correct) with and without the NR algorithms. This facilitates a direct comparison between predicted and experimentally measured effects of noise reduction algorithms on speech recognition. The experimental results were used to evaluate nine different predictive measures, one in two variants. The best predictions were found with the Coherence Speech Intelligibility Index (CSII) [Kates and Arehart (2005), J. Acoust. Soc. Am. **117**(4), 2224–2237]. In general, measures using correlation between the clean speech and the processed noisy speech, as well as other measures that are based on short-time analysis of speech and noise, seemed most promising © *2014 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.4892766]

PACS number(s): 43.71.Gv, 43.71.Ky, 43.66.Ts, 43.72.Dv [ICB] Pages: 1363–1374

I. INTRODUCTION

Most hearing impairments reduce the ability to understand speech in background noise (e.g., Moore, 1996). Hearing-aid users often need to cope with noisy listening situations (Wagener *et al.*, 2008), and difficulties hearing in these situations constitute a major source of dissatisfaction (Kochkin, 2010). Therefore manufacturers of hearing devices are trying to find improved noise-reduction (NR) methods, which need to be evaluated for their effects on speech recognition.

The purpose of the present study was to evaluate a number of measures in terms of their capacity to predict the effect that NR algorithms have on listeners' speech recognition ability in noise. The work reported here was part of a larger study, where both speech intelligibility and sound quality of NR processed speech were evaluated. The sound-quality work has been reported elsewhere (Smeds *et al.*, 2010b).

A number of strategies can be used in hearing instruments to improve the signal-to-noise ratio (SNR). The present study is concerned with effects of so-called "single-channel" NR, commonly used in modern hearing aids, by itself or in combination with microphone arrays for spatial beam-forming. These single-channel NR algorithms can be designed based on various rationales. Previous studies have shown that hearing-aid NR algorithms function in very different ways (Chung, 2004; Hoetink *et al.*, 2009; Bentler, 2006; Smeds *et al.*, 2010a; Brons *et al.*, 2013).

Although improved speech recognition in noise is an appropriate NR design goal, it has been difficult to find evidence for NR algorithms that can actually achieve this goal (Bentler *et al.*, 2008; Luts *et al.*, 2010; Hu and Loizou, 2007). One exception was found in a study where Peeters *et al.* (2009) showed improved speech recognition for listeners with impaired hearing.

A number of studies show that hearing-aid users prefer to use NR in noisy situations (e.g., Boymans and Dreschler, 2000; Luts *et al.*, 2010) and that NR algorithms can improve listening comfort (Bentler *et al.*, 2008). Even when an NR algorithm is designed mainly to improve listening comfort, it is important to evaluate its effects on speech recognition. At least one study has shown that listeners with impaired hearing subjectively judged speech clarity to be better with NR even if measured speech recognition actually decreased with NR processing (Dahlquist *et al.*, 2005).

The effects of NR algorithms are usually evaluated in listening tests with participants with or without hearing impairment (e.g., Bentler *et al.*, 2008; Luts *et al.*, 2010; Hu and Loizou, 2007). It would be of great value if some predictive measure could be used to indicate the effect of various NR algorithms prior to laboratory or field testing with listeners.

Methods to predict speech intelligibility, using measures of speech and noise characteristics, have been important design tools ever since the early development of telephone communications (Fletcher, 1929; Fletcher and Galt, 1950). In the telephone-system application, the most important forms of transmission distortion were stationary noise and a non-uniform linear frequency response with a very limited frequency bandwidth. The Articulation Index, later developed into the Speech Intelligibility Index (SII) (ANSI, 1997)

[a] Author to whom correspondence should be addressed. Electronic mail: Karolina.Smeds@orca-eu.info
[b] Also at: KTH School of Electrical Engineering, Stockholm, Sweden.

turned out to be a useful predictive measure. More recently, the SII has also been used for theoretical prescription of gain characteristics of non-linear hearing instruments (Byrne *et al.*, 2001; Keidser *et al.*, 2011). Extensions of the SII and several new predictive measures have been proposed for use in situations with highly modulated noise and with advanced forms of non-linear signal processing.

Some approaches are related to the Speech Transmission Index (STI), originally proposed by Houtgast and Steeneken (1973) and later standardized (IEC, 2011). The STI is focused on the degree with which the temporal modulation characteristics of the clean speech are preserved in the presented noisy speech. The STI was designed to account better than the SII for the effects of modulated noise, reverberation, and some forms of non-linear signal transmission like peak-clipping and automatic gain control (Steeneken and Houtgast, 1980).

When a predictive measure of speech recognition is used for preliminary evaluations of a new or modified NR algorithm for hearing aids, the most important concern is whether the measure can predict *small changes in the right direction*. If the predictive measure indicates an improvement with a new NR algorithm, but tests with listeners show no improvement, or even a reduction in speech-recognition performance, then obviously the preliminary predictive evaluation might guide algorithm development in the wrong direction.

Several previous studies have evaluated the ability of predictive measures to estimate listeners' performance with NR processed speech (e.g., Ma *et al.*, 2009; Taal *et al.*, 2011b; Xia *et al.*, 2012). These studies present deviations and correlations between predicted and measured intelligibility scores across a wide range of noise types and SNRs and processing algorithms. However, as the overall prediction errors are pooled across NR algorithms, it is difficult to see if the predictive measures can correctly reveal small differences between NR algorithms. In fact, in some cases the presented scatter plots (Taal *et al.*, 2011b; Xia *et al.*, 2012) suggest that some predictive measures may give systematically different errors for different NR algorithms.

In the studies mentioned in the preceding text, only normal-hearing listeners participated. Predictive measures intended to be used to evaluate NR algorithms in hearing aids need to be evaluated using participants with impaired hearing.

The present evaluation included listeners with impaired and normal hearing. The study was designed to test various predictive measures in listening conditions with speech-to-noise ratios individually adjusted to give equal speech-recognition performance across different NR methods. This approach is sensitive in revealing the ability of the predictive measure to correctly indicate also small differences between test conditions.

II. LABORATORY TEST

A. Method

Twenty listeners with sensorineural hearing loss and ten listeners with normal hearing participated in a laboratory test where speech recognition was tested with sound files that were either unprocessed or pre-processed using three software-based NR algorithms. For the participants with impaired hearing, linear hearing aids were individually fitted to compensate for each participant's hearing loss. The test results were SNRs corresponding to equal speech recognition performance for all test conditions (unprocessed and three NR algorithms).

1. Participants

Twenty listeners with symmetrical, sensorineural, mild-to-moderate hearing loss (Fig. 1, left panel), 11 women and 9 men, were recruited from a research database at ORCA Europe. Symmetrical hearing was defined as a maximum threshold difference between the ears of 15 dB at a maximum of three adjacent audiometric test frequencies (including 1.5, 3, and 6 kHz). In the group, two participants had a threshold difference of 15 dB at two adjacent frequencies and for more than half of the participants the threshold differences were less than 15 dB at all test frequencies. The participants' ages ranged from 62 to 82 yr (mean, 72 yr), and they were all fluent in Swedish.

The participants were all experienced users of binaural hearing aids. All but one participant had more than 1 yr of hearing-aid experience (median, 3.5 yr). The person with shorter experience had used the hearing aids regularly for

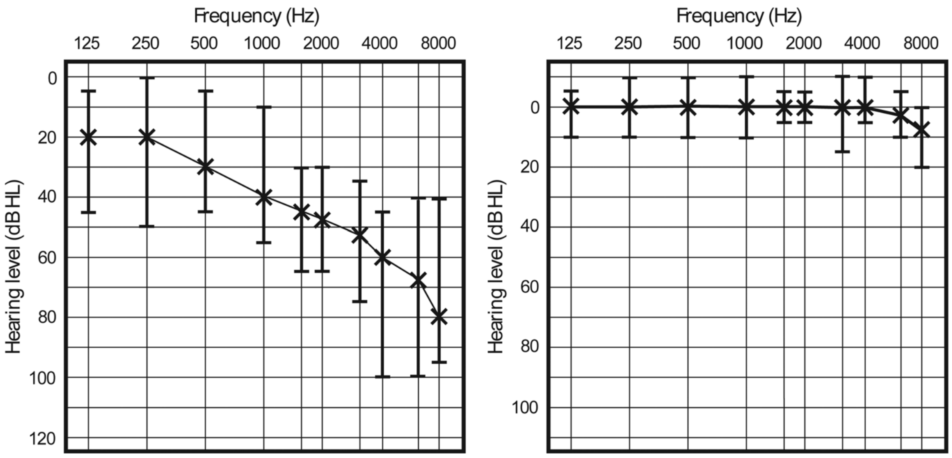

FIG. 1. Median thresholds (crosses) and range of hearing thresholds (bars) for 20 listeners (40 ears) with impaired hearing (left panel) and 10 listeners (20 ears) with normal hearing (right panel).

more than 6 months. The participants were not paid for their participation, but they received a small gift at the last visit.

Later ten listeners with normal hearing (Fig. 1, right panel), six women and four men, were recruited by advertising at the Stockholm University. Their ages ranged from 19 to 28 yr (mean, 23 yr), and they were all fluent in Swedish. They were paid for their participation.

The study design agreed with the Declaration of Helsinki in which ethical principles for human medical experiments are outlined. All participants were given written and verbal information about the study, and they signed an informed consent form. Participation was voluntary and could be terminated if a person would decide to do so. For the participants with impaired hearing, commonly available CE marked hearing aids were fitted according to a conservative clinical practice.

2. Hearing aids

High-quality hearing aids (Inteo 9, Widex A/S), linearly programmed according to the NAL-R prescription (Byrne and Dillon, 1986), but with gain reduced by 6 dB across the frequency range, were fitted bilaterally to the participants with hearing impairment. The gain reduction was motivated by binaural loudness summation, by the fact that the speech presentation levels used in this study were slightly higher than normal, and by the fact that a number of studies have indicated that NAL-R prescribes more gain than listeners prefer (e.g., Humes *et al.*, 2002; Leijon *et al.*, 1990). The hearing aids were used during the laboratory experiment without any prior gain acclimatization.

All advanced signal processing in the hearing aids was switched off. The hearing aids were used with tight ear molds. The hearing-aid fittings were verified using real-ear insertion gain measurements (REM, INTERACOUSTICS EQUINOX SUITE 2.03) and the linearity of the programmed hearing aids was confirmed using coupler-gain measurements (INTERACOUSTICS EQUINOX SUITE 2.03) at a wide range of input levels.

3. Noise reduction algorithms

Three software-implemented NR algorithms were used. All methods work by fast-varying adaptive adjustments of the gain frequency response. The short-time spectra of signal and noise are estimated and the algorithms reduce the gain in time-frequency bins where the SNR is low. The algorithms differ in their SNR estimation methods and in the speed and range of gain adjustments.

The purpose of this study was not to evaluate any particular NR algorithm. The algorithms were selected primarily because they produced clearly different perceptual qualities evaluated by normal-hearing listeners in an informal listening test.

Two of the algorithms are described in the textbook by Loizou (2007), and the MATLAB code on the CD attached with the textbook was used. The method called "WEDM" uses a Bayesian noise estimator based on the weighted Euclidean distortion measure (function stsa_weuclid.m). The method called "Wiener" applies conventional Wiener filtering based on *a priori* SNR estimation (function wiener_as.m). The third algorithm, perceptually tuned spectral subtraction algorithm with low-pass filtered spectral filter coefficients (PSSLP), was developed and evaluated for hearing-aid use (Luts *et al.*, 2010). A fourth test condition, "unprocessed," with no NR processing, was used for comparison.

To illustrate the long-term effects of the NR processing on speech and noise levels, these levels were measured for an example condition with 0 dB SNR at the input to the algorithms. Compared with the speech and noise levels for the unprocessed condition (0.0, 0.0) dB, all algorithms reduced both speech and noise levels to (−2.0, −2.8) dB with PSSLP, to (−3.1, −7.2) dB with WEDM, and to (−0.9, −3.2) dB with the Wiener processing. Thus WEDM was the most aggressive algorithm, reducing the long-term speech level by 3.1 dB and increasing the long-term SNR by 4.1 dB, but it also resulted in quite audible distortion.

These changes of the long-term speech and noise levels are somewhat different at other SNRs. However, these changes were considered as a built-in consequence of the NR algorithms and were included in all calculations of predicted intelligibility measures.

4. Word recognition test

Speech Recognition Thresholds (SRTs) were measured. In this study, the term SRT denotes the SNR at which a participant reaches a pre-defined performance criterion, here 80% correct. A Swedish adaptive sentence test using five-word sentences with a fixed syntax ("matrix text") spoken by a female talker was used (Hagerman, 1982). Artificial babble noise was derived by superimposing the International Speech Test Signal (ISTS) (Holube *et al.*, 2010) eight times with randomly varying starting points and the levels pairwise decreased by 2, 4, and 6 dB relative to the first pair. The babble noise was then filtered to the long-term average spectrum of the speech sentences. The sentences were mixed with the artificial babble in SNRs from −12 to +15 dB with a step size of 1 dB. These mixed speech and babble files were then processed by the three NR algorithms.

The sound files were presented at a fixed speech level of 70.5 dB(A) re 20 μPa for the unprocessed version. The overall levels of the processed speech were slightly lower as shown in Sec. II A 3. The reported results are the nominal SNRs used, i.e., the SNR at the input to the NR algorithms.

The adaptive speech testing started with a training list (10 sentences), which was used to familiarize the participants with the speech material. The adaptive procedure was designed so that the participants would be close to 80% correctly identified words at the end of the training list. During the actual testing, the SNR was kept un-changed if the participant recognized four of five words correctly. The SNR was decreased by 1 dB if the participant recognized five words. The SNR was increased by 1 dB if the participant recognized two or three words and by 2 dB if the participant recognized no or one word. The test was terminated when one of the following two conditions was fulfilled: (1) After seven reversals in the up-down procedure (and the result was then calculated as the mean of the SNRs at the last four reversals) or (2) if the participant achieved 80% correctly identified words for seven consecutive sentences (in which case this SNR was used as the result). All participants reached one of these stop criteria within three test lists (30 sentences). Data were collected twice at two visits.

5. Instrumentation and calibration

The speech and noise sound files for the adaptive speech test were stored on a PC and played back with an external 24-bit RME Fireface 800 sound card and power amplifier Rotel type RMB-1075. The listening test was performed in a sound-proof booth (3.2 m × 3.1 m × 2.0 m). The participants listened binaurally under sound-field conditions using one loudspeaker (Jamo D400) placed 1 m in front of the listener. The measured transfer function from the digital signal to the listening position was included in all calculations of predictive measures.

The frequency response of the complete playback system, from the stored audio files, via D-A converter, amplifier and loudspeaker, to the listening position in the test room, was measured by presenting a sound file with white Gaussian noise. The sound at the listening position was recorded by a free-field microphone (Brüel and Kjær 4189) connected to a pre-amplifier (Brüel and Kjær ZC 0032) and the AD converter of a RME Fireface 800 sound card, and the resulting signal was stored as an audio file.

The presented and recorded signals were analyzed in 1/3-octave bands ranging from 100 Hz to 10 kHz. The room frequency response was estimated as the difference (in dB) between the mean power levels in each band of the recorded and presented signals.

For use in the predictive measures, the individual hearing-aid insertion-gain frequency responses for listeners with hearing impairment were analyzed in a similar way and were represented with the same frequency resolution as the room transfer function. The insertion-gain responses were smoothed by power averaging in 1/3-octave frequency bands.

For the absolute sound pressure level calibration (performed daily), a special calibration signal (stationary Gaussian noise with approximately speech-shaped spectrum) was presented in the same way as the test sounds. The A-weighted sound pressure level was measured at the listening position using a Rion NL-32 sound-level meter with microphone NH-21. The measured calibration level was 72 dB(A) re 20 μPa.

All presented sound files were stored digitally with fixed amplitude in relation to the calibration signal. The playback equipment was left in the same state as during the calibration measurements. As shown in Sec. II A 3, the NR algorithms modified the overall amplitude of the digital speech and noise signals. These level effects were included in the subsequent calculations.

6. Statistical analysis

The individual SNR values for 80% word recognition (SRT) were averaged across the two test sessions. Friedman's two-way analysis of variance by ranks test (MATLAB Statistical Toolbox, v. 8.2) was then used to determine whether there were any statistically significant rank-order differences ($p < 0.05$) across NR algorithms, for each of the two groups of listeners.

The overall reliability of the speech test results was quantified as follows: The mean squared test-retest SRT difference, d^2, averaged across listeners and test conditions, was used as a conservative estimate of the variance of the test-retest difference. Because the variance of the sum of two independent measurements is equal to the variance of their difference, the variance of the mean of the two test-retest results is estimated as $d^2/4$, and the corresponding standard deviation is $d/2$. The standard error of group results for N listeners is then estimated as $d/2\sqrt{N}$.

B. Results

The results from the adaptive speech test showed that the participants with impaired hearing (Fig. 2, left panel) achieved 80% correct word recognition at substantially higher SNRs and with greater inter-individual variation than the participants with normal hearing (Fig. 2, right panel). Friedman's two-way analysis of variance by ranks showed that there were significant ($p < 0.05$) differences among the test conditions (three NR algorithms and one unprocessed) within both groups of participants.

The root-mean-square test-retest SRT difference, averaged across listeners and test conditions, was $d = 2.71$ dB-units for the group with impaired hearing and 1.76 dB-units for the normal-hearing listeners. As described in Sec. II A 6, the standard error of a group result in one test condition is estimated as $d/2\sqrt{N}$, i.e., about 0.30 dB-units for $N = 20$ listeners with impaired hearing and about 0.28 dB-units for the $N = 10$ listeners with normal hearing.

The test-retest results indicate that the procedure could reveal quite small systematic group differences between test conditions. This is confirmed by the results of the Friedman test, which also accounts for the random individual variations.

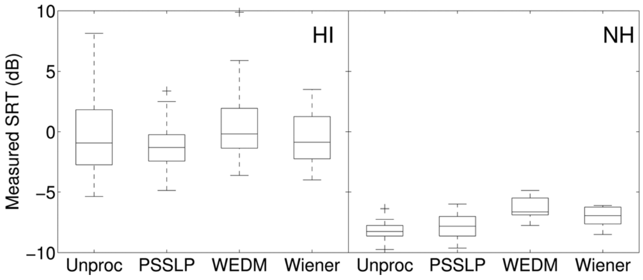

FIG. 2. Speech recognition thresholds (SRT) for 20 listeners with impaired hearing (HI, left panel) and 10 listeners with normal hearing (NH, right panel) for three NR algorithms (PSSLP, WEDM, Wiener) and one reference condition with unprocessed signals (Unproc). The SRT is defined as the SNR required for 80% correct responses in the adaptive speech test. Each individual result is the average across two test sessions. Each box shows inter-quartile values across all listeners, and the median is represented by the line in the box. Outliers (+) are defined as values outside 1.5 times the box length, and the whiskers extend to the highest and lowest values when outliers are excluded.

III. PREDICTIVE MEASURES

Nine different predictive measures were applied (with two variants of one measure). The measures are described in Sec. III A 1 to Sec. III A 9.

All measures apply an initial audio frequency analysis meant to mimic the frequency analysis of the inner ear. The first four methods (SII, ESII, STSII, and Glimpses) use the processed speech and noise signals to calculate a weighted average of the audibility of processed speech across time and frequency.

The other measures analyze other characteristics of the output signal from each audio channel. Two methods (fwSNRseg-a, -b) calculate speech-to-distortion ratios using the processed noisy speech and the corresponding clean speech. Two methods (STOI and CSII) use measures of correlation between the processed noisy speech and the corresponding clean speech. The last two methods (sEPSM and mr-sEPSM) calculate SNRs in the modulation frequency domain, using the envelopes of the processed noisy speech and the processed noise.

A. Method

Speech and noise files with input SNRs ranging from −12 to + 15 dB in 1-dB steps were prepared for each test condition (three NR algorithms and one unprocessed).

Some of the predictive measures need separate speech and noise signals in the calculation. As the WEDM and Wiener algorithms processed the noisy speech signals, the processed output signal was separated into speech and noise files using the phase-inversion method of Hagerman and Olofsson (2004). The PSSLP algorithm was implemented with so-called "shadow filtering," i.e., the code processed separate input speech and noise signals, but the processing of both signals was determined by the internally mixed speech plus noise. Therefore no extra separation step was necessary for this NR algorithm.

In all calculations, the power spectra for the speech and noise signals were adjusted to include the overall level calibration and the room transfer function, measured as described in Sec. II A 5. Individual hearing threshold loss and hearing-aid insertion gain values were used for each ear. For the listeners with normal hearing, the hearing loss and the insertion gain values were set to 0 dB. In this way, all calculations of predictive measures accounted for the absolute sound pressure level and spectrum presented at each ear of each listener.

The predictive measure was calculated separately for each ear, and the highest of the two predicted values was used when displaying the data and in the statistical analyses. The participants had symmetrical hearing, and therefore the differences between predictions for the two ears were small.

Some predictive measures were not designed to take the absolute presentation level and the hearing thresholds into account. To apply these methods for listeners with impaired hearing, the absolute levels and spectra of the presented sounds were related to the absolute hearing thresholds as described later for each measure.

1. SII

The Speech Intelligibility Index (SII) quantifies audibility of speech based on long-term estimates of the speech and noise spectrum levels (spectral power density per hertz in decibels) and the hearing threshold levels (ANSI, 1997). Audibility was determined from the long-term speech and noise spectrum levels as specified in the standard, using critical-band frequency resolution and the band importance function for average speech (ANSI, 1997, Table I).

The threshold-equivalent reference spectrum levels for normal hearing were taken from the critical-band version of the SII standard (ANSI, 1997, Table I). For the listeners with impaired hearing, the hearing threshold loss in decibel hearing level (HL) was interpolated to the critical-band center frequencies and added to the normal reference spectrum levels. A non-standard desensitization factor, proposed by Pavlovic *et al.* (1986), was applied to incorporate suprathreshold deficits associated with sensorineural hearing loss.

2. ESII

Rhebergen *et al.* (2006) and Rhebergen *et al.* (2010) presented an extended Speech Intelligibility Index (ESII) method with the purpose of predicting speech recognition in fluctuating noise. This extension to the SII has shown promising results for various types of fluctuating noise. The ESII is determined from the long-time average speech spectrum together with effective short-time noise spectra, calculated with critical-band frequency resolution. The effective noise spectra are calculated to include the effects of forward masking.

A sequence of short-time SII values are estimated with fine temporal resolution, and these values are averaged to give the final ESII value. The short-time SII values were calculated by the critical-band variant of the standard method, using the band importance function for average speech (ANSI, 1997, Table I). The calculations used MATLAB code kindly made available by Koenrad Rhebergen in Oct. 2013.

The individual hearing thresholds were included in the same way as for the standard SII (Sec. III A 1). No desensitization factor was applied.

3. STSII

As part of the current study, a simple short-time SII version (STSII) was implemented. This method calculates an SII value using the short-time speech and noise spectra estimated in Hamming-weighted time windows of 50 ms with 50% overlap, i.e., 25 ms time resolution. The SII calculation for each time window used the same standard method as in Sec. III A 1. No effects of forward masking were included. The final result was an average of all STSII values. A similar approach was proposed by Kates (1987).

The individual hearing thresholds were included in the same way as for the standard SII (Sec. III A 1), including the non-standard desensitization factor.

4. Glimpses

Cooke (2006) showed good correlation between a "glimpses" measure and consonant recognition in babble noise with normal-hearing listeners. The glimpses measure is the proportion of time-frequency bins where the speech magnitude exceeds the noise magnitude by a predetermined amount. The time-frequency representations are calculated separately for the speech and the noise components in the processed signal.

The spectral analysis was performed by time-domain filtering in a complex Gammatone filterbank with 40 bands with approximately normal auditory equivalent rectangular bandwidths (ERBs), covering the frequency range from 50 Hz to 7.5 kHz. The signal envelopes in each band were smoothed by a first-order low-pass filter with 8-ms time constant and then down-sampled to a time resolution of about 10 ms. This analysis was done using MATLAB software kindly provided by Martin Cooke in Feb. 2011. The final glimpses measure was then calculated simply as the relative number of time-frequency bins where the SNR was greater than 0 dB.

The original glimpses measure did not take any hearing-threshold effects into account. To apply the method for listeners with impaired hearing, the effective noise spectrum levels were limited by a noise floor at the spectrum levels representing the individual absolute hearing threshold. The threshold-equivalent spectrum levels for normal hearing were taken from the critical-band version of the SII standard (ANSI, 1997, Table I), interpolated to the center frequencies of the filters used in the calculation. For the listeners with impaired hearing, the hearing threshold loss (in decibel HL) was interpolated in the same way and added to the normal reference spectrum levels.

5. *fwSNRseg*

In an evaluation of several predictive intelligibility measures for noise-corrupted speech, processed with several NR algorithms, frequency-weighted segmental signal-to-noise ratio (fwSNRseg) approaches came out among the most promising ones (Ma *et al.*, 2009).

The fwSNRseg is a weighted average of speech-to-distortion ratios, in decibels, calculated from the ratio between the time-frequency representations of the unprocessed clean speech and the distortion. The distortion is calculated in each time-frequency bin as the difference between the envelopes of the clean speech and the processed noisy speech. As opposed to the SII-based measures, this distortion measure includes not only the noise but also envelope modifications caused by the NR algorithm.

Ma *et al.* (2009) tried several methods to determine the weights. One of these variants have been used here, and one modification, based on a later suggestion by Loizou and Kim (2011) has also been included

(a) Weight factors proportional to the clean speech time-frequency magnitude, called "$p = 1$" by Ma *et al.* (2009, Eq. 6).

(b) The same weight factors ("$p = 1$") except that the weights were doubled in all time-frequency bins where the magnitude of the processed noisy speech was more than 6 dB higher than the clean-speech magnitude. This level region was called "Region III" by Loizou and Kim (2011).

Loizou and Kim (2011) suggested that NR processing should be constrained so that it does not produce output in "Region III," but they did not interpret this criterion as a predictive intelligibility measure. In the time-frequency bins in "Region III," the signal-to-distortion ratios (in decibels) are always negative, so increasing the weights is equivalent to applying a higher cost for any distortion in this region. Several ways to modify the weights in (b) have been tested in the current study of which the presented variant seemed to be the most promising.

The present implementation calculated the time-frequency signal representations by first segmenting the time-domain signals into 20-ms Hamming-windowed frames with 50% overlap, i.e., 10-ms time resolution. The short-time spectra were calculated in 25 frequency bands with normal auditory ERBs covering the frequency range from 50 Hz to 8 kHz.

The studies by Loizou and co-workers did not take any hearing-threshold effects into account. To apply the methods for listeners with impaired hearing, the spectrum levels of the processed noisy speech were compared to the corresponding spectrum levels representing the individual absolute hearing threshold. The threshold-equivalent reference spectrum levels for normal hearing were taken from the critical-band version of the SII standard (ANSI, 1997, Table I), interpolated to the frequencies used in the calculation. For the listeners with impaired hearing, the hearing threshold loss (in decibel HL) was interpolated in the same way and added to the normal reference spectrum levels. The final weighted sum included only those time-frequency bins where the processed noisy speech levels were higher than the corresponding threshold levels.

6. *STOI*

Taal *et al.* (2011a,b) developed a short-time objective intelligibility measure (STOI) of the correlation between band envelope magnitudes of clean speech and processed noisy speech. The method has shown good agreement with speech recognition results obtained for normal-hearing listeners tested with noisy speech processed by a number of NR algorithms.

A short-time spectral analysis is performed by segmenting the input signals in Hamming-weighted blocks with 25.6 ms duration with 50% overlap, i.e., 12.8 ms time resolution. Short-time band power spectra are calculated for each segment in 15 third-octave bands with center frequencies from 150 Hz to 3.8 kHz. Using the sequence of envelope magnitude values in each frequency band, the linear correlation coefficient between clean speech and processed noisy speech is calculated within overlapping time segments of about 400 ms, after scaling to equalize the power of clean and processed noisy speech and clipping to limit the signal-to-distortion ratio. The correlation coefficients are then averaged across time segments and frequency bands. All these calculations were performed by MATLAB code kindly provided by Cees Taal in Dec. 2009, also available at http://msp.ewi.tudelft.nl/content/software-and-data.

As the original STOI calculation does not take hearing loss into account, the calculations were modified by adding the third-octave band power values of a threshold-equivalent noise to the band power values of the processed noisy speech, before calculating the correlation. The reference spectrum levels of the threshold-equivalent noise for normal hearing was taken from the SII standard (ANSI, 1997, Table I), interpolated to the band center frequencies used in the calculation, and adjusted to third-octave band levels. For the listeners with impaired hearing, the hearing threshold loss (in decibel HL) was interpolated in the same way and added to the normal reference spectrum levels.

7. CSII

Kates and Arehart (2005) presented a three-level coherence SII (CSII) based on the magnitude-squared coherence function between the clean speech and the processed noisy speech. This method has shown promising results for noisy speech subjected to peak- and center-clipping distortion.

The coherence values are integrated within each auditory frequency band represented by rounded-exponential frequency responses with normal auditory frequency resolution. The resulting signal-to-distortion ratios (SDR) in each auditory filter band are used as audibility measures in the critical-band version of the standard SII. The importance weighting for "average speech" (ANSI, 1997, Table I) was used.

Before calculating the coherence measures and the SDR, the signal segments are separated into three level regions based on the clean-speech root-mean-square (RMS) values. The high-level segments are those at or above the overall RMS level. The mid-level segments are those between 0 and 10 dB below the overall RMS level, and the low-level segments are those between 10 and 30 dB below the overall RMS level.

The signals are segmented into 16-ms Hamming-windowed frames with 50% overlap, i.e., 8-ms time resolution. The results for the three level regions were weighted by factors 1.84 for the low-level segments, 9.99 for mid-level region, and 0.0 for the high level region, as Kates and Arehart (2005) found these weight factors to be optimal. A MATLAB implementation was kindly made available by Georg Stiefenhofer in 2010.

The hearing loss is accounted for by a threshold-equivalent internal masking noise spectrum exactly as in the standard SII procedure (Sec. III A1). No desensitization factor was applied.

8. sEPSM

Jørgensen and Dau (2011) proposed an intelligibility measure based on long-term SNRs in the modulation-frequency domain, called speech-based envelope power spectrum model (sEPSM), and found good agreement with speech-recognition results in stationary noise and reverberation and with a variant of NR by spectral subtraction.

This method uses both the processed speech-plus-noise and the processed noise-only signals as input. The signals are first spectrally analyzed in a filter bank with 22 fourth-order complex gammatone filters with 1/3-octave spacing, covering the frequency range from 63 Hz to 8 kHz. The envelopes of the output signals in each filter band are analyzed in a modulation filter bank with seven filters, one low-pass filter with 1-Hz cutoff frequency and six overlapping bandpass filters with center frequencies with octave spacing from 2 to 64 Hz. The long-term power in each modulation filter band, and each audio frequency channel, is calculated in the frequency domain for the processed noisy speech and for the processed noise only, and the ratio between these values is the SNR. The final overall modulation SNR is calculated by power summation across modulation frequencies and audio frequency channels. All the calculations were done based on a MATLAB implementation kindly made available by Søren Jørgensen in Feb. 2012.

To account for the individual hearing loss, the final summation includes only those audio channels where the long-term third-octave band level of the processed noisy speech exceeds the corresponding threshold-equivalent reference levels for third-octave filtered noise in a diffuse sound field (ISO, 2005). Compared to the method described in the original article, this is a slight revision suggested by Jørgensen (April 8, 2014, personal communication).

For the listeners with impaired hearing, the hearing threshold loss (in decibel HL) was interpolated to the center frequencies of the audio channels and added to the normal reference threshold. The hearing thresholds and insertion gain values were used only to determine which audio channels to include for each ear.

9. mr-sEPSM

Jørgensen *et al.* (2013) presented a refined version, multi-resolution speech-based envelope power spectrum model (mr-sEPSM), of the envelope modulation power spectrum model to better account for fluctuating noise, reverberation, and non-linear noise reduction. The new version is similar to the previous sEPSM approach. However, now the modulation filter bank has nine channels, one low-pass and eight bandpass filters with octave-spaced center frequencies from 2 to 256 Hz.

The main difference is that the output from the modulation filters is now represented in the time domain. The filtered envelope signals for the noisy speech and the noise in each modulation band is segmented into non-overlapping blocks. For the low-pass filter, the block duration is 1 s, and for the band-pass filters, the durations are equal to the inverse of the band-pass filter center frequencies. A modulation SNR is calculated for each time segment, each modulation band, and each audio channel.

The final modulation SNR is calculated by power summation across modulation frequencies, time segments, and audio frequency channels, including only those audio channels where the long-term level of the processed noisy speech exceeded the hearing threshold. All these calculations were done using MATLAB code kindly made available by Søren Jørgensen in Dec. 2012.

The hearing loss was accounted for in the same way as for the sEPSM (Sec. III A 8).

B. Data presentation

1. Performance indicators

In the following, the data analysis necessary to compare measured and predicted benefit of NR processing is described. The performance indicators were selected to quantify the predicted and measured effects of NR algorithms on the same decibel scale, such that the prediction errors could be directly compared across different measures. Meyer and Brand (2013) used a similar method to calculate predicted SRTs.

The measured benefit B_{na} was quantified as the reduction in the speech recognition threshold (SRT, in dB) for the nth listener using NR algorithm a,

$$B_{na} = SRT_{nu} - SRT_{na}, \quad (1)$$

where subscript u indicates the unprocessed condition and subscript a processing with noise reduction algorithm a. Thus a positive value of B_{na} indicates that the NR algorithm improved the listener's speech-recognition ability, yielding a lower SRT with NR than without.

Each predictive measure was then used to estimate a corresponding predicted benefit. The method for this estimation is illustrated in Fig. 3 by an example using one predictive measure (SII) for one listener with impaired hearing and one of the NR algorithms (WEDM).

The upper panel of Fig. 3 shows that the measured NR benefit was -1.9 dB. The "benefit" is negative as the SRT was higher with NR (filled square) than without (filled circle).

The lower panel shows how the SII is used to calculate two corresponding predicted measures of NR benefit. For all available speech and noise signals, the SII was calculated as a function of SNR, for the unprocessed condition (dashed line in the lower panel of Fig. 3) and for the WEDM processed condition (solid line) for SNRs from -12 to $+15$ dB in 1-dB steps.

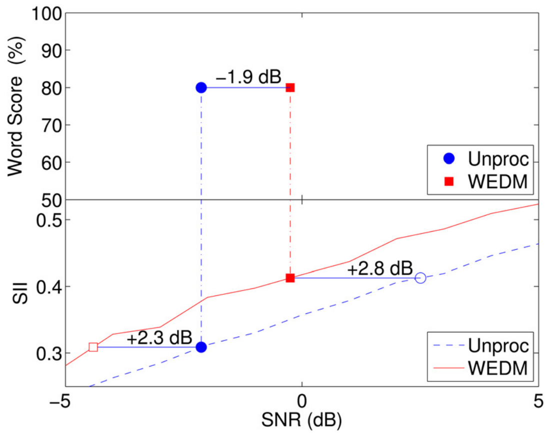

FIG. 3. (Color online) Example of measured (upper panel) and predicted (lower panel) speech recognition performance for one listener with impaired hearing, using the WEDM algorithm and the SII predictive measure. Upper panel: Measured SRT for the unprocessed signals (filled circle) and with the WEDM processed signals (filled square). A negative "benefit" of -1.9 for the WEDM was seen in this example. Lower panel: Calculated SII values are shown as a function of the signal-to-noise ratio (SNR) with a dashed curve for the unprocessed condition and a solid curve for the WEDM processed condition. Filled symbols mark the calculated SII values at the measured SRTs. The unfilled square shows the SNR giving the same SII with WEDM as the SII at the measured SRT with unprocessed signals. The unfilled circle shows the SNR giving the same SII with the unprocessed signals as the SII at the measured SRT with WEDM processing. The horizontal lines in the lower panel show two predicted values of the SRT change caused by the WEDM, indicating a positive benefit of $+2.3$ and $+2.8$ dB in this example.

The first predicted benefit was determined by calculating an SII value using the measured SRT for the unprocessed condition (filled circle in the lower panel of Fig. 3). A corresponding SNR, at the same SII, was then determined for the WEDM processed condition (unfilled square). In this example, the calculated benefit was about $+2.3$ dB.

The second predicted benefit was determined by calculating an SII value using the measured SRT for the WEDM processed condition (filled square in the lower panel of Fig. 3). A corresponding SNR, at the same SII, was then determined for the unprocessed condition (unfilled circle). This calculated benefit was about $+2.8$ dB.

A final measure of predicted benefit, PB_{na}, was then calculated as the average of these two values, i.e., $PB_{na} \approx (2.8 + 2.3)/2 \approx +2.5$ dB. In this example, the predicted benefit was positive because the SII predicts a lower SRT with NR than without.

The predicted benefit was calculated in this way for each listener, each NR algorithm, and each of the predictive measures.

The prediction error D_{na} was then calculated as

$$D_{na} = PB_{na} - B_{na}, \quad (2)$$

where PB_{na} is the predicted and B_{na} the measured benefit for the nth listener with NR algorithm a. Thus in the example in Fig. 3, the prediction error is $D_a \approx 2.5 - (-1.9) = +4.4$ dB.

The larger the prediction error, the worse the predictive measure works. The error was also considered more serious if the predicted benefit pointed in the wrong direction, i.e., predicted a positive NR benefit when the real benefit was negative (as in the example) or vice versa. This type of error was quantified by calculating the un-normalized cross-correlation (C_n) between measured and predicted benefit values for each listener as

$$C_n = \sum_a B_{na} \cdot PB_{na}, \quad (3)$$

where B_{na} is the measured and PB_{na} the predicted benefit for the nth listener with NR algorithm a.

2. Statistical tests

A good predictive measure should give distributions of individual prediction errors (D_{na}) with medians near zero for all NR conditions. Friedman's two-way analysis of variance by ranks test (MATLAB Statistical Toolbox, v. 8.2) was used to determine, for each predictive measure, whether there were any statistically significant rank-order differences ($p < 0.05$) between prediction errors across the three NR conditions and the unprocessed condition, where the prediction error, by definition, was zero for all participants. This test reveals if a measure gives prediction errors that deviate systematically from zero for at least one NR algorithm.

To identify predictions in the wrong direction, a signed-rank Wilcoxon test (MATLAB Statistical Toolbox, v. 8.2) was used to test if the median of the cross-correlation C_n, across listeners, was negative ($p < 0.05$, one-tailed).

C. Results

The distributions of measured and predicted benefit are presented in Fig. 4 (for participants with impaired hearing, HI) and Fig. 5 (for participants with normal hearing, NH). Predictive measures that did not show any statistically significant rank-order differences between prediction errors across processing conditions (Friedman two-way analysis of variance by ranks test, $p < 0.05$) have been marked "OK" in the figure. These methods did not show prediction errors deviating significantly from zero for any of the NR algorithms.

Only one of the predictive measures, the CSII (Kates and Arehart, 2005), passed this test for both listener groups. For the listeners with impaired hearing (Fig. 4), the following measures also passed the test: STSII, STOI, and mr-sEPSM.

Table I presents a single-number performance indicator for each measure. The median prediction error was first calculated across listeners for each NR algorithm, and the largest prediction error among the three algorithms is shown. The CSII showed prediction errors less than 1 dB for both groups of listeners. The STOI showed errors less than 1 dB for the group with impaired hearing.

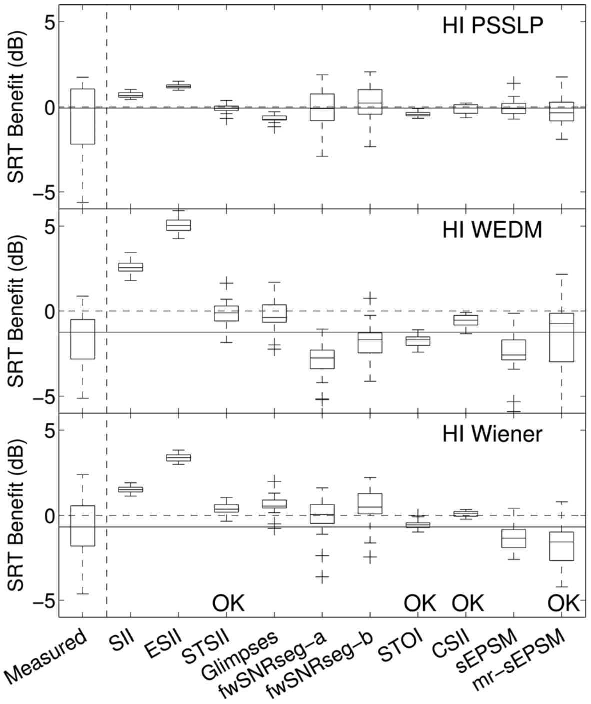

FIG. 4. Measured and predicted benefit with three NR algorithms (PSSLP, WEDM, Wiener) for 20 listeners with impaired hearing (HI). The NR benefit is quantified by the reduction of the SRT in noise, defined here as the SNR required for 80% correct responses in the adaptive speech test. The left-most box-plot shows the distribution of measured benefit. The solid horizontal line indicates the median measured benefit. The other box-plots show the predicted benefit for each measure. Each boxplot shows interquartile values, and the median is represented by the line in the box. Outliers (+) are defined as values outside 1.5 times the box length, and the whiskers extend to the highest and lowest values when the outliers are excluded. Predictive measures that did not show significant prediction errors, according to the Friedman test, described in Sec. III B 2, have been marked "OK."

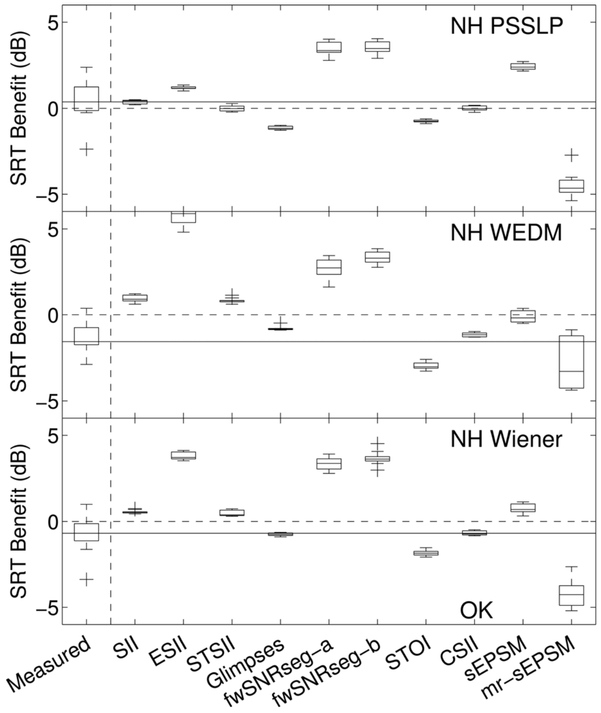

FIG. 5. Measured and predicted benefit with three NR algorithms for 10 listeners with normal hearing (NH), displayed as in Fig. 4.

Statistically significant predictions in the wrong direction were found for the SII and the ESII for both groups of listeners, and for the STSII for the NH group. These measures are clearly not good predictors of the effect the NR algorithms have on speech recognition in noise.

TABLE I. SRT prediction errors for two groups of listeners, 20 with impaired hearing (HI) and 10 with NH. The median prediction error was calculated across listeners, and the result with the largest magnitude among the three noise reduction algorithms is shown. Bold numbers indicate that the prediction errors were significantly ($p < 0.05$) different from zero for at least one noise reduction algorithm, as indicated by the Friedman test described in Sec. III B 2. Underlined numbers indicate that the predicted benefit was also in the wrong direction compared to the measured benefit and that this discrepancy was statistically significant ($p < 0.05$), as indicated by the correlation test described in Sec. III B 2.

		Prediction error (dB)	
Measure	Reference	HI	NH
SII	ANSI (1997)	**3.8**	**2.5**
ESII	Rhebergen *et al.* (2006)	**6.4**	**7.0**
STSII	This paper	1.2	**2.4**
Glimpses	Cooke (2006)	**1.4**	**–1.5**
fwSNRseg-a	Ma *et al.* (2009)	**–1.6**	**4.1**
fwSNRseg-b	This paper	**1.0**	**4.8**
STOI	Taal *et al.* (2011a)	–0.6	**–1.6**
CSII	Kates and Arehart (2005)	0.9	0.4
sEPSM	Jørgensen and Dau (2011)	**–1.2**	**2.0**
mr-sEPSM	Jørgensen *et al.* (2013)	–1.2	**–5.0**

IV. DISCUSSION

Only one of the predictive measures, the CSII (Kates and Arehart, 2005), correctly predicted the effects of NR processing on the speech recognition threshold (SRT) for both groups of listeners. All other measures showed statistically significant differences between the predicted and measured benefit for at least one of the listener groups. The CSII has shown favorable results in some other evaluations (e.g., Ma *et al.*, 2009; Xia *et al.*, 2012). In contrast, Taal *et al.* (2011b) found that the CSII performed worse than a majority of their evaluated measures. That study used NR based on so-called ideal time-frequency segregation and generally lower SNRs than the current evaluation, including one extreme condition with a −60 dB SNR.

Some other measures, using short-time analysis of both speech and noise (STSII, STOI, and mr-sEPSM), showed small and statistically non-significant prediction errors for the HI group (Fig. 4 and Table I). These methods seem to work better than measures using long-term analysis (SII and ESII).

The NR algorithms increase the short-term modulations of both speech and noise. These artificial modulations were probably mostly perceived as distortion and did not improve speech recognition for the listeners. WEDM was the algorithm that led to the largest increase in long-term SNR (Sec. II A 3), but this algorithm also produced the largest amount of distortion.

The SII uses long-term spectra for both speech and noise, and the ESII uses the long-term spectrum for speech. These measures clearly overestimated the predicted benefit of the WEDM processing because they are sensitive to the improvement in long-term SNR but insensitive to the distortion produced. The ESII was designed to account for the beneficial effects of short-term noise fluctuations, where the speech signal is not distorted. This was probably a disadvantage for the ESII when applied to NR-processed signals and might explain why the ESII showed even larger prediction errors than the standard SII in the current study.

Meyer and Brand (2013) evaluated the standard SII and five STSII variants, including the ESII, for their ability to predict speech recognition thresholds in fluctuating noise with different forms of modulation for listeners with normal and with impaired hearing. In that evaluation, the ESII performed better than the standard SII. They also found that a STSII version similar to the current STSII (but using frequency-dependent time windows) gave slightly better correlations with measured SRTs than the ESII, but there was still a large unexplained variability among listeners with impaired hearing.

The measures using correlation between the clean speech and the processed noisy speech, CSII and STOI, did not systematically overestimate the performance with WEDM, probably because they take the speech distortion into account. The STOI has also shown promising results in previous evaluations (e.g., Taal *et al.*, 2011a; Xia *et al.*, 2012).

In practical evaluations of NR algorithms, a measured improvement of the SRT by about 1 dB would be considered a clearly interesting and valuable result. The current evaluation method was designed to be very sensitive in revealing if a predictive measure might indicate false effects of NR algorithms.

Other studies (e.g., Ma *et al.*, 2009; Taal *et al.*, 2011b; Xia *et al.*, 2012) have evaluated predictive measures in terms of correlation, or deviations, between measured and predicted speech recognition scores across a wide range of listening conditions. These more traditional indicators of merit may be more appropriate if a predictive measure is used mainly to indicate overall effects of varying acoustic conditions. For that purpose, it might be less important if the method does not accurately predict small effects of changes in the signal processing.

The prediction errors in Table I should be interpreted with some caution, considering how the values were obtained. The idea was to quantify the predicted benefit of NR by the horizontal distance (in decibels) between the curves in the lower panel of Fig. 3, showing the predictive measure as a function of SNR for NR-processed and for unprocessed stimuli. However, the two curves are not exactly parallel because both the NR processing and the calculation of the predictive measures are affected by the SNR. In the presented results, each calculated value of the predicted benefit of NR was the mean of the values at two different reference SNRs [Fig. 3 lower panel, where the predicted benefit was (2.8 + 2.3)/2 dB]. Alternatively, the predicted benefit could have been quantified with reference only to the measured SNR at the unprocessed condition, i.e., using only the +2.3 dB benefit indicated in the lower panel of Fig. 3. A test using this alternative method showed results that were very similar to those presented because the predicted benefit was similar at both reference SNRs for most predictive measures. However, the prediction errors for the NH group, presented in Table I, would have been slightly smaller for ESII (but still deviating significantly from zero and still with a prediction in the wrong direction) and slightly larger for fwSNRseg-a, fwSNRseg-b, and for mr-sEPSM. Thus the conclusions would have been exactly the same if the alternative method had been used. The selected method, using the average of two predicted benefit values, is more balanced and also tends to reduce the random variability in the predicted results.

The present study evaluated the effects on speech recognition at a single point on the listeners' psychometric function, at the SNR where the listeners achieved 80% correct responses in the test. This criterion was selected to achieve SNRs which were reasonably realistic (Smeds *et al.*, 2014) for hearing-aid users (Fig. 2). Testing at a more conventional performance level of 50% correct would result in very low SNRs with the Hagerman speech test material.

To quantify the predicted effects in terms of recognition scores at a fixed SNR, it would have been necessary to also consider the transfer function from the measured signal characteristics to a predicted recognition score. As this transfer function usually depends on the speech material, it might introduce an additional source of error. With the present study design, using the change in SRT as performance measure, the unprocessed condition served as an individual reference condition for each listener. This was considered an advantage, since this design avoided the difficulty of estimating an optimal transfer function.

V. CONCLUSIONS

Nine predictive measures of speech intelligibility, one measure in two versions, were evaluated with regard to their ability to predict the speech-recognition benefit of single-channel noise reduction processing. Two groups of listeners participated, one with and one without hearing impairment. The speech-to-noise ratio was adjusted in an adaptive speech recognition test so that all listeners achieved equal word recognition scores across four test conditions with three different single-channel NR algorithms and one unprocessed condition.

Only one of the predictive measures, CSII (Kates and Arehart, 2005), correctly predicted the effect of the currently tested noise reduction algorithms on the speech recognition threshold within both groups of listeners. In general, measures using correlation between the clean speech and the processed noisy speech, as well as other measures that are based on short-time analysis of speech and noise, seemed most promising.

ACKNOWLEDGMENTS

The authors thank Martin Dahlquist, Josefina Larsson, and Martin Hansen for valuable discussions and the participants in the experiment for their work. Koenrad S. Rhebergen, Martin Cooke, Cees H. Taal, Søren Jørgensen, and Georg Stiefenhofer kindly made essential parts of their MATLAB code available for the current analyses. The authors want to express their sincere gratitude for those very valuable contributions.

ANSI (**1997**). S3.5, *American National Standard Methods for the Calculation of the Speech Intelligibility Index* (Acoustical Society of America, 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**(8), 447–460.

Bentler, R. A. (**2006**). "Digital noise reduction: An overview," Trends Amplif. **10**(2), 67–82.

Boymans, M., and Dreschler, W. A. (**2000**). "Field trials using a digital hearing aid with active noise reduction and dual-microphone directionality," Audiology **39**(5), 260–268.

Brons, I., Houben, R., and Dreschler, W. A. (**2013**). "Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort," Ear Hear. **34**(1), 29–41.

Byrne, D., and Dillon, H. (**1986**). "The National Acoustic Laboratories (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear Hear. **7**(4), 257–265.

Byrne, D., Dillon, H., Ching, T., Katsch, R., and Keidser, G. (**2001**). "NAL–NL1 Procedure for fitting nonlinear hearing aids: Characteristics and comparisons with other procedures," J. Am. Acad. Audiol. **12**(1), 37–51.

Chung, K. (**2004**), "Challenges and recent developments in hearing aids: I. Speech understanding in noise, microphone technologies and noise reduction algorithms," Trends Amplif. **8**(4), 83–124.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise." J. Acoust. Soc. Am. **119**(3), 1562–1573.

Dahlquist, M., Lutman, M. E., Wood, S., and Leijon, A. (**2005**). "Methodology for quantifying perceptual effects from noise suppression systems," Int. J. Audiol. **44**(12), 721–732.

Fletcher, H. (**1929**). *Speech and Hearing* (Van Nostrand, New York), 331 pp.

Fletcher, H., and Galt, R. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**(2), 89–151.

Hagerman, B. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. **11**(2), 79–87.

Hagerman, B., and Olofsson, A. (**2004**). "A method to measure the effect of noise reduction algorithms using simultaneous speech and noise," Acta Acust. Acust. **90**(2), 356–361.

Hoetink, A. E., Körössy, L., and Dreschler, W. A. (**2009**). "Classification of steady state gain reduction produced by amplitude modulation based noise reduction in digital hearing aids," Int. J. Audiol. **48**(7), 444–455.

Holube, I., Fredelake, S., and Vlaming, M. (**2010**). "Development and analysis of an international speech test signal (ISTS)," Int. J. Audiol. **49**(12), 891–903.

Houtgast, T., and Steeneken, H. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica **28**(1), 66–73.

Hu, Y., and Loizou, P. (**2007**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**(32), 1777–1786.

Humes, L. E., Wilson, D. L., Barlow, N. N., and Garner, C. (**2002**). "Changes in hearing-aid benefit following 1 or 2 years of hearing-aid use by older adults," J. Speech Lang. Hear. Res. **45**(4), 772–782.

IEC (**2011**). 60268-16, *Sound System Equipment—Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index*, 4.0 ed. (International Electrotechnical Commission, Geneva, Switzerland).

ISO (**2005**). 389-7, *Acoustics—Reference Zero for the Calibration of Audiometric Equipment. Part 7: Reference Threshold of Hearing Under Free-Field and Diffuse-Field Listening Conditions* (International Organisation for Standardisation, Geneva, Switzerland).

Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**(3), 1475–1487.

Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**(1), 436–446.

Kates, J. (**1987**). "The short-time articulation index," J. Rehabil. Res. Dev. **24**(4), 271–276.

Kates, J. M., and Arehart, K. H. (**2005**). "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**(4), 2224–2237.

Keidser, G., Dillon, H., Flax, M., Ching, T., and Brewer, S. (**2011**). "The NAL-NL2 prescription procedure," Audiol. Res. **1**(1S), e24.

Kochkin, S. (**2010**). "MarkeTrak VIII: Consumer satisfaction with hearing aids is slowly increasing," Hear. J. **63**(1), 11–19.

Leijon, A., Lindkvist, A., Ringdahl, A., and Israelsson, B. (**1990**). "Preferred hearing aid gain in everyday use after prescriptive fitting," Ear Hear. **11**(4), 299–305.

Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL), 608 pp.

Loizou, P. C., and Kim, G. (**2011**). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," IEEE Trans. Audio Speech Lang. Proc. **19**(1), 47–56.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Büchler, M., Dillier, N., Houben, R., Dreschler, W., Froelich, M., Puder, H., Grimm, G., Hohmann, V., Leijon, A., Lombard, A., Mauler, D., Moonen, M., and Spriet, A. (**2010**). "Multicenter evaluation of signal enhancement algorithms for hearing aids," J. Acoust. Soc. Am. **127**(3), 1491–1505.

Ma, J., Hu, Y., and Loizou, P. C. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new bandimportance functions," J. Acoust. Soc. Am. **125**(5), 3387–3405.

Meyer, R. M., and Brand, T. (**2013**). "Comparison of different short-term speech intelligibility index procedures in fluctuating noise for listeners with normal and impaired hearing," Acta Acust. Acust. **99**(3), 442–456.

Moore, B. C. (**1996**). "Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids," Ear Hear. **17**(2), 133–161.

Pavlovic, C., Studebaker, G., and Sherbecoe, R. (**1986**). "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," J. Acoust. Soc. Am. **80**(1), 50–57.

Peeters, H., Kuk, F., Lau, C.-C., and Keenan, D. (**2009**). "Subjective and objective evaluation of noise management algorithms," J. Am. Acad. Audiol. **20**(2), 89–98.

Rhebergen, K., Versfeld, N. J., and Dreschler, W. A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**(6), 3988–3997.

Rhebergen, K. S., Versfeld, N. J., de Laat, J., and Dreschler, W. A. (**2010**). "Modelling the speech reception threshold in non-stationary noise in hearing-impaired listeners as a function of level," Int. J. Audiol. **49**(11), 856–865.

Smeds, K., Bergman, N., and Nyman, T. (**2010a**), "Noise reduction in modern hearing aids—Long-term and short-term measurements," in *Binaural Processing and Spatial Hearing*, edited by J. M. Buchholz, T. Dau, J. C. Dalsgaard, and T. Poulsen (The Danavox Jubilee Foundation, Helsingør, Denmark), pp. 445–452.

Smeds, K., Wolters, F., Nilsson, A., Båsjö, S., Hertzman, S., and Leijon, A. (**2010b**), "Objective measures to quantify the perceptual effects of noise reduction in hearing aids," in *Proceedings of AES 38th International Conference*, Piteå, Sweden, pp. 101–108.

Smeds, K., Wolters, F., and Rung, M. (**2014**). "Estimation of signal-to-noise ratios in realistic sound scenarios," J. Am. Acad. Audiol. (in press).

Steeneken, H., and Houtgast, T. (**1980**). "A physical method of measuring speech-transmission quality," J. Acoust. Soc. Am. **67**(1), 318–326.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011a**), "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Proc. **19**(7), 2125–2136.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011b**), "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," J. Acoust. Soc. Am. **130**(5), 3013–3027.

Wagener, K. C., Hansen, M., and Ludvigsen, C. (**2008**). "Recording and classification of the acoustic environment of hearing aid users," J. Am. Acad. Audiol. **19**(4), 348–370.

Xia, R., Li, J., Akagi, M., and Yan, Y. (**2012**). "Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4465–4468.