Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index

Henning Schepker[a] and Jan Rennies
Project Group Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology IDMT, D-26129 Oldenburg, Germany

Simon Doclo[b]
Signal Processing Group, Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg, D-26111 Oldenburg, Germany

(Received 6 March 2015; revised 27 July 2015; accepted 20 September 2015; published online 3 November 2015)

In many speech communication applications, such as public address systems, speech is degraded by additive noise, leading to reduced speech intelligibility. In this paper a pre-processing algorithm is proposed that is capable of increasing speech intelligibility under an equal-power constraint. The proposed *AdaptDRC* algorithm comprises two time- and frequency-dependent stages, i.e., an amplification stage and a dynamic range compression stage that are both dependent on the Speech Intelligibility Index (SII). Experiments using two objective measures, namely, the extended SII and the short-time objective intelligibility measure (STOI), and a formal listening test were conducted to compare the *AdaptDRC* algorithm with a modified version of a recently proposed algorithm in three different noise conditions (stationary car noise and speech-shaped noise and non-stationary cafeteria noise). While the objective measures indicate a similar performance for both algorithms, results from the formal listening test indicate that for the two stationary noises both algorithms lead to statistically significant improvements in speech intelligibility and for the non-stationary cafeteria noise only the proposed *AdaptDRC* algorithm leads to statistically significant improvements. A comparison of both objective measures and results from the listening test shows high correlations, although, in general, the performance of both algorithms is overestimated.
 [http://dx.doi.org/10.1121/1.4932168]

[MAH] Pages: 2692–2706

I. INTRODUCTION

In many speech communication applications, such as public address systems, mobile telephony, or any audio device with speech output, a high quality of communication needs to be provided. To achieve a high communication quality, first a high speech intelligibility must be ensured. However, in typical communication situations speech is degraded by additive noise and/or reverberation. The influence of these disturbances may lead to reduced speech intelligibility and increased listening effort (Beutelmann and Brand, 2006; Bronkhorst, 2000; George *et al.*, 2010; Morimoto *et al.*, 2004; Rennies *et al.*, 2014). A straightforward solution to obtain a high speech intelligibility in noise is to increase the speech level to achieve a good signal-to-noise ratio (SNR). A similar phenomenon can be observed in human speech production in noisy environments referred to as the Lombard effect (Van Summers *et al.*, 1988). The Lombard effect is characterized not only by an increase in speech level but it has also been found that, among several other modifications, the spectral tilt of speech is reduced (Lu and Cooke, 2008; Van Summers *et al.*, 1988). Subjective listening tests comparing speech intelligibility of Lombard speech and normal speech have shown that Lombard speech is better intelligible compared to normal speech even when presented at the same physical level (Lu and Cooke, 2008).

Although the aforementioned broadband amplification can easily be implemented, it may lead to an overload of the amplification system and/or loudspeaker or to unpleasantly high sound levels. Therefore, it is desirable to design algorithms that are able to increase speech intelligibility in such a way that they maintain equal powers of both the unprocessed signal and the processed signal.

In general, algorithms that modify the speech signal prior to presentation can be classified into one of the following five categories: (1) algorithms that change the frequency characteristics by applying spectral modification techniques (e.g., Brouckxon *et al.*, 2008; Kleijn *et al.*, 2015; Sauert and Vary, 2010b, 2012; Taal and Jensen, 2013; Taal *et al.*, 2014), (2) algorithms that explicitly employ non-linear modifications such as dynamic range compression (DRC, e.g., Licklider and Pollack, 1948; Niederjohn and Grotelueschen, 1976; Zorila *et al.*, 2012; Zorila and Stylianou, 2014), (3) algorithms that selectively enhance signal components (e.g., Arai *et al.*, 2010; Ortega and Huckvale, 2000; Skowronski and Harris, 2006), (4) algorithms that aim at enhancing the modulation of speech signals (e.g., Kusumoto *et al.*, 2005),

[a]Current address: Signal Processing Group, Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg, D-26111 Oldenburg, Germany. Electronic mail: henning.schepker@uni-oldenburg.de
[b]Also at: Project Group Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology IDMT, D-26129 Oldenburg, Germany.

and (5) algorithms that modify the time-scale of the speech signal (e.g., Tang and Cooke, 2011; Verhelst, 2000). In the following, we will focus on the first two categories, i.e., spectral modification and DRC.

In addition these algorithm can be classified into noise-adaptive algorithms, which take into account the characteristics of the near-end noise (e.g., Brouckxon *et al.*, 2008; Sauert and Vary, 2010b, 2012; Taal *et al.*, 2014; Tang and Cooke, 2011), and noise-independent algorithms (e.g., Kusumoto *et al.*, 2005; Licklider and Pollack, 1948; Niederjohn and Grotelueschen, 1976; Ortega and Huckvale, 2000; Zorila *et al.*, 2012; Zorila and Stylianou, 2014).

One of the first analyses of signal pre-processing algorithms (i.e., processing prior to presentation) was made by Licklider and Pollack (1948). They investigated the effects of high-pass and low-pass filtering as well as peak-clipping and several combinations of these on speech intelligibility in quiet. They found only minor effects on speech intelligibility for most of the considered combinations, while intelligibility was severely degraded when low-pass filtering was followed by peak-clipping.

Recently, Sauert and Vary (2010b) proposed an algorithm, in which time- and frequency-dependent amplification of the speech signal is carried out aiming to maximize the Speech Intelligibility Index (SII). Although maximizing the SII, they found that their approach suffered from spectral adaptation to the noise characteristics. This is especially undesired for noises with band-pass characteristics where the speech signal after processing exhibits the same band-pass characteristics as the noise and is hence strongly distorted. To circumvent this problem, Sauert and Vary (2012) proposed to use a transition between an SII-based weighting proposed in Sauert and Vary (2010b) and unity-weighting. Other algorithms have also been proposed that aim at maximizing other objective measures such as a low-complexity distortion measure (Taal *et al.*, 2014), a loudness metric (Shin and Kim, 2007), a glimpse portion measure (Tang and Cooke, 2012), or the speech magnitude distortions (Crespo and Hendriks, 2013).

In the class of non-linear modification algorithms Niederjohn and Grotelueschen (1976) proposed to use high-pass filtering followed by static rapid amplitude compression. They reported an increase in speech intelligibility in white noise over the unprocessed signal at the same SNR. The idea of combining frequency-shaping, i.e., linear filtering, and DRC was recently adopted by Zorila *et al.* (2012). They used a speech signal-dependent frequency-shaping and a static broadband DRC scheme and reported improvements in speech-shaped noise (SSN) and a competing speaker for three different SNRs.

The previously mentioned approaches use either only frequency-shaping or in addition a static broadband compression characteristic. While these approaches yield increased intelligibility over a wide range of SNRs, they also modify the speech signal in conditions of good speech intelligibility. This may be disadvantageous, since any modification of the speech signal may also lead to a degradation in perceived speech quality. Therefore, in the recently proposed *AdaptDRC* algorithm (Schepker *et al.*, 2013) a processing scheme is employed that combines a time- and frequency-dependent frequency-shaping and a time- and frequency-dependent dynamic range compression characteristic, preserving the original speech signal in cases of good intelligibility.

In a recent study, Cooke *et al.* (2013a) compared a large variety of pre-processing algorithms in a subjective listening test. Within their study also the algorithms of Sauert and Vary (2010a) and Zorila *et al.* (2012) as well as Lombard speech were included. They found that, for a stationary SSN, despite their differences in processing (frequency-shaping and/or DRC), all algorithms that modified the clean speech signal improved speech intelligibility at equal output powers. However, for a non-stationary speech masker not all algorithms were able to improve speech intelligibility, but the DRC algorithm of Zorila *et al.* (2012) and Lombard speech showed largest improvements. The study of Cooke *et al.* (2013a) was extended in Cooke *et al.* (2013b) in the so-called 2013 Hurricane Challenge, where similar noise and SNR conditions were used but different algorithms were evaluated. Results indicated that algorithms that used a DRC stage showed large improvements in the considered stationary and non-stationary noises, while most gain amplification algorithms only improved speech intelligibility for stationary noises.

In this paper a more detailed description of the *AdaptDRC* algorithm of Schepker *et al.* (2013) is presented. Extensive evaluations using objective measures and a formal listening test in three different real-world noises are performed that provide insights into the performance of the *AdaptDRC* algorithm in stationary as a well as non-stationary noisy environments. Thus, the results presented by Cooke *et al.* (2013b) and Schepker *et al.* (2013) are extended using additional noises and different speech material. Furthermore, a correlation analysis is carried out to investigate the performance of the objective measures to predict speech intelligibility for signals modified by state-of-the-art pre-processing algorithms.

The remainder of this paper is organized as follows. In Sec. II the considered scenario and some important assumptions are discussed. In Sec. III the proposed *AdaptDRC* algorithm is described. In Sec. IV the proposed algorithm is evaluated using two objective measures (extended SII and the short-time objective intelligibility measure, STOI) and a formal listening test. In Sec. V both evaluation methods are compared by means of a correlation analyses and the results are discussed.

II. SCENARIO

Consider the acoustic scenario depicted in Fig. 1. The unprocessed (clean) speech signal $s[k]$ at discrete time k is modified using the processing stage $W\{\cdot\}$ and the modified speech signal $\tilde{s}[k]$ is played back via a loudspeaker. A microphone picks up the disturbed speech signal $y[k]$, which is the mixture of the modified speech signal $\tilde{s}[k]$ convolved with the room impulse response $h[k]$ between the loudspeaker and the microphone and the additive noise disturbance $r[k]$, i.e.,

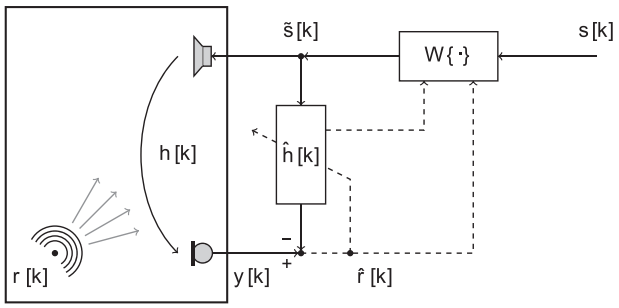

FIG. 1. Considered acoustic scenario.

$$y[k] = \tilde{s}[k] * h[k] + r[k], \tag{1}$$

where $*$ denotes convolution. An estimate of the noise signal $\hat{r}[k]$ can be obtained by using, e.g., adaptive filtering techniques to model the room impulse response $\hat{h}[k]$ (Haensler and Schmidt, 2008) and subtracting $\tilde{s}[k] * \hat{h}[k]$ from the microphone signal $y[k]$. Using the estimated noise signal $\hat{r}[k]$, the estimated room impulse response $\hat{h}[k]$, and the clean speech signal $s[k]$, the processed speech signal $\tilde{s}[k]$ is then computed as

$$\tilde{s}[k] = W\{s[k], \hat{r}[k], \hat{h}[k]\}. \tag{2}$$

The goal of a signal pre-processing algorithm is to derive a processing stage $W\{\cdot\}$ that enhances the intelligibility of $\tilde{s}[k] + r[k]$ compared to $s[k] + r[k]$ under an equal power constraint, i.e., the power of $\tilde{s}[k]$ is equal to the power of $s[k]$. To provide an insight into the optimal performance of the proposed algorithm, estimation errors will be neglected in this paper and no reverberation is assumed to be present (i.e., $\hat{h}[k] = h[k] = \delta[k]$) and hence we assume a perfect noise estimate to be available (i.e., $\hat{r}[k] = r[k]$). An overview of the notation used in this paper is provided in Table I.

III. PRE-PROCESSING ALGORITHMS

In this section a detailed description of the proposed *AdaptDRC* pre-processing algorithm as well as the implementation of a state-of-the-art algorithm of Sauert and Vary (2012) used as a reference algorithm is provided. In Sec. III A the used processing framework and some general definitions are provided. In Sec. III B the proposed *AdaptDRC* pre-processing algorithm is described in detail and our modified implementation of the reference algorithm by Sauert and Vary (2012) referred to as *ModSau* in the remainder is discussed as a special case of the amplification stage of the proposed *AdaptDRC* algorithm.

A. Processing framework and definitions

The signal $s[k]$ is first split into N subband signals $s_n[k]$, $n = 1, \ldots, N$, using a real-valued non-decimated filterbank. In our implementation, we have used an all-pass filterbank based on doubly-complementary IIR filters (Regalia *et al.*, 1987), splitting the signals into $N = 8$ octave-bands with center frequencies ranging from 125 Hz to 16 kHz. Each subband signal $s_n[k]$ is framed into non-overlapping blocks of length M with the lth block denoted as $s_n^l[m] = s_n[lM + m]$, $m = 0, \ldots, M-1$. The speech power of the lth block in the nth subband is equal to

$$\phi_{s,n}[l] = \frac{1}{M} \sum_{m=0}^{M-1} \left(s_n^l[m]\right)^2. \tag{3}$$

TABLE I. Overview of symbols and parameters used in the algorithmic description.

Parameter	Description	Parameter	Description
k	Discrete time index	$s[k]$	Speech signal
l	Discrete block index	$\tilde{s}[k]$	Processed speech signal
m	Discrete sample index in block	$r[k]$	Noise signal
M	Block length	$\hat{r}[k]$	Estimate of $r[k]$
n	Subband index	$h[k]$	Room impulse response
N	Number of subbands	$\phi_s[l]$	Broadband speech power
i_n	Band importance function	$\phi_{s,n}[l]$	Subband speech power
u_n	Stand. equiv. speech spectrum level	$e_n[l]$	Equiv. speech spectrum level
v_n	Subband dependent weighting	$d_n[l]$	Equiv. dist. spectrum level
α_a	Attack smoothing constant	$\hat{SII}[l]$	Estimated SII
α_r	Release smoothing constant	$\hat{a}(e_n[l], d_n[l])$	Approx. audibility function
ν	Conversion constant from dB FS to dB SPL	$q(e_n[l], d_n[l])$	Mapping function
$cr_{(\max)}$	Maximum compression ratio	$\hat{g}(d_n[l])$	Speech distortion function
α_b	IOC parameter smoothing constant	$w_n[l]$	Amplification gain
α_p	Compressive gain smoothing constant	$\theta[l]$	Transition parameter
α_L	Broadband level smoothing constant	$p(\boldsymbol{\lambda}_n, (\bar{s}_n[k])^2)$	Compressive gain function
		$\bar{p}(\bar{\boldsymbol{\lambda}}_n, (\bar{s}_n[k])^2)$	Smoothed compressive gain function
		$\boldsymbol{\lambda}_n[l]$	IOC parameter vector
		$\bar{\boldsymbol{\lambda}}_n[l]$	Smoothed IOC parameter vector
		$\bar{s}_n[k]$	Estimated speech envelope
		$\gamma_{n,i}[l]$	i-th input power of IOC parameter vector
		$\xi_{n,i}[l]$	i-th output power of IOC parameter vector
		$cr_n[l]$	Time-varying compression ratio

The broadband input speech power $\phi_s[l]$ in the lth block is defined as the sum over all subband powers. Similarly, the broadband output speech power $\phi_{\tilde{s}}[l]$ is calculated from the output speech signal $\tilde{s}[k]$. Furthermore, the equivalent speech spectrum level is defined as (ANSI, 1997)

$$e_n[l] = 10\log_{10}\left(\frac{\phi_{s,n}[l]}{\Delta f_n}\right) = 10\log_{10}\left(\frac{1}{M\Delta f_n}\sum_{m=0}^{M-1}\left(s_n^l[m]\right)^2\right), \quad (4)$$

with Δf_n the bandwidth of the nth subband. Similarly, the equivalent disturbance spectrum level is defined as

$$d_n[l] = 10\log_{10}\left(\frac{1}{M\Delta f_n}\sum_{m=0}^{M-1}\left(\hat{r}_n^l[m]\right)^2\right), \quad (5)$$

where $\hat{r}_n^l[m]$ is defined similarly as $s_n^l[m]$, i.e., $\hat{r}_n^l[m] = \hat{r}_n[lM+m]$, $m = 0, \ldots, M-1$.

B. Proposed algorithm: AdaptDRC

The proposed *AdaptDRC* algorithm is depicted in Fig. 2. It combines a time- and frequency-dependent amplification stage with a time- and frequency-dependent DRC stage, where both the amplification and the DRC are either functions of the short-term SII estimate or employ intermediate results from the short-term SII estimation. In the following first the short-term SII estimation procedure and second the amplification stage and the DRC stage as well as the gain-smoothing procedure will be described.

1. Short-term SII estimation

The short-term SII estimation procedure is adopted from Sauert and Vary (2010b) but has been slightly modified. The short-term SII in the lth block is estimated from the clean speech signal $s[k]$ and the estimated noise signal $\hat{r}[k]$ as

$$\hat{SII}[l] = \sum_{n=1}^{N} i_n \hat{a}(e_n[l], d_n[l]), \quad (6)$$

where i_n denotes the band importance function of the nth subband according to ANSI (1997) and $\hat{a}(e_n[l], d_n[l])$ approximates the audibility function defined in ANSI (1997), which estimates the audibility of the speech signal relative to the noise signal as

$$\hat{a}(e_n[l], d_n[l]) = q(e_n[l], d_n[l])\hat{g}(d_n[l]). \quad (7)$$

The function $q(e_n[l], d_n[l])$ is a mapping function of the SNR according to

$$q(e_n[l], d_n[l]) = \max\left\{\min\left\{\frac{e_n[l] - d_n[l] + 15\,\text{dB}}{30\,\text{dB}}, 1\right\}, 0\right\} \quad (8)$$

and the function $\hat{g}(d_n[l])$ describes the speech distortion for louder levels. Assuming the particular case of $e_n[l] = d_n[l] + 15$ dB, maximizing $q(e_n[l], d_n[l])$ in Eq. (8), $\hat{g}(d_n[l])$ may be approximated as (Sauert and Vary, 2010b)

$$\hat{g}(d_n[l]) = \min\left\{1 - \frac{d_n[l] + 15\,\text{dB} - u_n - 10\,\text{dB}}{160\,\text{dB}}, 1\right\}, \quad (9)$$

which depends on the so-called standardized equivalent speech spectrum level u_n, which has been defined in ANSI (1997) for a broadband speech level of 62.35 dB sound pressure level (SPL) for normal speech. Note that $\hat{a}(e_n[l], d_n[l])$ in Eq. (7) differs from the approximation of the audibility function proposed by Sauert and Vary (2010b) in that it takes into account the flooring and ceiling effect of the function $q(e_n[l], d_n[l])$ in Eq. (8), which was neglected in Sauert and

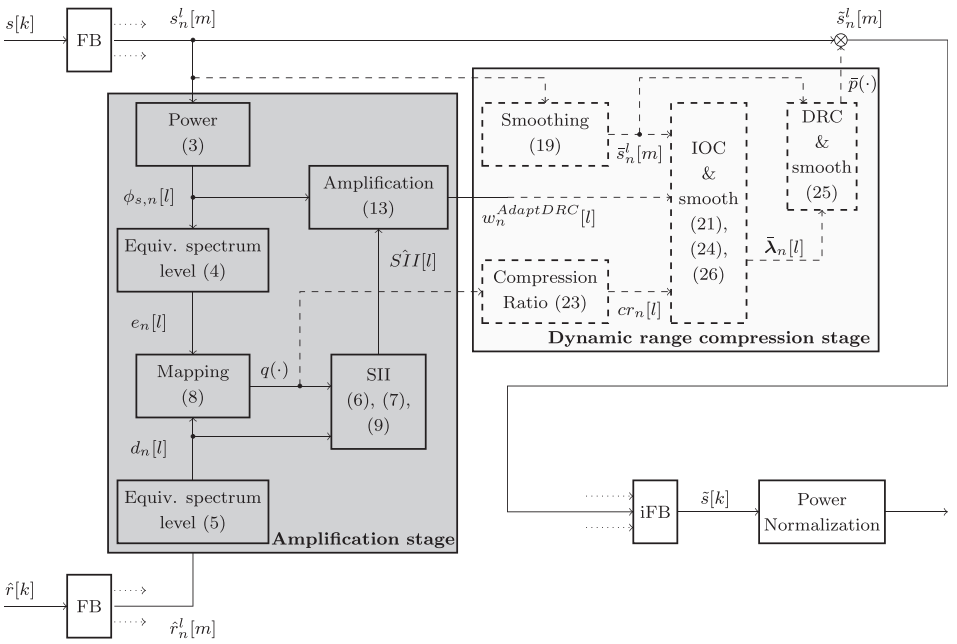

FIG. 2. Detailed block diagram of the *AdaptDRC* algorithm.

Vary (2010b). Additionally, by using the approximation $\hat{g}(d_n[l])$ in Eq. (9) it is assumed that $e_n[l] = d_n[l] + 15\,\text{dB} < u_n + 170\,\text{dB}$, which can be assumed to be valid in nearly all conditions. In conditions that would violate this assumption the equivalent speech spectrum level $e_n[l]$ in a particular subband would be at least 155 dB larger than the standardized equivalent speech spectrum level u_n, which is highly unlikely in realistic acoustic scenarios.

In addition, note that for the original definition of the SII in ANSI (1997) only octave-bands in the range of 250 Hz–8 kHz are considered. Since, in our implementation, we have also taken into account the lowest octave-band with center frequency 125 Hz, the values for i_1, i_2, u_1, and u_2 (i.e., the band importance function and the standardized equivalent speech spectrum levels in the two lowest octave-bands) were slightly changed. The values were chosen as $i_1 = 0.0083$, $i_2 = 0.0534$, $u_1 = 28.60\,\text{dB}$ and $u_2 = 34.75$ dB, being a trade-off between the original octave-band values and the values defined for third-octave-bands. An overview on the used band importance function and the standardized equivalent speech spectrum levels is depicted in Table II.

2. Amplification

In this section, first, the application of the amplification gain and its general description are provided. Second, the amplification gain of the proposed *AdaptDRC* algorithm (cf. Fig. 2) and the amplification gain of our implementation of the algorithm by Sauert and Vary (2012) are presented. The aim of the amplification stage for both algorithms is to provide a transition between modification of the speech signal in severe disturbance and no modification in low disturbance of the speech signal. The amplification gain $w_n[l]$ is applied to each samples of a block of samples of the input speech signal, i.e.,

$$\tilde{s}_n^l[m] = w_n[l]s_n^l[m], \quad m = 1, \ldots, M. \tag{10}$$

To achieve a trade-off between spectral modification and no spectral modification, a similar amplification gain as proposed by Sauert and Vary (2012) is used. This general amplification gain is defined as

$$w_n[l] = \sqrt{\frac{v_n^{1-\theta[l]}(\phi_{s,n}[l])^{\theta[l]}}{\sum_{\lambda=1}^{N} v_\lambda^{1-\theta[l]}(\phi_{s,\lambda}[l])^{\theta[l]}} \cdot \frac{\phi_s[l]}{\phi_{s,n}[l]}}, \tag{11}$$

where v_n is a subband dependent weighting with $\sum_{n=1}^{N} v_n = 1$ and $\theta[l]$ is a time-dependent transition parameter that can take values in the range $0 \le \theta[l] \le 1$. From Eq. (11) it can be shown that $\sum_{n=1}^{N} w_n^2[l]\phi_{s,n}[l] = \phi_s[l]$. Hence, Eq. (11) does preserve the broadband power of the input. In general, Eq. (11) has the following properties:

$$\begin{cases} \theta[l] = 0 \rightarrow w_n[l] = \sqrt{v_n \dfrac{\phi_s[l]}{\phi_{s,n}[l]}} \rightarrow \phi_{\tilde{s},n}[l] = v_n\phi_s[l], & \text{(12a)} \\ \theta[l] = 1 \rightarrow w_n[l] = 1. & \text{(12b)} \end{cases}$$

Hence, Eq. (11) provides a trade-off between distributing the broadband speech power $\phi_s[l]$ according to the weighting defined by v_n for $\theta[l] = 0$ and unity weighting for $\theta[l] = 1$.

a. AdaptDRC. Since speech signals in general contain more energy in lower frequency subbands compared to higher frequency subbands, on the one hand the amplification stage of the *AdaptDRC* algorithm aims at uniformly distributing the speech signal power over all subbands in case of low (predicted) speech intelligibility, i.e., for $\hat{SII}[l] \rightarrow 0$. On the other hand, the speech signal should not be altered to avoid any distortions in the case of high predicted speech intelligibility, i.e., for $\hat{SII}[l] \rightarrow 1$. Therefore, the subband dependent weighting is chosen as $v_n = 1/N$. To control the trade-off between distributing the speech signal power and avoiding distortions, we chose $\theta[l] = \hat{SII}[l]$ in (11), where $\hat{SII}[l]$ is defined in (6). Hence, the amplification gain of the *AdaptDRC* algorithm is defined as

$$w_n^{AdaptDRC}[l] = \sqrt{\frac{\phi_{s,n}^{\hat{SII}[l]}[l]}{\sum_{\lambda=1}^{N} \phi_{s,\lambda}^{\hat{SII}[l]}[l]} \cdot \frac{\phi_s[l]}{\phi_{s,n}[l]}}. \tag{13}$$

From Eq. (13) one observes the following properties:

$$\begin{cases} \hat{SII}[l] = 0 \rightarrow w_n^{AdaptDRC}[l] = \sqrt{\dfrac{1}{N}\dfrac{\phi_s[l]}{\phi_{s,n}[l]}} \rightarrow \phi_{\tilde{s},n}[l] = \dfrac{1}{N}\phi_s[l], & \text{(14a)} \\ \hat{SII}[l] = 1 \rightarrow w_n^{AdaptDRC}[l] = 1. & \text{(14b)} \end{cases}$$

TABLE II. Modified band importance functions and the modified standardized equivalent speech spectrum levels as a function of octave-band center frequency as used in the SII estimation.

f_n/Hz	125	250	500	1000	2000	4000	8000	16 000
i_n	0.0083	0.0534	0.1671	0.2373	0.2648	0.2142	0.0549	0.0
u_n/dB	28.60	34.75	34.75	25.01	17.32	9.33	1.13	0.00

Hence, a uniform distribution of the speech signal power $\phi_s[l]$ is achieved for $\hat{SII}[l] = 0$ and for $\hat{SII}[l] = 1$ the resulting gain is $w_n^{AdaptDRC}[l] = 1$, obtaining the desired properties. Note that the application of the amplification gain $w_n^{AdaptDRC}[l]$ in general leads to an increased speech power in higher frequency subbands, while in lower frequency subbands the output speech power is reduced.

b. Implementation of Sauert and Vary (2012). In contrast to the amplification gain of the proposed *AdaptDRC* algorithm, Sauert and Vary (2012) proposed to set the subband dependent weighting v_n according to the band importance function as defined in ANSI (1997) and the function $\hat{g}(d_n[l])$ in Eq. (9), i.e., $v_n = i_n \hat{g}(d_n[l])$ and used a heuristically chosen parameter $\theta[l]$. It was found that their particular choice of $\theta[l]$ resulted in a good compromise for the performance in different noise conditions. In our implementation of the algorithm proposed in Sauert and Vary (2012) we have modified the transition parameter $\theta[l]$. The transition parameter is chosen in a non-heuristic way based on the transformation of estimated SII values to the speech intelligibility of words as proposed by Beutelmann and Brand (2006), i.e.,

$$\theta[l] = \frac{0.0204}{0.01996 + e^{-20\cdot\hat{SII}[l]}} - 0.01996, \tag{15}$$

where $\hat{SII}[l]$ is defined in Eq. (6). Although the choice of $\theta[l]$ is different from Sauert and Vary (2012), no major differences between the two parameters were found when the algorithm output was compared by objective quality measures and informal listening tests. The amplification gain of the *ModSau* algorithm is thus defined as

$$w_n^{ModSau}[l] = \sqrt{\frac{\left(i_n \hat{g}(d_n[l])\right)^{1-\theta[l]} \phi_{s,n}^{\theta[l]}[l]}{\sum_{\lambda=1}^{N} \left(i_\lambda \hat{g}(d_\lambda[l])\right)^{1-\theta[l]} \phi_{s,\lambda}^{\theta[l]}[l]} \cdot \frac{\phi_s[l]}{\phi_{s,n}[l]}}, \tag{16}$$

where $\theta[l]$ is defined in Eq. (15). For the *ModSau* algorithm one observes the following properties:

$$\begin{cases} \theta[l] = 0 \rightarrow w_n^{ModSau}[l] = \sqrt{\left(i_n \hat{g}(d_n[l])\right) \dfrac{\phi_s[l]}{\phi_{s,n}[l]}} \rightarrow \phi_{\tilde{s},n}[l] = \left(i_n \hat{g}(d_n[l])\right) \phi_s[l], & \text{(17a)} \\ \theta[l] = 1 \rightarrow w_n^{ModSau}[l] = 1. & \text{(17b)} \end{cases}$$

Hence, for $\theta[l] = 0$ a distribution of the speech signal power $\phi_s[l]$ according to the band importance function i_n of the SII and $\hat{g}(d_n[l])$ is obtained and for $\theta[l] = 1$ the resulting gain is $w_n^{ModSau}[l] = 1$. According to Sauert and Vary (2010b), this is optimal with respect to the SII under the equal power constraint for $\theta[l] = 0$, but has to be deemed sub-optimal with respect to the SII for $\theta[l] > 0$. Nevertheless, the performance as evaluated by the SII is comparable to using the algorithm proposed by Sauert and Vary (2010b) aiming to optimize the SII under an equal power constraint (Sauert and Vary, 2012).

3. Dynamic range compression

The DRC stage of the *AdaptDRC* algorithm (cf. Fig. 2) aims at amplifying low-level signals that are assumed to be not well audible and attenuating high-level signals that are assumed to be well audible.

In general, in multiband DRC algorithms the processed speech signal $\tilde{s}_n[k]$ in the nth subband is computed as

$$\tilde{s}_n[k] = s_n[k] p(\boldsymbol{\lambda}_n, (\bar{s}_n[k])^2), \quad m = 0, \ldots, M-1, \tag{18}$$

where $p(\boldsymbol{\lambda}_n, (\bar{s}_n[k])^2)$ is a compressive gain function which computes a gain at each sample k by evaluating the input-output-characteristic (IOC) defined by the parameter vector $\boldsymbol{\lambda}_n$ for the input speech power $\bar{s}_n^2[k]$. The input speech power is estimated from the envelope of the nth subband signal, i.e.,

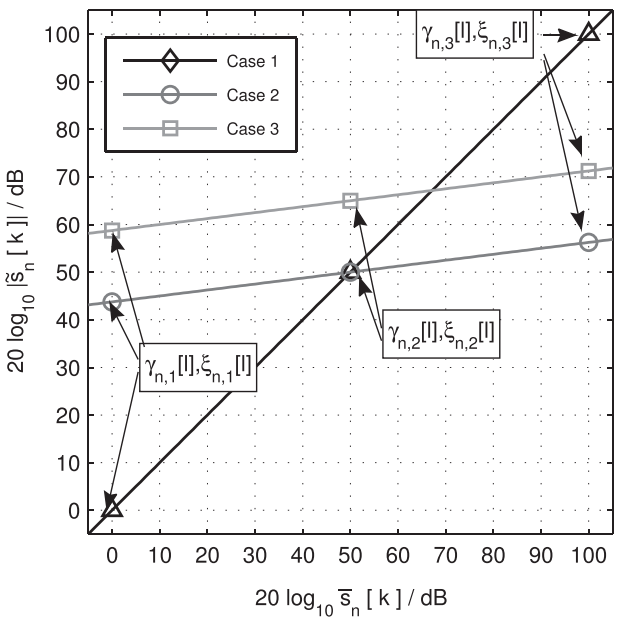

FIG. 3. Exemplary IOC for three different cases. Symbols indicate the points defined by the input and output powers $\gamma_{n,i}[l]$ and $\xi_{n,i}[l]$ and lines indicate $p(\boldsymbol{\lambda}_n[l], (\bar{s}_n[k])^2)$. It is assumed that $\phi_{s,n}[l] = 10^{50/10}$ and $cr_{\max} = 8$. The following cases are shown: case 1: $cr_n[l] = 1$; case 2: $cr_n[l] = cr_{\max}$; case 3: $w_n^{AdaptDRC} = 10^{15/20}$ and $cr_n[l] = cr_{(\max)}$.

$$\bar{s}_n[k] = \begin{cases} \alpha_a \bar{s}_n[k-1] + (1-\alpha_a)|s_n[k]| \\ \qquad \text{if} \quad |s_n[k]| \geq \bar{s}_n[k-1], \\ \alpha_r \bar{s}_n[k-1] + (1-\alpha_r)|s_n[k]| \\ \qquad \text{if} \quad |s_n[k]| < \bar{s}_n[k-1], \end{cases} \tag{19}$$

where α_a and α_r are attack and release smoothing constants. The IOC parameter vector $\boldsymbol{\lambda}_n$ usually contains a set of several input and output powers $\gamma_{n,i}$ and $\xi_{n,i}$. Figure 3 depicts exemplary IOCs for different parameter vectors $\boldsymbol{\lambda}_n$ defined by three input and output powers ($i = 1, 2, 3$).

In DRC algorithms typically fixed compression ratios are applied to either the broadband signal (Zorila *et al.*, 2012) or the subband signals. However, in case $s_n[k]$ is already well audible, DRC can lead to signal degradation that may harm the perceived quality or even reduce speech intelligibility.

Therefore, the novelty of the proposed *AdaptDRC* algorithm in comparison to previously proposed algorithms (Niederjohn and Grotelueschen, 1976; Zorila *et al.*, 2012) is the fact that the IOC changes for each block l, depending on the speech and noise characteristics. In the *AdaptDRC* algorithm three input and output powers are used to define the IOC, i.e.,

$$\boldsymbol{\lambda}_n[l] = [\,\gamma_{n,1}[l]\ \gamma_{n,2}[l]\ \gamma_{n,3}[l]\ \xi_{n,1}[l]\ \xi_{n,2}[l]\ \xi_{n,3}[l]\,], \tag{20}$$

with

$$\begin{aligned} \gamma_{n,1}[l] &= 1, \\ \gamma_{n,2}[l] &= \phi_{s,n}[l], \\ \gamma_{n,3}[l] &= \nu, \end{aligned} \tag{21}$$

and

$$\begin{aligned} \xi_{n,1}[l] &= (\phi_{s,n}[l])^{1-1/cr_n[l]}, \\ \xi_{n,2}[l] &= \phi_{s,n}[l], \\ \xi_{n,3}[l] &= (\phi_{s,n}[l])^{1-1/cr_n[l]} \nu^{1/cr_n[l]}, \end{aligned} \tag{22}$$

where ν is a conversion constant from dB FS to dB SPL chosen to be $\nu = 10^{(100/10)}$ and $cr_n[l]$ is the time-varying compression ratio. The time-varying compression ratio is computed independently for each subband and employs the SNR mapping function of the SII $q(e_n[l], d_n[l])$ as defined in Eq. (8), which is an intermediate step in the SII calculation, to provide a transition between maximum compression and no compression. It is defined as

$$cr_n[l] = \max\{cr_{(\max)} \cdot (1 - q(e_n[l], d_n[l])), 1\}, \tag{23}$$

where $cr_{(\max)}$ is assumed to be larger than 1 and denotes the maximum compression ratio. For subband SNRs lower than or equal to -15 dB, $q(e_n[l], d_n[l]) = 0$, such that $cr_{(\max)}$ will be applied, while for subband SNRs larger than or equal to $+15$ dB, $q(e_n[l], d_n[l]) = 1$, such that no compression will be applied, hence achieving the desired property of no modification in case of good speech intelligibility.

The amplification stage of the proposed *AdaptDRC* algorithm discussed in Sec. III B 2 can be directly incorporated into the IOC $\boldsymbol{\lambda}_n[l]$ by redefining $\xi_{n,i}[l]$ in Eq. (22) as

$$\begin{aligned} \xi_{n,1}[l] &= (\phi_{s,n}[l])^{1-1/cr_n[l]} w_n^{AdaptDRC}[l], \\ \xi_{n,2}[l] &= \phi_{s,n}[l] w_n^{AdaptDRC}[l], \\ \xi_{n,3}[l] &= (\phi_{s,n}[l])^{1-1/cr_n[l]} w_n^{AdaptDRC}[l] \nu^{1/cr_n[l]}. \end{aligned} \tag{24}$$

Figure 3 depicts exemplary IOCs for the proposed *AdaptDRC* assuming that $\phi_{s,n}[l] = 10^{50/10}$ and $cr_{\max} = 8$. Case 1 depicts a scenario where the SNR is sufficiently high (SNR $\geq +15$ dB) such that $cr_n[l] = 1$ and hence no compression is applied. For case 2, the opposite extremum is considered (SNR ≤ -15 dB) such that $cr_n[l] = cr_{(\max)} = 8$ leading to the desired compressive IOC where a large amplification is applied for low input levels and a strong attenuation is applied for high input levels. Case 3 assumes an additional amplification of $w_n^{AdaptDRC} = 10^{15/20}$ compared to case 2, resulting in a shift of the IOC.

4. Smoothing of gain functions and broadband normalization

Directly applying the gain derived from the IOC defined in Eqs. (21) and (24) may lead to noticeable artifacts due to large gain changes over time especially at block boundaries. To mitigate these artifacts, a two-stage smoothing procedure is applied. In a first step, each IOC parameter is smoothed independently, i.e.,

$$\bar{\lambda}_{n,j}[l] = \alpha_b \bar{\lambda}_{n,j}[l-1] + (1-\alpha_b)\lambda_{n,j}[l], \tag{25}$$

where $\lambda_{n,j}[l]$ is the jth element of the IOC parameter vector $\boldsymbol{\lambda}_n[l]$ and α_b is a smoothing constant. Note that the application of this smoothing procedure influences the appearance of the resulting IOC. Considering the exemplary IOCs in Fig. 3, the IOCs without smoothing can also be described by a straight line in the dB-domain. When smoothing is applied, only the connection between two neighboring elements of the IOC $\bar{\lambda}_n[l]$ can be described by a straight line. While this procedure leads to smooth changes of the IOC over time and reduces some major artifacts, it does not completely resolve the problem of larger gain changes at block boundaries. Thus, in a second step, the resulting gain is recursively smoothed with a smoothing factor α_p, i.e.,

$$\begin{aligned} \bar{p}(\bar{\boldsymbol{\lambda}}_n[l], (\bar{s}_n^l[m])^2) = \alpha_p \bar{p}(\bar{\boldsymbol{\lambda}}_n[l], (\bar{s}_n^l[m-1])^2) \\ + (1-\alpha_p) p(\bar{\boldsymbol{\lambda}}_n[l], (\bar{s}_n^l[m])^2). \end{aligned} \tag{26}$$

The processed subband signals $\tilde{s}_n[k]$ are then obtained by applying Eq. (26) to the input signal $s_n[k]$ in Eq. (18).

The application of Eq. (26) typically leads to changes in the broadband signal power and, therefore, does not satisfy the equal power constraint given in Sec. II. While in an offline procedure the signal could easily be rescaled after processing to satisfy this constraint, this is not possible in an online application. Therefore, after applying the inverse filterbank, a normalization gain is applied to the time-domain signal to yield approximately equal powers. This normalization gain is calculated by dividing the smoothed versions of the broadband input and output powers, i.e., $\sqrt{\bar{\phi}_s[l]/\bar{\phi}_{\tilde{s}}[l]}$,

where $\bar{\phi}_s[l]$ and $\bar{\phi}_{\tilde{s}}[l]$ are obtained by first-order recursive smoothing of the corresponding input speech power and output speech power, respectively, with smoothing constant α_L.

IV. EVALUATION

The proposed *AdaptDRC* algorithm was evaluated both using objective measures as well as by performing a formal subjective listening test. In both cases the speech material was taken from the Oldenburg Sentence Test (OLSA; Wagener *et al.*, 1999a,b; Wagener *et al.*, 1999c), which consists of 120 sentences spoken by a German male speaker. All sentences exhibit the same five word syntactical structure (*noun verb numeral adjective noun*) with ten possible alternatives each. The used sentences were generated by random combinations given the fixed syntactical structure, resulting in semantically unpredictable sentences. As additive disturbance three different noises were considered: (1) a stationary SSN that was generated by random superposition of the speech material, thus yielding the same long-term spectrum as the speech material, (2) a stationary (low-frequency) car noise, and (3) a non-stationary cafeteria noise, which comprises different speakers and some dish clanging. This allows to investigate the impact of noises that comparable in terms of their long-term spectra (SSN and the cafeteria noise) but differ in their temporal structure and noises that differ in their long-term spectra (SSN and the car noise) but are comparable in terms of their temporal structure. Hence, the results are expected to provide an indication about the influence of both spectral changes and temporal changes of the noise disturbance on the performance of both algorithms. For all evaluations, the parameters for the smoothing constants and block length given in Table III were used. Note that for clarity Table III depicts the corresponding integration time constants.

A. Objective evaluation

To quantitatively evaluate the performance of the proposed *AdaptDRC* algorithm, two different objective measures (ESII and STOI) were used that have shown high correlations with speech intelligibility in previous studies. Both the *ModSau* algorithm and the *AdaptDRC* algorithm and the unprocessed *Reference* condition were evaluated for a wide range of SNRs from −30 dB SNR to +30 dB SNR for all noises. All 120 sentences of the OLSA corpus were used as speech signals and an average speech level of 60 dB SPL was assumed, accordingly the noise was scaled to achieve the desired SNRs. Both measures are described briefly in Sec. IV A 1. The results are presented in Sec. IV A 2 and discussed in Sec. IV A 3.

TABLE III. Integration constants of the different smoothing parameters and the block length used in the evaluation.

Parameter	Value	Parameter	Value
τ_a	0.005 s	τ_p	0.250 s
τ_r	0.001 s	τ_b	0.250 s
τ_L	0.250 s	M	0.020 s
$cr_{(\max)}$	8		

1. Measures

To objectively predict the influence of the algorithms on speech intelligibility, the ESII measure (Rhebergen and Versfeld, 2005) and the STOI measure (Taal *et al.*, 2011) were used. Both objective measures have successfully been applied in previous studies (e.g., Taal *et al.*, 2014; Zorila *et al.*, 2012) to evaluate the impact of pre-processing algorithms.

The ESII measure (Rhebergen and Versfeld, 2005) can be interpreted as a sophisticated time-dependent SNR measure, taking into account the relative importance of different frequency regions for speech intelligibility and considering several properties related to speech perception, e.g., the smearing of speech at high levels and the upward-spread of masking as well as different time integration constants across frequency. The ESII measure essentially computes the SII in short time frames of speech and noise and averages the resulting time-dependent SII values, thus considering short-term fluctuations of both speech and noise.

The STOI measure (Taal *et al.*, 2011) is based on the correlation between the undisturbed speech signal and the disturbed speech signal. It is calculated in short time-frames and thus takes into account fluctuations of both the speech signal and the noise signal. Only the frequency range from 150 Hz to 4.3 kHz is taken into account. Both measures are computed as an index between 0 and 1, where 0 represents maximum disturbance and 1 represents minimum disturbance of the speech signal.

2. Results

Figure 4 shows the results of the objective evaluation for the ESII measure (top row) and the STOI measure (bottom row). The left column shows the results for the cafeteria noise, in the mid column the results for the SSN are shown, and the right column shows results for the car noise. The results show an increase in predicted speech intelligibility relative to the unprocessed *Reference* for both algorithms in all considered noises and for both objective measures. In general, for the ESII measure the proposed *AdaptDRC* algorithm shows a better performance than the *ModSau* algorithm at higher SNRs, while the *AdaptDRC* algorithm and the *ModSau* algorithm show almost the same performance at low SNRs, which is not observed in STOI. At higher SNRs the differences between both algorithms as measured by STOI tend to disappear for both stationary noises and small improvements for the *AdaptDRC* algorithm over the *ModSau* algorithm can be observed for the cafeteria noise.

3. Discussion

Two different algorithms were compared with respect to their performance by employing two objective measures that have shown high correlations to speech intelligibility in previous studies. Sauert and Vary (2012) evaluated their algorithm using the SII measure for a car noise and a speech

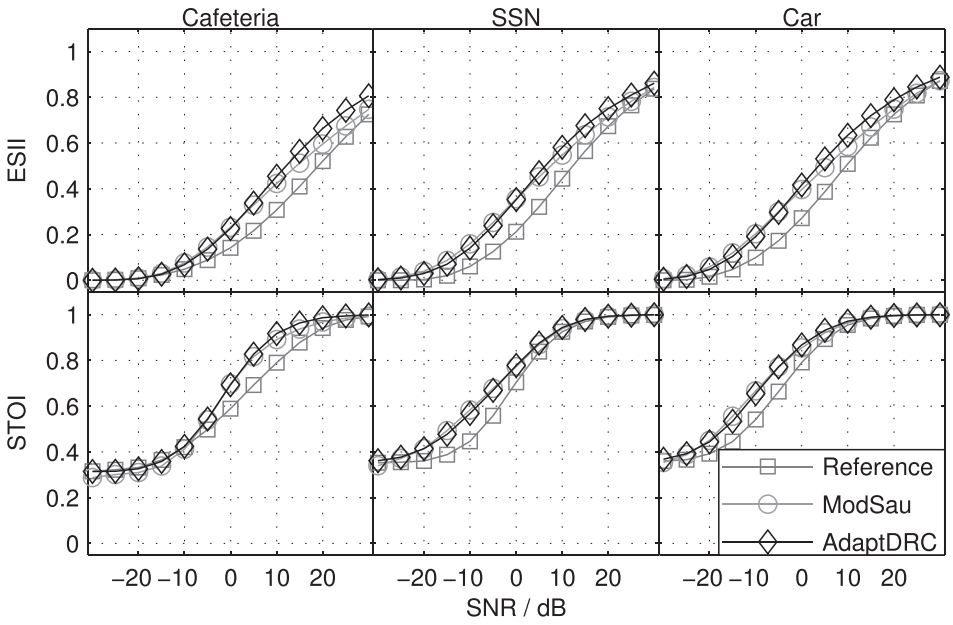

FIG. 4. Results from the objective evaluation using 120 sentences from OLSA speech material (Wagener *et al.*, 1999c) as a function of the SNR for the ESII measure (top row) and the STOI measure (bottom row). Results are shown for the cafeteria noise (left column), the SSN (mid column) and the car noise (right column).

babble noise. This is different from the present study where the ESII was used to evaluate the performance of both algorithm. For stationary noises the SII and ESII yield very similar results (Rhebergen *et al.*, 2006) when using an SSN to simulate the speech signal. However, in the present study the speech signals were used as an input to the ESII, hence, in the following only the improvement over the *Reference* is considered to compare the results from Sauert and Vary (2012) and the present study. Although the *ModSau* algorithm used in the present study differs from the original implementation of Sauert and Vary (2012), for the car noise they reported similar improvements over the unprocessed *Reference* compared to the improvements shown in Fig. 4. Furthermore, they reported improvements for the speech babble noise that are comparable to the results for the cafeteria noise in the present study. This leads to the conclusion that with respect to the evaluation using the ESII measure the proposed implementation and the original implementation of Sauert and Vary (2012) may be considered similar.

Both the proposed *AdaptDRC* algorithm as well as the *ModSau* algorithm show an increased performance in both objective measures when compared to the unprocessed *Reference*. At low SNRs, the *ModSau* algorithm shows about the same improvements in ESII values as the *AdaptDRC* algorithm. However, for larger SNRs the ESII shows an improved performance of the *AdaptDRC* algorithm over the *ModSau* algorithm. For the STOI measure these differences are not observed and both algorithms show similar results.

B. Subjective evaluation

The use of objective measures often provides a good indication about the performance of the algorithms and is therefore a valuable tool when designing algorithms. However, in some cases results from objective measures and formal listening tests indicate contradicting results, i.e., objective measures predict an increased speech intelligibility, while formal listening tests show a decreased speech intelligibility (Taal and Jensen, 2013) or vice versa. Hence, the impact of speech pre-processing algorithms on speech intelligibility can only be truly assessed using subjective listening tests. Therefore, a formal listening test was conducted to compare the performance in terms of speech intelligibility as measured by the number of correctly understood words for both algorithms. This section organizes as follows. In Sec. IV B 1 the used method is presented. In Sec. IV B 2 the results are presented and discussed in Sec. IV B 3.

1. Method

a. Subjects. The listening test was performed with eight normal-hearing subjects with pure-tone thresholds below 20 dB hearing level for all audiometric frequencies between 125 Hz and 8 kHz. The mean age of the subject group was 25.9 years with the youngest subject being 23 years and the oldest 28 years. The subjects participated voluntarily and were paid a small compensation for their time investment. None of the authors participated in the listening test.

b. Equipment. The measurements were conducted using a personal computer and MATLAB software. The stimuli were processed at the desired SNRs and stored on a hard-drive prior to the listening test. An RME Fireface UC Soundcard was used and the signals were presented via Sennheiser HD650 headphones in a soundproof booth. All stimuli were sampled at 44.1 kHz. The speech signals were presented at a level of 60 dB SPL, i.e., the time-average speech level was calibrated to 60 dB SPL, while the noise level was varied to achieve the desired SNRs.

c. Procedure. The focus of the formal listening tests was to measure the performance of the proposed algorithm over a wide range of SNRs and, hence, to cover condition ranging from low speech intelligibility to high speech intelligibility. Therefore, in a preliminary study with four of the eight subjects an adaptive procedure (Brand and Kollmeier, 2002) was applied to estimate the SNRs corresponding to 20%, 50%, and 80% correctly understood words for each of the three noises and the unprocessed *Reference* speech signals. The following SNRs were determined:

- Cafeteria noise: $\mathrm{SNR}_{20} = -14$, $\mathrm{SNR}_{50} = -10$, $\mathrm{SNR}_{80} = -6$ dB SNR.
- SSN: $\mathrm{SNR}_{20} = -11$, $\mathrm{SNR}_{50} = -9$, $\mathrm{SNR}_{80} = -7$dB SNR.
- Car noise: $\mathrm{SNR}_{20} = -18$, $\mathrm{SNR}_{50} = -16$, $\mathrm{SNR}_{80} = -14$ dB SNR.

These results indicate that SSN is the most difficult condition which can be explained by its large spectral overlap with the speech signal. For the car noise the opposite holds, due to its concentration of energy mostly in lower frequency regions.

The choice of these SNRs allowed to reliably fit psychometric functions for the *Reference* condition with the purpose of estimating the performance at other SNRs than the measured SNRs. From informal listening tests large improvements especially for low SNRs (corresponding to 20% and 50% speech intelligibility in the unprocessed *Reference* condition) were expected. For each combination of algorithm (*Reference*, *ModSau*, *AdaptDRC*) and noise (cafeteria noise, SSN, car noise) the simuli were presented at three different SNRs for each subject.

Aiming at reliably estimating psychometric functions for each subject, a semi-adaptive procedure for determining these three fixed SNRs for each combination and subject was employed in the listening test for the processed and unprocessed signals. For the sake of clarity, let $\Delta\mathrm{SNR} = \mathrm{SNR}_{50} - \mathrm{SNR}_{80}$, with SNR_{50} and SNR_{80} the SNRs corresponding to 50% and 80% speech intelligibility for the unprocessed speech signals as determined in the preliminary study, and let $SI(\mathrm{SNR}_\beta)$ denote the intelligibility measured at the SNR_β. For all subjects and combinations first the speech intelligibility at the SNR_{50} was measured. In order to allow for reliably fitting of the psychometric function, the subsequent SNRs for a specific combination were chosen based on the previous results for this combination. When speech intelligibility was considered large the next SNR decreased, whereas the next SNR was increased when speech intelligibility was considered small. More specifically, the following semi-adaptive procedure for choosing the SNRs was applied:

(1) Measure $SI(\mathrm{SNR}_{50})$.
(2) Choose the second SNR, $SNR_{(2)}$, as

$$\mathrm{SNR}_{(2)} = \begin{cases} \mathrm{SNR}_{50} - 2\Delta\mathrm{SNR} \text{ if } SI(\mathrm{SNR}_{50}) \geq 70\%, \\ \mathrm{SNR}_{50} + 2\Delta\mathrm{SNR} \text{ if } SI(\mathrm{SNR}_{50}) \leq 30\% \\ \text{randomly } \mathrm{SNR}_{20} \text{ or } \mathrm{SNR}_{80} \text{ otherwise.} \end{cases} \tag{27}$$

(3) Based on $SI(\mathrm{SNR}_{50})$ and $SI(\mathrm{SNR}_{(2)})$ choose the third SNR, $\mathrm{SNR}_{(3)}$, as

$$\mathrm{SNR}_{(3)} = \begin{cases} \mathrm{SNR}_{50} - 4\Delta\mathrm{SNR} & \text{if } SI(\mathrm{SNR}_{50}) \geq 70\% \text{ and } SI(\mathrm{SNR}_{(2)}) \geq 70\%, \\ \mathrm{SNR}_{50} - 3\Delta\mathrm{SNR} & \text{if } SI(\mathrm{SNR}_{50}) \geq 70\% \text{ and } 70\% > SI(\mathrm{SNR}_{(2)}) \geq 50\%, \\ \mathrm{SNR}_{50} - \Delta\mathrm{SNR} & \text{if } SI(\mathrm{SNR}_{50}) \geq 70\% \text{ and } SI(\mathrm{SNR}_{(2)}) < 50\%, \\ \mathrm{SNR}_{50} + 4\Delta\mathrm{SNR} & \text{if } SI(\mathrm{SNR}_{50}) \leq 30\% \text{ and } SI(\mathrm{SNR}_{(2)}) \leq 30\%, \\ \mathrm{SNR}_{50} + 3\Delta\mathrm{SNR} & \text{if } SI(\mathrm{SNR}_{50}) \leq 30\% \text{ and } 50\% \geq SI(\mathrm{SNR}_{(2)}) > 30\%, \\ \mathrm{SNR}_{50} + \Delta\mathrm{SNR} & \text{if } SI(\mathrm{SNR}_{50}) \leq 30\% \text{ and } SI(\mathrm{SNR}_{(2)}) > 50\%, \\ \mathrm{SNR}_{20} & \text{if } \mathrm{SNR}_{(2)} == \mathrm{SNR}_{80}, \\ \mathrm{SNR}_{80} & \text{if } \mathrm{SNR}_{(2)} == \mathrm{SNR}_{20}. \end{cases} \tag{28}$$

This procedure was applied for every combination of noise and algorithm, where the sequence of combinations was randomized. In addition, after one SNR of a particular combination of noise and algorithm was measured, a new combination was chosen randomly. In order to avoid any effect of training (Wagener *et al.*, 1999b), each subject was familiarized with the speech material prior to the listening test.

d. Statistical analyses. Statistical analyses were conducted using R statistics software. Shapiro-Wilk tests showed that not all data could be assumed to be normally distributed. Therefore, an aligned rank transform (ART; Wobbrock *et al.*, 2011) was employed before using standard analyses of variance (ANOVA) procedures. For a two-factor data-set the ART produces three different data-sets, one for each of the main factors and a third one for the interaction of the factors. Each data-set is then assumed to only depend on either one of the main factors or the interaction. For each of these data-sets, a two-way ANOVA is carried out while only the results for the dependent factor may be interpreted (Wobbrock *et al.*, 2011). *Post hoc* analyses were carried out (if appropriate) using the student t-test. Differences were assumed to be significant for p-values smaller than 0.05. The level of significance was adjusted using Bonferroni correction when multiple comparisons were conducted.

2. Results

The results of the formal listening test are shown in Fig. 5. The three different subplots show the results for the cafeteria noise (top), SSN (mid), and the car noise (bottom). Symbols indicate results for individual listeners, while lines show the average psychometric functions obtained by parametric averaging of the individual psychometric functions which were estimated according to Brand and Kollmeier (2002).

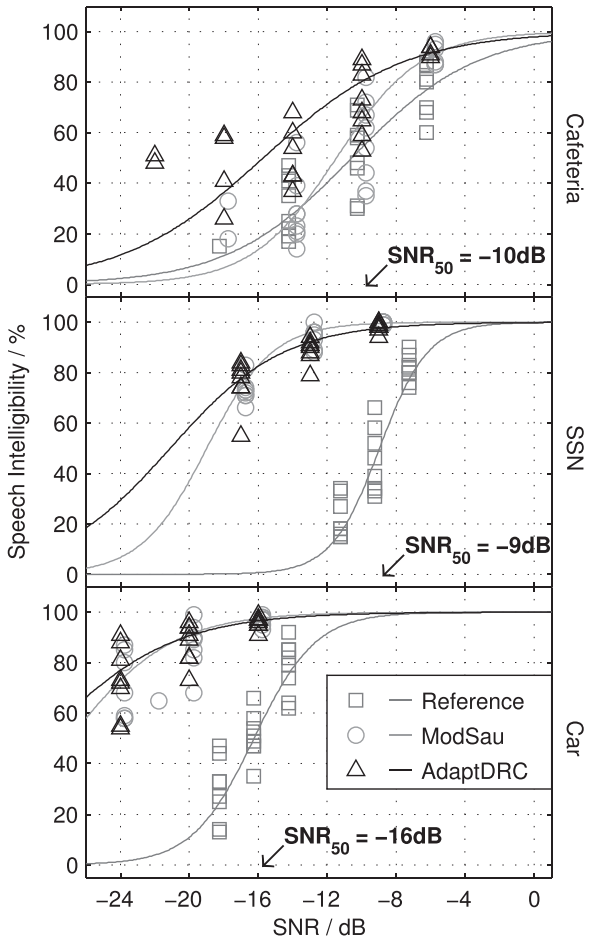

FIG. 5. Results of the formal listening test for the cafeteria noise (top panel), the SSN (mid panel), and the car noise (bottom panel) as a function of the SNR for the *Reference*, and both algorithms (*ModSau* and the proposed *AdaptDRC*). Individual points show results measured with subjects using the semi-adaptive procedure while straight lines indicate average psychometric function obtained by parametric averaging of individual psychometric functions. For each noise the corresponding SNR_{50} is indicated. Note that a slight offset is added to the individual data to increase visibility.

First, it could be observed that for all three noises in the unprocessed *Reference* condition an intelligibility of about 50% was measured on average at the SNR_{50} determined in the preliminary study, although interindividual variability was present. Furthermore, for the SSN and the car noise, both the *AdaptDRC* algorithm and the *ModSau* algorithm increased speech intelligibility at the SNR_{50} reaching almost perfect intelligibility. For the cafeteria noise at the SNR_{50}, the *AdaptDRC* algorithm showed the largest improvement, while the *ModSau* algorithm only showed a slight increase. To test for statistical significance of these findings, a two-factor ANOVA was performed after an ART of the data at the SNR_{50} with the factors of noise and algorithm. The statistical analysis showed a significant influence of both main factors as well as their interaction [noise: $F(2;14) = 26.82$, $p < 0.05$; algorithm: $F(2;14) = 284.44$, $p < 0.05$; noise × algorithm: $F(2;14) = 27.98$, $p < 0.05$]. Paired comparisons were performed on a Bonferroni-corrected significance level for the factor of algorithm, indicating that, at the SNR_{50}, the *AdaptDRC* algorithm improved speech intelligibility significantly over the *Reference* ($p < 0.0167$) and over the *ModSau* algorithm ($p < 0.0167$), and the *ModSau* algorithm significantly improved speech intelligibility over the *Reference* ($p < 0.0167$).

Considering each noise independently, it could be observed that for the car noise both the *AdaptDRC* algorithm and the *ModSau* algorithm showed a large improvement over the *Reference*. Furthermore, it was observed that due to the semi-adaptive procedure speech intelligibility was measured at different SNRs for different subjects. In order to allow for a statistical analysis, those SNRs that were measured for most subjects were chosen and data were interpolated using individual psychometric function when these SNRs were not explicitly measured. For the car noise SNRs of −16, −20, and −24 dB were chosen. A two-factor ANOVA after an ART of the data showed significant influence of both main factors SNR and algorithm as well as their interaction [SNR: $F(2;14) = 186.25$, $p < 0.05$; algorithm: $F(2;14) = 87.92$, $p < 0.05$; SNR × algorithm: $F(4;28) = 23.36$, $p < 0.05$]. Paired comparisons at the Bonferroni-corrected significance level for the factor algorithm showed a significant improvement in speech intelligibility for both algorithms over the *Reference* (*AdaptDRC*: $p < 0.0167$; *ModSau*: $p < 0.0167$) and no significant difference between both algorithms ($p = 0.072$).

For the SSN similar observations were made, i.e., both algorithms showed a large improvement over the *Reference*. Furthermore, for the SNRs under investigation no major differences could be observed between both algorithms. For the statistical analysis SNRs of −9, −13, and −17 dB were chosen and missing data points were interpolated using individual psychometric functions. A two-factor ANOVA after an ART confirmed the observations, revealing a significant influence of both main factors SNR and algorithm and their interaction [SNR: $F(2;14) = 221.86$, $p < 0.05$; algorithm: $F(2;14) = 65.21$, $p < 0.05$; SNR × algorithm: $F(4;28) = 44.70$, $p < 0.05$]. Paired comparisons at the Bonferroni-corrected significance level for the factor algorithm showed a significant increase in speech intelligibility for both algorithms over the *Reference* (*AdaptDRC*: $p < 0.0167$; *ModSau*: $p < 0.0167$) and no significant differences between the algorithms ($p = 0.65$).

For the cafeteria noise both algorithms showed improvements in speech intelligibility over the *Reference* for higher SNRs, while for lower SNRs the results indicated that only the *AdaptDRC* algorithm increased speech intelligibility over the *Reference*. An improvement of as large as 25% for an SNR of −14 dB could be observed. For the statistical analysis SNRs of −6, −10, and −14 dB were chosen and missing data points were interpolated using individual psychometric functions. A two-factor ANOVA after an ART showed a significant influence of both main factors and their interaction [SNR: $F(2;14) = 108.64$, $p < 0.05$; algorithm: $F(2;14) = 65.08$, $p < 0.05$; SNR × algorithm: $F(2;14) = 6.75$, $p < 0.05$]. Paired comparison at the Bonferroni-corrected significance level for the factor algorithm showed a significant improvement in speech intelligibility for the proposed *AdaptDRC* algorithm over both the *Reference*

($p < 0.0167$) and the *ModSau* algorithm ($p < 0.0167$), and no significant improvement for the *ModSau* algorithm over the *Reference* ($p = 0.066$).

3. Discussion

A semi-adaptive procedure was used to measure speech intelligibility in three different noises for the *AdaptDRC* algorithm, the *ModSau* algorithm, and the unprocessed *Reference*. The SNRs corresponding to an (average) speech intelligibility of 20%, 50%, and 80% as determined by a preliminary study yielded approximately the desired (average) speech intelligibility in this study for the unprocessed *Reference*.

To the best of our knowledge, a similar procedure has not been applied before to evaluate the performance of different speech pre-processing algorithms. By applying this procedure the subjects were first presented SNRs expected to results in 50% speech intelligibility for a specific combination of noise and algorithm. In the subsequently presented SNRs the SNR was typically decreased for the stationary noises and thus might have influenced the overall results in that subjects adapted to these severe conditions. Note that this may also have been the case if the SNRs had been typically increased. Since the order of the processing conditions was randomized, it is assumed that this procedure influenced all combination similarly and no combinations was particularly biased.

For the stationary SSN and car noise both the *AdaptDRC* algorithm and the *ModSau* algorithm were capable of increasing speech intelligibility by approximately the same amount. For the SSN at lower SNRs estimations based on parametrically averaged psychometric functions indicated that a larger speech intelligibility improvement may be achieved for the proposed *AdaptDRC* algorithm than for the *ModSau* algorithm. However, this is likely to be an artifact of the fitted psychometric function due to the limited spread of speech intelligibility data for this noise. This has to be verified using additional measurements in future studies.

For the non-stationary cafeteria noise only the proposed *AdaptDRC* algorithm significantly increased speech intelligibility compared to the *ModSau* algorithm and the unprocessed *Reference*. Furthermore, it could be observed that the interindividual variability for the processed combinations was increased compared to both stationary noises. Hence, the relative increase in speech intelligibility was larger for some subjects than for others, i.e., some subjects were able to benefit more from processing with the *AdaptDRC* algorithm than others for the non-stationary cafeteria noise.

Comparing the cafeteria noise and the SSN results may also provide some insight into the influence of the different stages of the proposed *AdaptDRC* algorithm. Both noises can be considered to be comparable with respect to their long-term average spectrum, but are different in their temporal structure. While for the SSN the *ModSau* algorithm and the *AdaptDRC* algorithm yielded approximately the same speech intelligibility, for the cafeteria noise only the *AdaptDRC* algorithm achieved significant improvements. On the one hand spectral changes introduced by both algorithms apparently did not lead to a different performance for the stationary SSN, while, on the other hand, for the non-stationary cafeteria noise the performance did differ. Hence, one may conclude that one factor causing this performance difference is the DRC stage of the *AdaptDRC* algorithm. To further investigate this, Fig. 6 depicts exemplary spectrograms for the *Reference*, the *ModSau* algorithm, the *AdaptDRC* algorithm, and the cafeteria noise. The used SNR was −18 dB. As can be seen, both algorithms increase the speech power in the frequency range from about 2 to about 10 kHz compared to the unprocessed *Reference*. Comparing the *AdaptDRC* algorithm and the *ModSau* algorithm, the *AdaptDRC* algorithm leads to a larger increase in speech power in the high frequencies as expected by the different amplification functions in Eqs. (13) and (16). Additionally, transients, e.g., at approximately 0.6 and 1.5 s, are amplified much stronger by the *AdaptDRC* algorithm, which may be attributed to the DRC stage.

Similar observations were made by Cooke *et al.* (2013b), where several algorithms were compared in the framework of the 2013 Hurricane Challenge. It was observed that two out of three algorithms that led to a significant improvement in a fluctuating speech masker contained a DRC stage, including the proposed *AdaptDRC* algorithm.

V. GENERAL DISCUSSION

In Sec. IV, the proposed *AdaptDRC* algorithm and the *ModSau* algorithm were compared by means of objective measures and a subjective listening test. Results from the listening test showed improvements of up to 70% in speech intelligibility as measured by correctly understood words. These improvements were considerably larger than

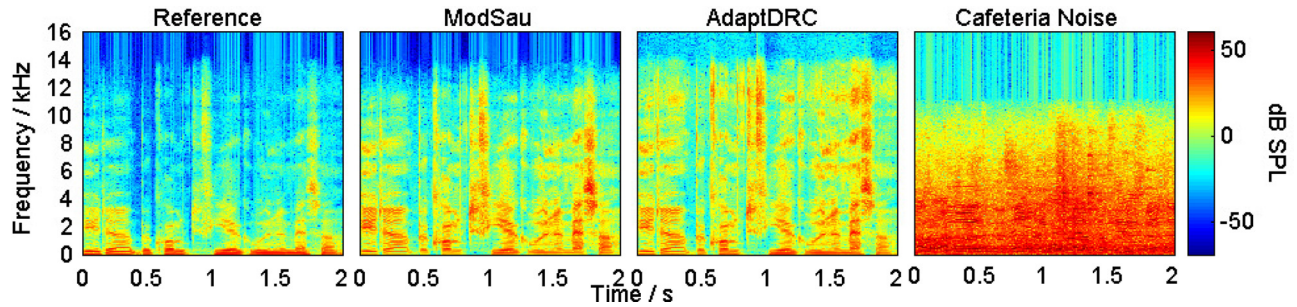

FIG. 6. (Color online) Exemplary spectrograms of a speech signal for the unprocessed *Reference* (left panel), the *ModSau* algorithm (mid left panel), and the *AdaptDRC* algorithm (mid right panel) for the cafeteria noise (right panel) mixed at an SNR of −18 dB.

improvements for similar noises (e.g., SSN) reported previously (e.g., Cooke *et al.*, 2013a; Cooke *et al.*, 2013b). One factor explaining this difference may be found in the employed sentence material. Cooke *et al.* (2013a) and Cooke *et al.* (2013b) used everyday sentences of the Harvard corpus (Rothauser *et al.*, 1969) which are unpredictable in their semantics and in their syntactical structure. In contrast, the sentences of the OLSA speech material have a clear syntactical structure but are semantically unpredictable. The average psychometric function for these two types of speech material can be considered different. The OLSA speech material was optimized in order to obtain a steep psychometric function, i.e., around 50% speech intelligibility small changes in SNR lead to a large change in speech intelligibility. For everyday sentences such as those of the Harvard corpus, the psychometric function is, in general, less steep, such that the same change in SNR leads to a smaller change in speech intelligibility for the HST speech material than for the OLSA speech material. The *AdaptDRC* algorithm was also evaluated as part of the study presented by Cooke *et al.* (2013b) and led to significant improvements, which supports the conclusion that differences in absolute speech intelligibility improvements may be mainly due to the different speech material.

When evaluating the performance using objective measures (see Sec. IV A), a large improvement in speech intelligibility was predicted for both algorithms for all noises while their overall performance was predicted to be rather similar. Only for medium to high SNRs the ESII predicted an improvement of the proposed *AdaptDRC* algorithm compared to the *ModSau* algorithm. However, the results from the formal listening test (see Sec. IV B) showed a large improvement for both algorithms in stationary noise, while only the *AdaptDRC* algorithm led to significant improvements for the non-stationary cafeteria noise. This raises the question of how to reliably interpret the results from the objective measures, which are commonly employed during the development stage of algorithms. Hence, in the following the predictive ability of both objective measures will be investigated using correlation analysis techniques between the objective and the subjective data.

The data are analyzed in terms of the rank correlation and linear correlation as well as the prediction bias. To calculate the linear correlation and the prediction bias values the results from the objective measures were transformed to account for a possible non-linear relationship between objective measures and subjective intelligibility scores. Based on the results for the unprocessed *Reference* a logistic function was fitted for each noise (Beutelmann and Brand, 2006), i.e.,

$$P(o) = \frac{m}{a + e^{-b \cdot o}} + c, \tag{29}$$

where o indicates the results obtained from either objective measure. The parameters m and b where fitted from the data and the parameter a and c were calculated from the boundary conditions of $P(0) = 0\%$ and $P(1) = 100\%$. Note that this procedure does not change the rank correlation since Eq. (29) is a monotonous function.

Figure 7 depicts an exemplary scatter plot for the measured speech intelligibility and the predicted speech intelligibility by the ESII after transformation using Eq. (29) for the cafeteria noise. Different symbols indicate different SNRs, while different gray scales indicate different algorithms, i.e., *Reference* (gray), *AdaptDRC* (black), and *ModSau* (light gray). Open symbols indicate individual results for each subject and filled symbols indicate average data across subjects (small symbols) or average data across subjects and processing condition (large symbols). Note that due to the semi-adaptive procedure for each subject the speech intelligibility was generally measured at different SNRs and, hence, average data for each SNR were not computed on the same number of individual data points. A correlation analysis was carried out for individual data (i.e., open symbols), as well as averaged data across subjects (i.e., small filled symbols) and averaged data across subjects and algorithm (i.e., large filled symbols).

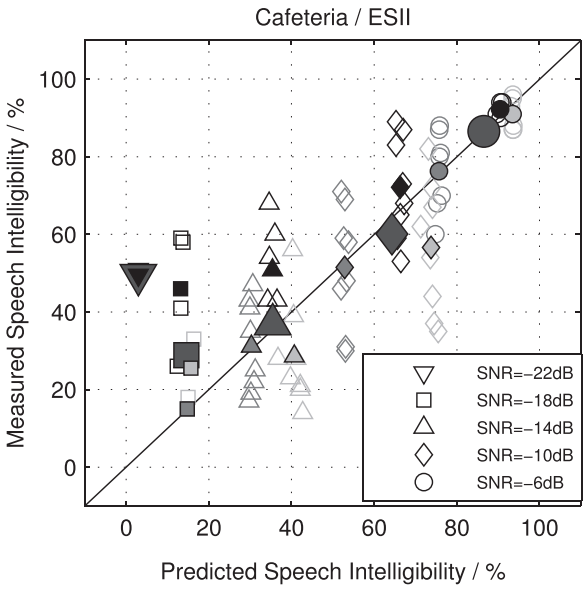

FIG. 7. Exemplary scatter plot of the predicted speech intelligibility by the ESII after transformation using Eq. (29) and the speech intelligibility measured in the formal listening test for the cafeteria noise at different SNRs. Gray symbols indicate data for the *Reference* condition, black symbols indicate data for the proposed *AdaptDRC* algorithm, and light grey symbols indicate data for the *ModSau* algorithm. Open symbols indicate individual data, filled symbols indicate average data across subjects, and large filled symbols indicate average data across subjects and algorithms (including the *Reference* condition).

Table IV provides an overview of the rank and linear correlations and the prediction bias between measured speech intelligibility and the predicted speech intelligibility by the ESII and STOI. Footnotes indicate statistically significant correlations. Values provided in parentheses are based on averaged data across subjects. In general, rank correlations for both objective measures were very similar, with values larger than 0.8 based on the averaged data. This indicates that both objective measures had a similar performance in terms of predicting the ranking of the subjective data, where in most conditions STOI achieved a slightly better predictive performance than the ESII. The best performance to predict the ranking of all algorithms was achieved for the SSN. Similar results were obtained for the linear correlation. For most

TABLE IV. Overview of rank correlation, linear correlation and the prediction bias for the comparison of measured speech intelligibility and predicted speech intelligibility by ESII and STOI.

		Rank correlation		Linear correlation		Prediction bias	
Noise	Algorithm	ESII	STOI	ESII	STOI	ESII	STOI
Cafeteria	Reference	0.79[a] (1.00[a])	0.80[a] (1.00[a])	0.85[a] (1.00)	0.84[a] (1.00)	0%	0%
	AdaptDRC	0.78[a] (0.90[a])	0.82[a] (0.90[a])	0.80[a] (0.95)	0.84[a] (0.99)	15%	23%
	ModSau	0.76[a] (1.00)	0.84[a] (1.00[a])	0.85[a] (0.94)	0.89[a] (1.00)	−10%	10%
	All	0.73[a] (0.82[a])	0.75[a] (0.86[a])	0.73[a] (0.83[a])	0.79[a] (0.89[a])	2%	11%
SSN	Reference	0.94[a] (1.00)	0.90[a] (1.00)	0.94[a] (0.99)	0.95[a] (0.99)	0%	0%
	AdaptDRC	0.88[a] (1.00)	0.89[a] (1.00)	0.85[a] (1.00)	0.83[a] (0.99)	24%	19%
	ModSau	0.88[a] (1.00)	0.90[a] (1.00[a])	0.96[a] (1.00)	0.97[a] (1.00)	12%	12%
	All	0.84[a] (0.87[a])	0.88[a] (0.92[a])	0.75[a] (0.77[a])	0.82[a] (0.84[a])	12%	10%
Car	Reference	0.86[a] (1.00[a])	0.82[a] (1.00[a])	0.89[a] (1.00)	0.88[a] (1.00)	0%	0%
	AdaptDRC	0.69[a] (0.90)	0.73[a] (1.00)	0.72[a] (0.99)	0.69[a] (0.99)	26%	26%
	ModSau	0.81[a] (1.00)	0.83[a] (0.80)	0.78[a] (0.87)	0.81[a] (0.90)	14%	20%
	All	0.74[a] (0.81[a])	0.69[a] (0.75[a])	0.65[a] (0.71[a])	0.63[a] (0.70[a])	13%	15%
All	Reference	0.88[a] (1.00[a])	0.86[a] (1.00[a])	0.90[a] (1.00[a])	0.90[a] (1.00[a])	0%	0%
	AdaptDRC	0.77[a] (0.88[a])	0.85[a] (0.95[a])	0.71[a] (0.85[a])	0.78[a] (0.91[a])	22%	23%
	ModSau	0.82[a] (0.85[a])	0.91[a] (0.93[a])	0.71[a] (0.82[a])	0.84[a] (0.91[a])	5%	14%
	All	0.76[a] (0.84[a])	0.80[a] (0.89[a])	0.71[a] (0.79[a])	0.77[a] (0.84[a])	9%	12%

[a]Significant correlations.

conditions STOI achieved a higher linear correlation than the ESII, especially when considering all noises together.

The prediction bias represents a measure for the offset of the predicted speech intelligibility relative to the measured speech intelligibility. A positive value indicates that the subjective data are underestimated, while a negative value indicates that the subjective data are overestimated. The results for the prediction bias showed that both objective measures tended to underestimate speech intelligibility. Note that the prediction bias for the *Reference* condition was 0% since the parameters of the transformation in Eq. (29) were fitted based on these data. Both the ESII and STOI measures tended to underestimate the measured speech intelligibility by up to 15%, which has to be considered when interpreting the data.

Tang and Cooke (2011) also reported correlation values between objective measures and measured speech intelligibility for several pre-processing algorithms. Similar to the present study they observed a large bias for the STOI measure, although it has to be mentioned that they did not consider a non-linear transformation as in Eq. (29). Furthermore, they only calculated one correlation value that included their complete set of conditions, i.e., all noises, SNRs and processing algorithms.

In this study an ideal scenario was considered where both the influence of reverberation and the influence of noise estimation errors were neglected. However, as argued in Sec. II, this provides information about the optimal performance and the general applicability of such algorithms. Nevertheless, reverberation and estimation errors are inevitable in real world scenarios, such that these influences have to be addressed in future studies.

VI. CONCLUSION

In this paper we have presented a novel speech pre-processing algorithm aiming to enhance speech intelligibility in noisy scenarios. The proposed *AdaptDRC* algorithm combines a time- and frequency-dependent amplification and dynamic range compression stage, where both stages depend on the SNR mapping function used in the SII. The main novelty is the fact that the input-output-characteristic of the DRC stage depends on the short-term characteristic of the speech and the disturbing noise. A comparison using objective measures showed only minor differences in performance compared to a state-of-the-art algorithm (Sauert and Vary, 2012) that only employs time- and frequency-dependent amplification. The results of a formal listening test, however, showed that the proposed *AdaptDRC* algorithm is capable of significantly increasing speech intelligibility and outperforms the state-of-art algorithm in non-stationary environments.

ACKNOWLEDGMENTS

The authors would like to thank Moritz Wächtler for his help in conducting the formal listening test and the anonymous reviewers for their helpful comments. This work was supported in part by the Research Unit FOR 1732 "Individualized Hearing Acoustics" and the Cluster of Excellence 1077 "Hearing4All," funded by the German Research Foundation (DFG).

ANSI (**1997**). S3.5, *Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

Arai, T., Hodoshima, N., and Yasu, K. (**2010**). "Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners," IEEE Trans. Audio Speech Lang. Process. **18**(7), 1775–1780.

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**(1), 331–342.

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimate for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**(6), 2801–2810.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multi-talker conditions," Acta Acust. **86**, 117–128.
Brouckxon, H., Verhelst, W., and Schuymer, B. (**2008**). "Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments," in *Proceedings of Interspeech*, Brisbane, Australia (September 2008), pp. 557–560.
Cooke, M., Mayo, C., and Valentini-Botinhao, C. (**2013b**). "Intelligibility-enhancing speech modifications: The Hurricane challenge," in *Proceedings of Interspeech*, Lyon, France, August 2013, 3552–3556.
Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (**2013a**). "Evaluating the intelligibility benefit of speech modifications in known noise conditions," Speech Commun. **55**(4), 572–585.
Crespo, J. B., and Hendriks, R. C. (**2013**). "Multizone near-end speech enhancement under optimal second-order magnitude distortion," in *Proceedings IEEE Workshop Applied Signal Processes Audio Acoustics (WASPAA)*, New Paltz, NY (October 2013).
George, E. L. J., Goverts, S. T., Festen, J. M., and Houtgast, T. (**2010**). "Measuring the effects of reverberation and noise on sentence intelligibility in hearing-impaired listeners," J. Speech Lang. Hear. Res. **53**, 1429–1439.
Haensler, E., and Schmidt, G. (**2008**). *Speech and Audio Processing in Adverse Environments* (Springer-Verlag, Berlin, Germany).
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.
Kusumoto, A., Arai, T., Kinoshita, K., Hodoshima, N., and Vaughan, N. (**2005**). "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," Speech Commun. **45**(2), 101–113.
Licklider, J. C. R., and Pollack, I. (**1948**). "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech," J. Acoust. Soc. Am. **20**(1), 42–51.
Lu, Y., and Cooke, M. (**2008**). "Speech production modifications produced by competing talkers, babble, and stationary noise," J. Acoust. Soc. Am. **124**(5), 3261–3275.
Morimoto, M., Sato, H., and Kobayashi, M. (**2004**). "Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces," J. Acoust. Soc. Am. **116**(3), 1607–1613.
Niederjohn, R., and Grotelueschen, J. (**1976**). "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," IEEE Trans. Acoust. Speech **24**(4), 277–282.
Ortega, V. H. M., and Huckvale, M. (**2000**). "Automatic cue-enhancement of natural speech for improved intelligibility," Speech Hear. Lang.: Work Prog. **12**, 42–56.
Regalia, P., Mitra, S., Vaidyanathan, P., Renfors, M., and Neuvo, Y. (**1987**). "Tree-structured complementary filter banks using all-pass sections," IEEE Trans. Circuits Syst. **34**(12), 1470–1484.
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**(5), 2642–2653.
Rhebergen, K. S., and Versfeld, N. J. (**2005**). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**(4), 2181–2192.
Rhebergen, K. S., 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.
Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., Nordby, K. S., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**(3), 225–246.
Sauert, B., and Vary, P. (**2010a**). "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proceedings of the European Signal Processing Conference*, Aalborg, Denmark (August 2010), pp. 1919–1923.
Sauert, B., and Vary, P. (**2010b**). "Recursive close-form optimization of spectral audio power allocation for near end listening enhancement," in *Proceedings of the ITG Conference on Speech Communication*, Bochum, Germany (October 2010).
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 (September 2012), pp. 195–198.
Schepker, H., Rennies, J., and Doclo, S. (**2013**). "Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression," in *Proceedings of Interspeech*, Lyon, France (August 2013), pp. 3577–3581.
Shin, J. W., and Kim, N. S. (**2007**). "Perceptual reinforcement of speech signal based on partial specific loudness," IEEE Signal Process. Lett. **14**(11), 887–890.
Skowronski, M. D., and Harris, J. G. (**2006**). "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," Speech Commun. **48**(5), 549–558.
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**(4), 858–872.
Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136.
Taal, C. H., and Jensen, J. (**2013**). "SII-based speech preprocessing for intelligibility improvement in noise," in *Proceedings Interspeech*, Lyon, France (August 2013), 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 (August 2011), pp. 345–348.
Tang, Y., and Cooke, M. (**2012**). "Optimising spectral weightings for noise-dependent speech intelligibility enhancement," in *Proceedings of Interspeech*, Portland, OR (September 2012).
Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**(3), 917–928.
Verhelst, W. (**2000**). "Overlap-add methods for time-scaling of speech," Speech Commun. **30**, 207–221.
Wagener, K., Brand, T., and Kollmeier, B. (**1999a**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests" ("Development and evaluation of a German sentence test II: Optimization of the Oldenburg sentence test"), Z. Audiol. **38**, 44–56.
Wagener, K., Brand, T., and Kollmeier, B. (**1999b**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests" ("Development and evaluation of a German sentence test III: Evaluation of the Oldenburg sentence test"), Z. Audiol. **38**, 86–95.
Wagener, K., Kühnel, V., and Kollmeier, B. (**1999c**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests" ("Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test"), Z. Audiol. **38**, 4–15.
Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J. (**2011**). "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the ACM Conference Human Factors in Computing Systems*, Vancouver, Canada (May 2011), pp. 143–146.
Zorila, T.-C., Kandia, V., and Stylianou, Y. (**2012**). "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proceedings of Interspeech*, Portland, OR (September 2012), pp. 635–638.
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 (September 2014), pp. 2050–2054.