Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements[a)]

J. Rennies,[1,b)] A. Pusch,[2] H. Schepker,[3] and S. Doclo[3]
[1]*Department of Speech, Language and Hearing Sciences and Hearing Research Center, Boston University, Boston, Massachusetts 02215, USA*
[2]*Project Group Hearing Speech and Audio Technology, Fraunhofer-Institute for Digital Media Technology IDMT and Cluster of Excellence Hearing4All, Oldenburg, Germany*
[3]*Signal Processing Group, Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg, Oldenburg, Germany*
jan.rennies@idmt.fraunhofer.de, arne.pusch@idmt.fraunhofer.de, henning.schepker@uni-oldenburg.de, simon.doclo@uni-oldenburg.de

Abstract: Previous studies showed that near-end listening enhancement (NELE) algorithms can significantly improve speech intelligibility in noisy environments. This study investigates the benefit of the NELE algorithm AdaptDRC in normal-hearing listeners at signal-to-noise ratios (SNRs) for which speech intelligibility is at ceiling, by evaluating listening effort for processed and unprocessed speech in the presence of speech-shaped and cafeteria noise. The results suggest that the NELE algorithm is able to reduce listening effort over a wide range of SNRs. Hence, listening effort seems to be applicable for evaluating NELE algorithms over a much wider SNR range than speech intelligibility.

[DDO]
Date Received: June 12, 2018 **Date Accepted:** October 2, 2018

1. Introduction

The perception of speech played back via sound reinforcement systems may be considerably impaired by ambient noise in the listening room, e.g., for public announcement systems or mobile phone calls in a noisy environment. In such conditions it is usually not possible to actively reduce the noise at the listener's end of the communication system (i.e., the "near-end"). However, since the target speech is available to the communication system, it can be pre-processed to enhance speech intelligibility in noisy conditions. Such pre-processing algorithms are referred to as near-end listening enhancement (NELE) algorithms and have received increased attention in the past years. Various approaches have been proposed, e.g., based on frequency-shaping (e.g., Kleijn *et al.*, 2015; Sauert and Vary, 2012; Taal *et al.*, 2014) and broadband or frequency-dependent dynamic range compression (DRC) (e.g., Schepker *et al.*, 2015; Zorila and Stylianou, 2014). Considering the large variety of NELE algorithms, a reliable and widely applicable evaluation procedure for comparing their performance is desirable. The most established evaluation procedure is to measure the speech intelligibility benefit of NELE algorithms in formal listening tests by comparing human speech recognition performance for unprocessed and processed speech. This is typically done by evaluating the change in the percentage of correctly recognized speech items (e.g., words or sentences), or by evaluating the equivalent intensity change (Cooke *et al.*, 2013), i.e., the difference in signal-to-noise ratio (SNR) which leads to the same speech recognition performance (e.g., 50% correctly understood speech items). Several studies have reported speech intelligibility results of NELE algorithms for normal-hearing listeners (e.g., Schepker *et al.*, 2015; Taal and Jensen, 2013; Tang and Cooke, 2011) and for hearing-impaired listeners (e.g., Rennies *et al.*, 2017). In a large-scale comparison study, Cooke *et al.* (2013) measured word recognition performance for a large variety of NELE algorithms. While the benefit differed considerably between algorithms (some algorithms even decreased speech intelligibility) as well as between different types of background noise, one general observation was that speech intelligibility could be

---

[a)]Portions of this work have been presented at the 44th Annual Meeting of the German Acoustical Society 2018 in Munich, Germany.

[b)]Also at: Project Group Hearing, Speech and Audio Technology, Fraunhofer-Institute for Digital Media Technology IDMT and Cluster of Excellence Hearing4All, Oldenburg, Germany. Author to whom correspondence should be addressed.

significantly increased (without increasing the speech level) by algorithms comprising frequency-shaping and DRC.

Measuring speech intelligibility has a fundamental limitation in that only rather adverse listening conditions can be assessed. For example, ceiling effects in speech intelligibility (i.e., performance close to optimum) for unprocessed speech of the Oldenburg sentence test (Wagener *et al.*, 1999) occurred at SNRs of about −4 dB for normal-hearing listeners. For speech processed by the NELE algorithm *AdaptDRC* (Schepker *et al.*, 2015), Rennies *et al.* (2017) reported that ceiling effects were reached at SNRs between −16 and −6 dB, depending on the type of background noise. This means that at larger SNRs it is not possible to derive meaningful conclusions about algorithm performance from speech intelligibility measurements, because listeners are always able to recognize most speech items, regardless of the type of processing. Since many everyday listening conditions do not include such adverse SNRs (Smeds *et al.*, 2015), the validity of speech intelligibility measurements for evaluating algorithm performance in these conditions may be limited.

This potential limitation has motivated recent studies to evaluate listening effort rather than speech intelligibility, i.e., the cognitive effort associated with extracting the relevant speech information from the target signal even when intelligibility is very good (e.g., Krueger *et al.*, 2017; Rennies *et al.*, 2014; Sato *et al.*, 2012; Schepker *et al.*, 2016). Although listening effort has been used to evaluate, e.g., single-channel noise reduction algorithms (e.g., Luts *et al.*, 2010) or other algorithms, to the best of our knowledge, listening effort has not yet been measured in the context of NELE algorithms. However, this may be of particular interest for this kind of algorithm, since many potential application scenarios (e.g., announcements in train stations) involve a large range of SNRs. The goal of this study is therefore to apply an established listening effort assessment method (Krueger *et al.*, 2017) to assess a NELE algorithm over a wide range of SNRs, and to compare the results to speech intelligibility measurements.

2. Methods

2.1 Subjects

Eleven normal-hearing subjects (nine male and two female) participated in the experiments. All were native German speakers and had normal audiograms with pure-tone averages lower than 25 dB hearing level. The subjects were between 24 and 36 yr old (median age 27 yr).

2.2 Stimuli and equipment

The speech material was taken from the Oldenburg sentence test (Wagener *et al.*, 1999). It consists of sentences of five words with the fixed syntactical structure *name verb numeral adjective object*, e.g., "Peter hat drei teure Autos" (Engl. "Peter has three expensive cars"). For each word group ten alternatives are available, which are randomly combined to result in syntactically correct, but semantically unpredictable sentences. Both unprocessed speech and speech processed by the *AdaptDRC* algorithm (Schepker *et al.*, 2015, see below) was used. The speech level was always fixed at 60 dB sound pressure level. Two different noise types were used: stationary speech-shaped noise (SSN) with the same average long-term spectrum as the unprocessed sentences, and cafeteria noise which contained more envelope fluctuations and was previously used by Schepker *et al.* (2015) and taken from the data base of Kayser *et al.* (2009). The noise levels were varied to achieve the desired SNRs.

The signals were digitally mixed in MATLAB, D/A-converted (RME ADI-8 PRO, Chemnitz, Germany), and amplified (DT HB7, Tucker-Davis Technologies, Alachua, FL). To be comparable to previous studies, the stimuli were presented diotically to the subjects via headphones (Sennheiser HD650, Wedemark, Germany) in a sound-attenuated booth.

2.3 NELE

Details of the *AdaptDRC* algorithm can be found in Schepker *et al.* (2015). Briefly, the algorithm processes speech signals in time frames of 20 ms and consists of two processing stages. The first is a frequency-shaping stage, where the known speech signal and the (estimated) environmental noise signal are divided into eight octave bands, centered at 125 Hz to 16 kHz. From the sub-band levels a simplified version of the Speech Intelligibility Index (SII) is computed, based on which the sub-band speech levels are weighted. For an SII equal to 1, no weighting is applied such that the spectral shape is not modified. For an SII equal to 0, the weighting results in equal level in all sub-bands. In typical scenarios, this weighting corresponds to an amplification of high frequencies, since speech usually has a sloping frequency spectrum. For SII values between 0 and 1, a continuous transition between a non-modified and a flat octave-band spectrum is applied. The second stage is a DRC stage, where in each sub-band softer parts are amplified relative to more intense parts in order to increase audibility. The compression ratio depends on the sub-band SNR and is between 1:1 (for SNR ≥ 15 dB) and 1:8 (for SNRs ≤ -15 dB). For intermediate sub-band SNRs a continuous transition of the compression ratio is applied. The root-mean-square (rms)-power of the re-combined output speech signal is normalized to the rms-power of the unprocessed speech signal, which is a typical constraint for evaluating NELE algorithms.

2.4 Procedures

All subjects started with the listening effort measurement, followed by the speech intelligibility measurement. Listening effort was measured in four conditions (unprocessed and processed speech both in SSN and cafeteria noise) over a wide range of SNRs (-15, -10, -5, 0, 2.5, 5, 7.5, 10 dB) using a constant-stimuli procedure. Each data point was measured six times by each subject. For each trial a randomly selected sentence and noise start sample were used. The trials were divided into six blocks, where each block contained one stimulus of each condition. The order of the conditions and the SNRs within each block were randomized for each block and subject. The task of the subjects was to rate the listening effort on a categorical 13-point scale ranging from "no effort" (1 Effort Scaling Categorical Unit, ESCU) to "extreme effort" (13 ESCU) (Krueger *et al.*, 2017) on a graphical user interface. In addition, a 14th category ("only noise") was available for trials in which subjects could not detect any speech signal in the presented mixture. The stimulus of each trial was played in a loop and the subjects were instructed to listen at least once to the whole sentence before making their choice.

Speech intelligibility was measured using a list of 20 sentences for each condition. For the SSN masker, speech intelligibility was measured at SNRs of -10 and -6 dB (unprocessed) and at SNRs of -22 and -14 dB (*AdaptDRC*), respectively. For the cafeteria masker, intelligibility was measured at SNRs of -14 and -5 dB (unprocessed) and at SNRs of -19 and -10 dB (*AdaptDRC*). These SNRs were selected based on a previous study (Schepker *et al.*, 2015) to produce about 20% and 80% correctly recognized words. Conditions were randomized for each subject, but all 20 sentences of one condition were finished before starting the next condition. Every sentence was played once. The task of the subjects was to orally repeat the words they had understood. The experimenter marked the correct responses and no feedback was provided. To train the subjects, two lists of 20 sentences using unprocessed stimuli were measured before the main measurement started.

3. Results

3.1 Speech intelligibility measurements

The bottom panels of Fig. 1 show mean speech intelligibility data across subjects (symbols) for the SSN (left) and cafeteria noise (right). Errorbars represent plus and minus one standard deviation. For each subject psychometric functions were estimated for each noise type and processing type by fitting a sigmoid function (Brand and Kollmeier, 2002) to the data points. Mean psychometric functions (lines in Fig. 1) were obtained by averaging the individual parameters of the fits across subjects. For both noise types a considerable improvement in speech intelligibility was found as indicated by the leftward shift of the psychometric functions. The mean shift at the speech reception threshold (SRT, i.e., the SNR for 50% speech intelligibility) was 10.8 dB for SSN and 5.8 dB for cafeteria noise, respectively.

To further analyze the algorithm benefit in terms of a shift along the SNR axis, SNRs corresponding to 20%, 50%, and 80% speech intelligibility were derived from each subject's psychometric functions by interpolation (this corresponds to finding the intersection between the psychometric functions with horizontal lines at these intelligibility values). These SNRs were analyzed for normal distribution and used as independent variables in a three-factor repeated-measures analysis of variance (ANOVA) with factors noise type, processing type, and intelligibility value (see, e.g., Kleinbaum *et al.*, 2013). The significance level was 0.05 and the degrees of freedom were Greenhouse-Geisser corrected. The ANOVA showed that all three factors had a significant influence on SNRs [noise: $F(1,10) = 26.697$, $p < 0.001$; processing: $F(1,10) = 1049.046$, $p < 0.001$; intelligibility: $F(1,10.001) = 108.717$, $p < 0.001$]. All two-factor interactions were also significant, indicating that the algorithm benefit depended on noise type [noise*processing:

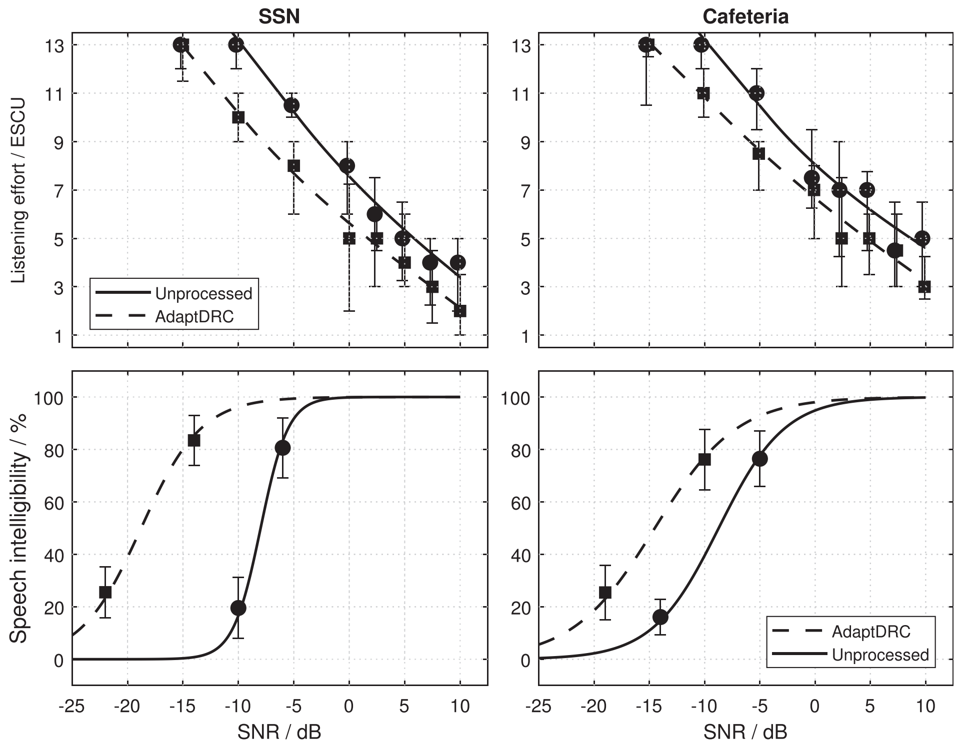

Fig. 1. Top: Median listening effort ratings, errorbars indicate interquartile ranges. Data points are slightly shifted horizontally to increase readability. Bottom: Mean speech intelligibility data, errorbars indicate plus and minus one standard deviation.

$F(1,10) = 105.958$, $p < 0.001$], that the slope of the psychometric functions depended on noise type [noise*intelligibility: $F(1,10.005) = 20.735$, $p = 0.001$], and that the algorithm benefit depended on intelligibility value [processing*intelligibility: $F(1,10.002) = 11.547$, $p = 0.007$]. The three-factor interaction was not significant [$F(1,10.001) = 0.019$, $p = 0.892$]. As *post hoc* analyses, separate two-factor ANOVAs with the factors noise type and processing type were conducted for each intelligibility value. For all intelligibility values, the factor processing type and the interaction between processing type and noise type were significant, indicating that the algorithm benefit depended on noise type over a large range of the psychometric function. The main effect of noise type was significant for intelligibility values of 50% and 80%, but not for 20%.

As a representative value for intelligibility close to ceiling, 95% was selected and the corresponding SNRs were derived from the mean psychometric functions. The left part of Table 1 summarizes these SNRs, which were −10.8 and −3.6 dB for *AdaptDRC*-processed speech in SSN and cafeteria noise, respectively, and −4.3 and +0.1 dB for unprocessed speech.

3.2 Listening effort scaling

The top panels of Fig. 1 show median listening effort ratings across subjects and repetitions (symbols) for SSN (left) and cafeteria noise (right). Errorbars represent interquartile ranges. Lines represent psychometric functions fitted to the median data (without the "noise only" ratings). The same function as employed by Krueger *et al.* (2017) was used, i.e., a function consisting of two straight lines intersecting at a categorical

Table 1. Left part: SNRs at which speech intelligibility was 95% and the mean listening effort ratings at these SNRs. Right part: SNRs at which listening effort was 12 ESCU and the mean speech intelligibility at these SNRs.

Condition	SNR/dB for 95% speech intelligibility	Corresponding perceived listening effort / ESCU	SNR/dB for listening effort of 12 ESCU	Corresponding speech intelligibility/%
SSN, unprocessed	−4.3	9.8	−8.0	50
SSN, *AdaptDRC*	−10.8	10.6	−13.3	88
Cafeteria, unprocessed	+0.1	8.0	−7.8	58
Cafeteria, *AdaptDRC*	−3.6	8.1	−12.6	63

listening effort of 7 ESCU and a smooth transition (Bézier function) between them. For unprocessed speech, data collected at the lowest SNR of −15 dB were omitted during the fitting process, because the maximum rating of 13 ESCU was already reached at the next highest SNR (−10 dB).

In general, listening effort decreased with increasing SNR except in conditions where listening effort was at ceiling (very low SNRs for the unprocessed conditions). At a given SNR, listening effort was always lower for *AdaptDRC*-processed speech than for unprocessed speech. For SSN, this difference was about 3 ESCU for SNRs of −10, −5, and 0 dB, and about 1 ESCU for SNRs of 2.5, 5, and 7.5 dB. At the lowest SNR of −15 dB both unprocessed and processed speech were rated with 13 ESCU (*extreme effort*). For the cafeteria noise, similar trends were observed. The differences between unprocessed and *AdaptDRC*-processed speech were slightly smaller than for SSN (2 to 2.5 ESCU for SNRs of −10 and −5 dB, respectively).

Similarly as for the intelligibility data, SNRs corresponding to the same listening effort value were derived from the individual psychometric functions, which had been fitted to all listening effort ratings of each subject. Because some data points for ratings of 13 ESCU had been omitted due to ceiling effects (as described above), and because only few data points were available for ratings below 4 ESCU, this analysis was limited to the range between 4 and 12 ESCU. SNRs were derived for listening effort ratings in steps of 1 ESCU and then used as independent variables in a three-factor repeated measures ANOVA with factors noise type, processing type, and listening effort value. The main effects of processing type [$F(1,10) = 52.196$, $p < 0.001$] and listening effort value [$F(1.088,10.882) = 80.886$, $p < 0.001$] were significant, while the main effect of noise type was not [$F(1,10) = 2.098$, $p = 0.178$]. The interaction between processing type and noise type was not significant, indicating that the algorithm benefit was similar for both noise types. The interaction of processing type and listening effort value was significant, indicating that the algorithm benefit depended on the position along the psychometric function. The interaction between listening effort value and noise type was also significant, indicating that the slope of the psychometric functions depended on noise type. The three-factor interaction was also significant. As *post hoc* analyses, both noise types were considered separately, and paired *t*-tests were conducted to test if the algorithm benefit was significant at the different listening effort values (resulting in a total of 18 tests, i.e., nine for each noise type). To account for multiple comparisons, the significance level was adjusted to $0.05/18 = 0.0028$. For SSN, differences for all listening effort values were significant ($p < 0.001$) except 4 ESCU ($p = 0.005$). For cafeteria noise, differences for listening effort values from 6 to 11 ESCU were significant ($p \leq 0.002$), while differences at 12 ESCU ($p = 0.008$), 5 ESCU ($p = 0.007$), and 4 ESCU ($p = 0.039$) were not.

To analyze the relation between listening effort and speech intelligibility in conditions where intelligibility approaches ceiling, the listening effort ratings for SNRs at which speech intelligibility was 95% were derived from the group psychometric functions (Table 1). For the cafeteria noise, listening effort ratings of both unprocessed and processed speech were about 8 ESCU. For SSN, the listening effort ratings were somewhat higher (about 10 ESCU). Analogously, the SNRs at which listening effort approached ceiling (12 ESCU) and the corresponding intelligibility values were derived (right part of Table 1). For unprocessed speech, these SNRs corresponded to an average speech intelligibility of 50%–60% for both noise types. For processed speech, these SNRs corresponded to an average intelligibility of 88% (SSN) and 63% (cafeteria noise).

4. Discussion

The algorithm benefit measured in this study as SRT-shift at 50% speech intelligibility was within 1.5 dB of the data of previous studies employing the same noise type and algorithm (Rennies *et al.*, 2017; Schepker *et al.*, 2015). Similarly, the listening effort ratings of this study were within 1.3 ESCU of the data reported by Rennies *et al.* (2014) for SNRs between −10 and +6 dB for the same speech material and SSN. This suggests that the methods employed here to assess speech intelligibility and listening effort produce highly comparable results across different groups of normal-hearing listeners.

To the best of our knowledge, the benefit of NELE algorithms has not been evaluated by simultaneous speech intelligibility and listening effort measurements before. The comparison of intelligibility and effort can provide insights into the SNR range in which either measure can provide information about subjective perception and potential algorithm benefit. In line with previous studies (Krueger *et al.*, 2017;

Rennies *et al.*, 2017), the present data show that intelligibility of unprocessed speech reaches ceiling (95%) at SNRs of −4.3 and 0.1 dB in SSN and cafeteria noise, respectively. For *AdaptDRC*-processed speech, these SNRs were as low as −10.8 and −3.6 dB, respectively (Table 1). This means that listening scenarios at higher SNRs cannot be assessed by means of speech intelligibility measurements. In contrast, listening effort ratings at these SNRs were never lower than 7.5 ESCU, indicating that more than half of the available listening effort scale was still available to make assessments of algorithm benefits at higher SNRs. The data indicated a significant reduction in listening effort due to the NELE processing up to SNRs corresponding to ratings of "low effort," i.e., well into the range of positive SNRs. The lowest listening effort ratings at which a significant algorithm benefit was found were 5 and 6 ESCU for SSN and cafeteria noise, respectively, which corresponded to about 5.5 dB SNR for unprocessed speech for both noise types. For SSN, this was more than 10 dB above than the SNR at which intelligibility was at ceiling (95%) for SSN, while for the cafeteria noise this was about 5.5 dB above the SNR at which intelligibility was at ceiling (95%). In other words, the use of listening effort scaling increased the usable SNR range for assessment toward higher SNRs by about the same amount as the SRT shift (10.8 and 5.8 dB for SSN and cafeteria noise, respectively) for the conditions tested in this study. At even higher SNRs, listening effort ratings did not differ significantly between unprocessed and processed stimuli, which reflects the intended adaptive behavior of *AdaptDRC* to gradually reduce the degree of processing as listening conditions improve.

Toward the lower end of the SNR range, listening effort ratings reached ceiling (12 ESCU) at SNRs of about −8 dB (unprocessed) and −13 dB (processed). At these SNRs, intelligibility was about 50%–60% for unprocessed speech, and about 60% and 90% for processed speech in cafeteria noise and SSN, respectively. For both noise types, a significant SNR reduction due to the processing was found at an intelligibility value of 20%. The corresponding SNRs were lower than the SNRs at which listening effort was at ceiling (12 ESCU) by 1.7 dB (unprocessed speech, SSN), 5.1 dB (unprocessed speech, cafeteria), 9.2 dB (processed speech, SSN), and 7.1 dB (processed speech, cafeteria). The small additional SNR range of unprocessed speech in SSN is a result of the steep slope of the corresponding psychometric function, which was optimized in this way to allow for accurate SRT measurements (Wagener *et al.*, 1999).

In summary, listening effort is well suited to evaluate speech perception and algorithm performance at high SNRs (well above SNRs of close-to-optimal intelligibility), while speech intelligibility is well suited to evaluate speech perception and algorithm performance at very low SNRs (well below SNRs of maximum listening effort). Taken together, speech intelligibility and listening effort measurements allowed us to measure speech perception and a significant algorithm benefit over an SNR range of about 28 dB (SSN) and 25 dB (cafeteria noise), indicating that the combined assessment is suitable to cover all ecologically relevant listening conditions. Although only a single NELE algorithm was tested in this study, there is no reason to assume why these general observations regarding evaluation methods should not be valid for other algorithms.

Acknowledgments

This research was supported by the German Research Foundation (DFG) funding the Research Unit FOR132 "Individualized Hearing Acoustics," the Cluster of Excellence 1077 "Hearing4All," and Grant No. RE 4160/1-1.

References and links

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**(6), 2801–2810.

Cooke, M., Mayo, C., and Valentini-Botinhao, C. (**2013**). "Intelligibility enhancing speech modifications: The Hurricane Challenge," in *Proceedings of Interspeech*, Lyon, France, pp. 3552–3556.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., and Kollmeier, B. (**2009**). "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," EURASIP J. Adv. Signal Process. **2009**, 298605.

Kleijn, W. B., Crespo, J. B., Hendriks, R. C., Petkov, P., Sauert, B., and Vary, P. (**2015**). "Optimizing speech intelligibility in a noisy environment: A unified view," IEEE Signal Process. Mag. **32**(2), 43–54.

Kleinbaum, D., Kupper, L., Nizam, A., and Rosenberg, E. (**2013**). *Applied Regression Analysis and Other Multivariable Methods*, 5th ed. (Cengage Learning, Boston, MA).

Krueger, M., Schulte, M., Zokoll, M., Wagener, K., Meis, M., Brand, T., and Holube, I. (**2017**). "Relation between listening effort and speech intelligibility in noise," Am. J. Audiol. **26**, 378–392.

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

Rennies, J., Drefs, J., Hülsmeier, D., Schepker, H., and Doclo, S. (**2017**). "Extension and evaluation of a near-end listening enhancement algorithm for listeners with normal and impaired hearing," J. Acoust. Soc. Am. **141**, 2526–2537.

Rennies, J., Schepker, H., Holube, I., and Kollmeier, B. (**2014**). "Listening effort and speech intelligibility in listening situations affected by noise and reverberation," J. Acoust. Soc. Am. **136**, 2642–2653.

Sato, H., Morimoto, M., and Wada, M. (**2012**). "Relationship between listening difficulty rating and objective measures in reverberant and noisy sound fields for young adults and elderly persons," J. Acoust. Soc. Am. **131**, 4596–4605.

Sauert, B., and Vary, P. (**2012**). "Near-end listening enhancement in the presence of bandpass noises," in *Proceedings of the ITG Conference on Speech Communication*, Braunschweig, Germany, pp. 195–198.

Schepker, H., Haeder, K., Rennies, J., and Holube, I. (**2016**). "Perceived listening effort and speech intelligibility in reverberation and noise for hearing-impaired listeners," Int. J. Audiol. **55**, 738–747.

Schepker, H., Rennies, J., and Doclo, S. (**2015**). "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," J. Acoust. Soc. Am. **138**, 2692–2706.

Smeds, K., Wolters, F., and Rung, M. (**2015**). "Estimation of signal-to-noise ratios in realistic sound scenarios," Am. J. Audiol. **26**, 183–196.

Taal, C. H., Hendriks, R. C., and Heusdens, R. (**2014**). "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," Comput. Speech Lang. **28**, 858–872.

Taal, C. H., and Jensen, J. (**2013**). "SII-based speech preprocessing for intelligibility improvement in noise," in *Proceedings Interspeech*, Lyon, France, pp. 3582–3586.

Tang, Y., and Cooke, M. (**2011**). "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proceedings of Interspeech*, Florence, Italy, pp. 345–348.

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 German sentence test Part III: Evaluation of the Oldenburg sentence test"), Z. Audiol. **38**(3), 86–95.

Zorila, T.-C., and Stylianou, Y. (**2014**). "On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement," in *Proceedings of Interspeech*, Singapore, pp. 2050–2054.