Optimal Near-End Speech Intelligibility Improvement Incorporating Additive Noise and Late Reverberation Under an Approximation of the Short-Time SII

Richard C. Hendriks, João B. Crespo, Jesper Jensen, and Cees H. Taal

***Abstract*—The presence of environmental additive noise in the vicinity of the user typically degrades the speech intelligibility of speech processing applications. This intelligibility loss can be compensated by properly preprocessing the speech signal prior to playout, often referred to as near-end speech enhancement. Although the majority of such algorithms focus primarily on the presence of additive noise, reverberation can also severely degrade intelligibility. In this paper we investigate how late reverberation and additive noise can be jointly taken into account in the near-end speech enhancement process. For this effort we use a recently presented approximation of the speech intelligibility index under a power constraint, which we optimize for speech degraded by both additive noise and late reverberation. The algorithm results in time–frequency dependent amplification factors that depend on both the additive noise power spectral density as well as the late reverberation energy. These amplification factors redistribute speech energy across frequency and perform a dynamic range compression. Experimental results using both instrumental intelligibility measures as well as intelligibility listening tests show that the proposed approach improves speech intelligibility over state-of-the-art reference methods when speech signals are degraded simultaneously by additive noise and reverberation. Speech intelligibility improvements in the order of 20% are observed.**

***Index Terms*—Additive noise, approximated speech intelligibility index (SII), late reverberation, speech intelligibility.**

I. Introduction

SPEECH communication systems are ubiquitous these days. Consider for example applications like mobile telephony, hearing aids and public address systems. These applications require that speech is well understood. However, due to presence of noise and reverberation, speech intelligibility can get degraded.

Manuscript received September 19, 2014; revised January 06, 2015; accepted February 17, 2015. Date of publication March 06, 2015; date of current version March 27, 2015. This work was supported in part by the Dutch Technology Foundation STW and Bosch Security Systems B.V., The Netherlands. Parts of this work were presented at ICASSP 2015 [47]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mads Christensen.

R. C. Hendriks and J. B. Crespo are with the Signal and Information Processing Lab, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: r.c.hendriks@tudelft.nl).

J. Jensen is with Oticon A/S, 2765 Smørum, Denmark, and also with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: jsj@oticon.dk; jje@es.aau.dk).

C. H. Taal is with Applied Sensor Technologies, Philips Research, 5656 AE Eindhoven, The Netherlands (e-mail: cees.taal@philips.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2409780

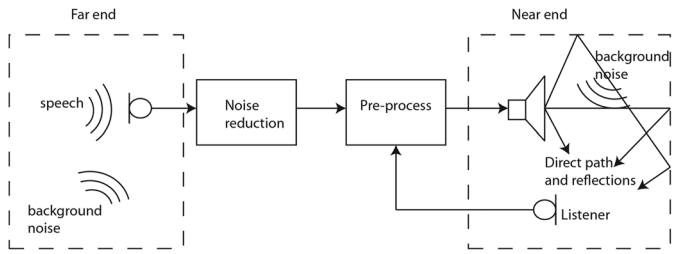

Fig. 1. Visualization of the application scenario.

In speech communication systems two different environments can be defined, often referred to as the near-end environment (the environment of the receiver) and the far-end environment (the environment of the sender). See Fig. 1 for a visualization. These two different environments give rise to two different reasons for intelligibility reduction for the near-end listener. At first, there is degradation due to noise sources that are present in the environment of the far-end talker. These sources get mixed with the target speech, and, when transmitted, lead to intelligibility reduction for the listener at the near-end. By applying single-microphone noise reduction (*e.g.*, [1], [2]) or multi-microphone noise reduction (*e.g.*, [3][4, Chs.43 - 53]) to the noisy signal recorded at the far-end, the effects of acoustical noise sources in this scenario can be reduced. Although single-microphone algorithms improve quality [5], they rarely lead to an intelligibility increase [6], [7], [8] (see [6], [9] for a few modest exceptions). Multi-microphone algorithms on the other hand are an effective means to improve both the quality as well as the intelligibility in this scenario [7].

Secondly, intelligibility for the near-end listener can also degrade when noise sources are present in the environment, while listening to the far-end talker. For example, when a train passes by at a station while announcements are made via a public address system. These noise sources in the environment of the near-end listener can not be reduced using a post-processor as usually done in the former scenario. A solution to this problem is to process the speech that is transmitted from the far-end to the near-end, such that speech intelligibility is maintained when the signal is exposed to environmental noise before reaching the observer at the near-end. We will refer to this scenario as near-end listening intelligibility improvement.

After some early initiatives on this problem in the previous century, (for example [10], [11]) it is only recently that this problem experienced a revival, see *e.g.* [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], and the numerous contributions

to the Hurricane challenge [22] for more examples. Many contributions are based on the empirical considerations that high frequencies and consonants are important for intelligibility, *e.g.*, [10], [11], [23], [12]. Although the empirical based approaches clearly improve the speech intelligibility, they often lack the use of a measure for speech intelligibility and cannot claim any optimality.

More recently, interest increased to improve the near-end speech intelligibility by optimizing more formal models of speech intelligibility. The methods in [18], [20], [21] optimize the glimpse proportion metric [24], which measures the proportion of spectro-temporal regions whose local signal-to-noise ratio (SNR) exceeds a pre-determined threshold. The approaches in [14], [15], [19] try to optimize the speech intelligibility index (SII) [25], and in [17], [48] it was proposed to optimally redistribute speech energy over frequency and time using a perceptual distortion measure based on a spectro-temporal auditory model presented in [26].

The aforementioned contributions all consider the situation where speech only gets degraded by additive noise. However, in many application scenarios, in particular the ones where speech is presented via loudspeakers, speech typically also gets degraded due to reverberation. It is important to emphasize here that despite the fact that reverberation can degrade the speech intelligibility, certain aspects of reverberation can also improve the speech intelligibility [27]. Typically, a distinction is made into the early reflections (up to say, 50 ms) and the late reflections. The presence of the early reflections can increase the effective SNR as observed by the listener and thus, can increase the intelligibility. This is in particular the case in conditions where the power of the direct sound is reduced. In the current paper, we focus on the presence of late reverberation, as, opposed to early reflections, this degrades the speech intelligibility.

Although there are some contributions where late reverberation is taken into account (*e.g.*, [28], [29]), the case with both additive noise and late reverberation has only rarely been treated in the context of near-end speech enhancement. In [30] an algorithm for speech reinforcement under noisy and reverberant conditions was presented by optimizing a slightly modified version of the perceptual distortion measure presented in [26]. The algorithm in [30] was derived under an energy constraint per frequency band across a segment of time-frames. The instrumental measure that was optimized in [30] minimizes the detectability of noise and late reverberation under early speech and led to a local optimal solution. In [31], the near-end intelligibility problem was considered in a multi-zone scenario, where a general optimization framework for intelligibility improvement in multiple zones was proposed. The signal model used in [31] allows to include effects of noise, reverberation as well as crosstalk between different zones.

In the current paper, we consider the single-zone scenario instead of the multi-zone scenario from [31] and optimize an instrumental measure predicting intelligibility instead of the perceptual distortion measure from [26] that was optimized in [30]. We investigate in this paper how effects of both reverberation and additive noise can jointly be taken into account. We make use of the approximation of the SII model presented in [19], which we refer to as the approximated SII (ASII). The SII model can very well predict the effects of additive stationary noise on intelligibility by comparing the average speech energy within one critical band with the average noise energy within the same critical band. The ASII [19] was presented to make constrained optimization of the SII model mathematical tractable, having been used in [19] to find the ASII optimal amplification per critical band under an energy constraint. The energy constraint can be used to satisfy loudspeaker power constraints or to overcome hearing discomfort due to loud sounds. The SII (and therefore also the ASII) was originally defined for stationary noise sources. As we also consider late reverberation, which is a non-stationary distortion by nature, we use the short-time variant of ASII, *i.e.*, $\mathrm{ASII}_{\mathrm{ST}}$. The $\mathrm{ASII}_{\mathrm{ST}}$ is thus an approximation of the original short-time variant of SII presented in [32], which is often referred to as extended SII (ESII). The ESII was proposed as an improvement on the SII, and can take the effect of non-stationary noise sources on intelligibility better into account. With this paper, we extend the results presented in [19] by taking also reverberation into account. As SII is not specifically developed for modeling the fine structure of reverberation, we only consider the average late reverberation energy, *i.e.*, the diffuse component of the reverberation in a critical band, which can then be modeled as an additional noise component.

II. Notation, Modeling Assumptions and Problem Formulation

Let $s(n)$ be a discrete-time speech signal with sampling index n. Let the processed version of $s(n)$ be denoted by $s_p(n)$, where the processing is intended to increase its intelligibility when played in a noisy and reverberant environment. Let h denote a time-invariant room impulse response. The reverberant speech is then given by $r(n) = (h{*}s_p)(n)$, where $*$ is the convolution operator. We assume that $r(n)$ can be split into a part containing the early reverberation and the late reverberation of the processed speech, that are, $e(n)$ and $z(n)$, respectively. Further, let $w(n)$ be additive noise, uncorrelated with $z(n)$ and $e(n)$. The observed noisy processed speech is then given by

$$x(n) = (h{*}s_p)(n) + w(n) = e(n) + z(n) + w(n). \qquad (1)$$

Processing $s(n)$ for intelligibility increase will be done per critical band and per time frame, similarly as in [19]. Let $g_i(n)$ denote the impulse response of the ith critical band filter, with $i \in \{1, \ldots, I\}$ and I the total number of critical band filters. The discrete Fourier transform (DFT) coefficient of $g_i(n)$ for frequency bin k is given by $G_i(k)$. Let $v(n)$ denote a window function with support N. The DFT coefficient of a windowed speech frame at the loudspeaker for frequency-bin k and a time frame starting at sampling-index m is in turn denoted by $S(m,k)$. The speech energy within one critical band i and time frame starting at sampling-index m is then defined as $\mathcal{S}^2(m,i) = \sum_k |S(m,k)|^2 |G_i(k)|^2$. Similarly we define the energy of the noise and late reverberation within one critical band as $\mathcal{W}^2(m,i) = \sum_k |W(m,k)|^2 |G_i(k)|^2$ and $\mathcal{Z}^2(m,i) = \sum_k |Z(m,k)|^2 |G_i(k)|^2$, respectively. The processed speech energy within one critical band is given by $\alpha^2(m,i)\mathcal{S}^2(m,i)$, with $\alpha(m,i)$ the (real and positive) amplification per time frame and critical band. Our interest is to find the amplification factors $\alpha(m,i)$ per time frame and critical band, which satisfy constrained optimality conditions with

respect to intelligibility, given in terms of the processed speech $\alpha^2(m,i)\mathcal{S}^2(m,i)$, as will be seen in Section IV.

The SII models the speech intelligibility as a function of the long-term SNR per critical band. This is defined in [25] as,

$$\xi_i = \frac{\sigma^2_{\mathcal{S}_i}}{\sigma^2_{\mathcal{W}_i}}, \tag{2}$$

where $\sigma^2_{\mathcal{S}_i}$ and $\sigma^2_{\mathcal{W}_i}$ were originally defined as long term averages of the speech and noise power spectrum, respectively. Assuming that speech is at comfortable level and neglecting masking effects, an approximate SII is calculated by transforming the SNRs ξ_i to the logarithmic domain, clipping them between -15 and $+15$ dB and normalizing them to ensure the outcome to be between zero and one, that is,

$$d(\xi_i) = \frac{\max(\min(10\log_{10}(\xi_i), 15), -15)}{30} + \frac{1}{2}. \tag{3}$$

The final SII score is then obtained by computing a weighted average over all critical bands as

$$SII = \sum_i \gamma_i d(\xi_i). \tag{4}$$

The weightings γ_i are known as the critical-band-importance functions and are given in [25].

The function $d(\xi_i)$ in (3) is not concave nor convex, which complicates constrained optimization. Therefore, it was suggested in [19] to approximate the SII by replacing $d(\xi_i)$ with the function

$$\tilde{d}(\xi_i) = \frac{\xi_i}{\xi_i + 1}. \tag{5}$$

The ASII is then given by

$$ASII = \sum_i \gamma_i \frac{\xi_i}{\xi_i + 1}, \tag{6}$$

which is concave as a function of ξ_i. We adopt the approximated SII in this paper and investigate how late-reverberation can be taken into account for speech reinforcement.

As reverberation is time-varying by nature, we do not consider the original ASII, but an approximation of the short-time variant of SII [32] also known as ESII. ESII is known to show higher correlation with intelligibility when non-stationary disturbances are present. Here, we assume speech and noise processes to be stationary and ergodic within one time-frequency unit, instead of across the entire signal. Let $E[\cdot]$ denote the statistical expectation operator and let us assume that speech and noise DFT coefficients are independent across frequency. The speech and noise variances per time-frequency unit are then given by $\sigma^2_{\mathcal{S}}(m,i) = E[\mathcal{S}^2(m,i)]$ and $\sigma^2_{\mathcal{W}}(m,i) = E[\mathcal{W}^2(m,i)]$, respectively. Similarly as for the speech and noise variances, we can define the variance of the late reverberation, that is, $\sigma^2_{\mathcal{Z}}(m,i) = E[\mathcal{Z}^2(m,i)]$. For completeness, we also define here the variances of the speech, noise and late reverberation in the DFT domain, that are, $\sigma^2_S(m,k) = E[S^2(m,k)]$, $\sigma^2_W(m,k) = E[W^2(m,k)]$ and $\sigma^2_Z(m,k) = E[Z^2(m,k)]$.

Using $\sigma^2_{\mathcal{S}}(m,i)$ and $\sigma^2_{\mathcal{W}}(m,i)$, the SNR per critical band and time frame is given by

$$\xi(m,i) = \frac{\sigma^2_{\mathcal{S}}(m,i)}{\sigma^2_{\mathcal{W}}(m,i)}. \tag{7}$$

The $\mathrm{ASII}_{\mathrm{ST}}$ is then given by

$$\mathrm{ASII}_{\mathrm{ST}} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\sum_i \gamma_i \frac{\xi(m,i)}{\xi(m,i)+1}, \tag{8}$$

where $|\mathcal{M}|$ indicates the cardinality of the set $\mathcal{M}$, which denotes the time-frame indices that are used to compute $\mathrm{ASII}_{\mathrm{ST}}$.

III. ASII based on Late Reverberation and Additive Noise

Since late reverberation and noise are known to decrease intelligibility (*e.g.,* [33]), we change the definition of SNR in (7) such that late reverberation generated by amplifying speech in a certain time-frame is also partly taken into account. We express $\sigma^2_{\mathcal{Z}}(m,i)$ in terms of the sum of the variances of $z(n)$ within a critical band in the DFT domain, that is,

$$\sigma^2_{\mathcal{Z}}(m,i) = \sum_k \sigma^2_Z(m,k)|G_i(k)|^2. \tag{9}$$

For simplicity, we neglect the early reflections in the impulse response, and set the processed signal that will be received via the direct path from loudspeaker to listener (corrected by the damping that it experiences) as the desired signal. Notice that this is a worst case scenario as early reflections can contribute to the intelligibility of speech (e.g., [27]). Let Δ model the attenuation of the direct sound. In the impulse response, Δ is given by the magnitude of the impulse in the response representing the direct path. We assume Δ is inversely proportional to the distance and define $n_\Delta = \Delta^{-1} f_s c^{-1}$ as the direct path delay from loudspeaker to listener location in samples with f_s the sampling frequency and $c = 343$ m/s the speed of sound. Further, let $n_0 = n_\Delta + \tau$ denote the sample index from which the late reflections of the impulse response start, with τ the pause between the direct path and the moment that the late reverberation starts [27]. The exact value of τ depends on the room acoustics, but a typical value is 50 ms [27]. From now on we will consider the late reverberation as noise and redefine the SNR such that late reverberation that is present n_0 samples after play out by the loudspeaker is also taken into account. We thus define SNR as the ratio between the speech variance per critical band and time frame after amplification by $\alpha(m,i)$ at the listener location (*i.e.*, after a delay of n_Δ samples and damping Δ), that is, $\alpha^2(m,i)\sigma^2_{\mathcal{S}}(m,i)\Delta^2$, and the sum of the noise variance and late reverberation present n_0 samples after play out by the loudspeaker. That is,

$$\begin{aligned}\xi(m+n_\Delta,i) &= \frac{\alpha^2(m,i)\sigma^2_{\mathcal{S}}(m,i)\Delta^2}{\sigma^2_{\mathcal{Z}}(m+n_0,i)+\sigma^2_{\mathcal{W}}(m+n_\Delta,i)} \\ &= \frac{\alpha^2(m,i)\sigma^2_{\mathcal{S}}(m,i)\Delta^2}{\sum_k \sigma^2_Z(m+n_0,k)|G_i(k)|^2 + \sum_k \sigma^2_W(m+n_\Delta,k)|G_i(k)|^2}.\end{aligned} \tag{10}$$

Eq. (10) shows that in order to take late reverberation into account, it is required to know both the power spectral densities (PSDs) $\sigma^2_Z(m,k)$ and $\sigma^2_W(m,k)$. Among other ways, one straightforward way to estimate the noise PSD $\sigma^2_W(m,k)$ is using a noise PSD estimation algorithm, *e.g.*, [34][35]. Alternatively, noise PSDs can be measured in noise-only regions when any reverberation due to speech activity has decayed (notice that this requires the noise source to be stationary). For the PSD of

the late reverberation, we derive in this section an expression under a stochastic model for the late reflections of the room impulse response.

Let $v(n)$ denote a window function with support N as defined in Section II. The late reverberation DFT coefficient $Z(m,k)$ is given by

$$Z(m,k) = \sum_{n=m}^{m+N-1} v(n-m) \sum_{l=n_0}^{+\infty} h(l) s_p(n-l) e^{-j2\pi k(n-m)/N}. \tag{11}$$

By switching the two summations we obtain

$$\begin{aligned} Z(m,k) &\approx \sum_{l=n_0}^{+\infty} h(l) \sum_{n=m}^{m+N-1} v(n-m) s_p(n-l) e^{-j2\pi k(n-m)/N} \\ &= \sum_{l=n_0}^{+\infty} h(l) S_p(m-l,k), \end{aligned} \tag{12}$$

where the approximation emphasizes that this is valid under the assumption that the impulse response is time-invariant during a time-frame [36], and where the subscript p in S_p indicates that this includes the amplification by $\alpha(m,i)$. To model the late reflections of the impulse response we use the Polack model [37], that is for $l \geq n_0$,

$$h(l) = a^{l-n_0} u(l-n_0), l \geq n_0, \tag{13}$$

with $u(l)$ an uncorrelated white stationary Gaussian noise process with variance σ_u^2 and a a damping factor. Let N denote the frame size and $R = N/2$ the frame shift. Assuming that u and S are two independent processes, we get for $\sigma_Z^2(m,k)$

$$\sigma_Z^2(m,k) = a^{-2n_0} \sigma_u^2 \sum_{l=n_0}^{+\infty} a^{2l} \sigma_{S_p}^2(m-l,k) \tag{14}$$

$$= a^{-2n_0} \sigma_u^2 \sum_{b=0}^{+\infty} \sigma_{S_p}^2(m-n_0-bR,k) \times \sum_{l=bR}^{bR+N-1} a^{2(l+n_0)}, \tag{15}$$

where we used the assumption that speech is stationary over a time frame, and where parameter b in Eq. (15) acts as a frame index. Using the geometric series this can be rewritten as

$$\sigma_Z^2(m,k) = \sigma_u^2 \sum_{b=0}^{+\infty} \sigma_{S_p}^2(m-n_0-bR,k) \frac{a^{2bR}(1-a^{2N})}{1-a^2}. \tag{16}$$

We can express σ_u^2 in terms of the diffuse room impulse response energy $\rho^2 = \sum_{l=n_0}^{+\infty} E[h^2(l)] \frac{1}{f_s}$ as $\sigma_u^2 = (1-a^2)\rho^2$ [31], finally leading to

$$\sigma_Z^2(m,k) = \rho^2 \sum_{b=0}^{+\infty} \sigma_{S_p}^2(m-n_0-bR,k) \left(a^{2bR}(1-a^{2N})\right). \tag{17}$$

Taking the amplification $\alpha^2(m,i)$ explicitly into account, we can write the late reverberation energy of the processed speech per critical band as

$$\begin{aligned} \sigma_{\mathcal{Z}}^2(m,i) &= \rho^2 \left(1-a^{2N}\right) \\ &\times \sum_{b=0}^{B-1} a^{2bR} \alpha^2(m-n_0-bR,i) \sigma_{\mathcal{S}}^2(m-n_0-bR,i), \end{aligned} \tag{18}$$

where, compared to (17) we have truncated the summation over b (reflecting the late reverberation) to B time frames. Substitution of $\sigma_{\mathcal{Z}}^2(m,i)$ from (18) into (10), we get the SNR per critical band given by

$$\xi(m+n_\Delta,i) = \frac{\alpha^2(m,i)\sigma_{\mathcal{S}}^2(m,i)\Delta^2}{\sigma_{\mathcal{W}}^2(m+n_\Delta,i) + \sigma_{\mathcal{Z}}^2(m+n_0,i)}, \tag{19}$$

where it is important to notice that $\sigma_{\mathcal{Z}}^2(m+n_0,i)$ depends on $\alpha^2(m-bR,i)$ with $b \in \{0,\ldots,B-1\}$ as given in (18).

IV. Energy Constrained Speech Intelligibility Maximization

In this section we derive optimal amplifications $\alpha^2(m,i)$ such that the $\mathrm{ASII}_{\mathrm{ST}}$ is maximized locally under an energy constraint r. The ultimate goal is to compute the $\alpha^2(m,i)$ such that they are globally optimal. Let $\mathcal{M}$ denote a set of time-frame indices m for which we would like to maximize the $\mathrm{ASII}_{\mathrm{ST}}$. Our objective function then follows by substitution of (19) into (8) as

$$\mathrm{ASII}_{\mathrm{ST}} = \frac{1}{|\mathcal{M}|} \times \sum_{m\in\mathcal{M}} \sum_i \times \frac{\gamma_i \alpha^2(m,i)\sigma_{\mathcal{S}}^2(m,i)\Delta^2}{\alpha^2(m,i)\sigma_{\mathcal{S}}^2(m,i)\Delta^2 + \sigma_{\mathcal{W}}^2(m+n_\Delta,i) + \sigma_{\mathcal{Z}}^2(m+n_0,i)}, \tag{20}$$

where we again emphasize that $\sigma_{\mathcal{Z}}^2(m+n_0,i)$ depends on $\alpha^2(m-bR,i)$ with $b \in \{0,\ldots,B-1\}$. The constrained optimization problem is then given by

$$\max_{\alpha^2(m,i)\forall m\in\mathcal{M},\forall i} \quad \mathrm{ASII}_{\mathrm{ST}} \tag{21}$$

$$s.t. \sum_{m\in\mathcal{M}} \sum_i \alpha^2(m,i)\sigma_{\mathcal{S}}^2(m,i) = r, \tag{22}$$

$$\alpha^2(m,i) \geq 0, \forall(m,i), \tag{23}$$

where the latter constraint is introduced to guarantee non-negative amplifications α^2.

However, the objective function $\mathrm{ASII}_{\mathrm{ST}}$ including the late reverberation as in (20) is not a concave function in $\alpha^2(m,i)$ (in fact, it is a sum of quasi-concave functions). Therefore, we consider a simplified problem where we compute only the locally optimal $\alpha^2(m,i)$. That is, we compute the amplifications $\alpha^2(m,i)$ for all frequency bands i and the time frames in a segment of B time-frames that maximize $\mathrm{ASII}_{\mathrm{ST}}$ locally for all frequency bands i in one time frame m. Let the starting samples of the B time frames in this segment be denoted by the set $\mathcal{L} = \{m-(B-1)R, m-(B-2)R, \ldots, m\}$. We then get the performance function

$$J_1 = \sum_i \frac{\gamma_i \alpha^2(m,i)\sigma_{\mathcal{S}}^2(m,i)\Delta^2}{\alpha^2(m,i)\sigma_{\mathcal{S}}^2(m,i)\Delta^2 + \sigma_{\mathcal{W}}^2(m+n_\Delta,i) + \sigma_{\mathcal{Z}}^2(m+n_0,i)}, \tag{24}$$

with $\sigma_{\mathcal{Z}}^2(m+n_0,i)$ defined as in (18), depending on $\alpha^2(\ell,i)\forall i, \ell \in \mathcal{L}$. Altogether, we then have the problem formulation

$$\max_{\alpha^2(\ell,i)\forall i,\ell\in\mathcal{L}} \quad J_1 \tag{25}$$

$$s.t. \sum_i \sum_{\forall \ell \in \mathcal{L}} \alpha^2(\ell,i)\sigma_{\mathcal{S}}^2(\ell,i) = r, \quad (26)$$

$$\alpha^2(m,i) \geq 0, \forall (m,i). \quad (27)$$

That is, we compute the $\alpha^2(\ell,i)$ for time frames $\ell \in \mathcal{L}$ and all frequency bands, such that the $\mathrm{ASII}_{\mathrm{ST}}$ in time-frame m is maximized under the given energy constraint. Notice that repeating this optimization for consecutive time-frames will each time lead to a set of gains, *i.e.*, $\alpha^2(m-(B-1)R,i), \alpha^2(m-(B-2)R,i), \ldots, \alpha^2(m,i)$ that partly overlap with the set of optimal gains that result from the optimization in the previous time-frame. Although there are several ways to combine these overlapping gains, we combine the multiple gains per TF-unit by taking their average.

Obviously, the goal is to maximize the performance function J_1 under the constraints in (26) and (27). J_1 is a function of the amplification factors $\alpha^2(\ell,i)\forall i, \ell \in \mathcal{L}$ via $\sigma_{\mathcal{Z}}^2(m+n_0,i)$. This problem can be simplified by realizing that setting $\alpha^2(\ell,i)\forall i, \ell \in \mathcal{L}\backslash l = m$ to zero will always increase J_1 as this reduces the reverberation generated by past time-frames. We can thus define another performance function where all $\alpha^2(\ell,i)\forall i, \ell \in \mathcal{L}\backslash l = m$ are set to zero. That is,

$$J_2 = \sum_i \frac{\gamma_i \alpha^2(m,i)\sigma_{\mathcal{S}}^2(m,i)\Delta^2}{\alpha^2(m,i)\sigma_{\mathcal{S}}^2(m,i)\left(\Delta^2+\rho^2\left(1-a^{2N}\right)\right)+\sigma_{\mathcal{W}}^2(m+n_\Delta,i)}, \quad (28)$$

for which it holds that

$$J_1 \leq J_2, \quad (29)$$

for all feasible $\alpha^2(m,i)$. More specifically, for each band i, all terms in both the numerator and denominator of the performance function J_1 defined in (24) are positive. Setting any of the $\alpha^2(m-bR,i)$ with $b \in \{1,\ldots,B-1\}$ to zero will always increase the value of the function in (29), reaching equality to J_2 in (28) if all $\alpha^2(m-bR,i)$ with $b \in \{1,\ldots,B-1\}$ are set to zero. This is due to the fact that the performance function only has a local view on the problem. Using the upper bound of J_1, given by J_2 in (28), the optimization problem in Eq. (25)–(27) can be stated to be equivalent to

$$\max_{\alpha^2(m,i)\forall i} J_2 \quad (30)$$

$$s.t. \sum_i \sum_{\forall \ell \in \mathcal{L}} \alpha^2(\ell,i)\sigma_{\mathcal{S}}^2(\ell,i) = r \quad (31)$$

$$\alpha^2(\ell,i) = 0, \text{ for } \ell \in \mathcal{L}\backslash \ell = m \text{ and } \forall i \quad (32)$$

$$\alpha^2(m,i) \geq 0, \forall i. \quad (33)$$

The optimization problem in Eq. (30)–(33) is a convex problem as the objective function is concave in $\alpha^2(m,i)$ and the constraints are all linear in $\alpha^2(\ell,i)$ with $\ell \in \mathcal{L}$. Consequently, its Karush-Kuhn-Tucker (KKT) conditions [38], [39]

$$-\frac{\gamma_i \sigma_{\mathcal{S}}^2(m,i)\Delta^2\sigma_{\mathcal{W}}^2(m+n_\Delta,i)}{\alpha^2(m,i)\sigma_{\mathcal{S}}^2(m,i)(\Delta^2+\rho^2\left(1-a^{2N}\right)+\sigma_{\mathcal{W}}^2(m+n_\Delta,i))^2} + \nu\sigma_{\mathcal{S}}^2(m,i) - \lambda(m,i) = 0, \forall i \quad (34)$$

$$\sum_i \sum_{\forall \ell \in \mathcal{L}} \alpha^2(\ell,i)\sigma_{\mathcal{S}}^2(\ell,i) = r \quad (35)$$

$$\alpha^2(\ell,i) = 0, \text{ for } \ell \in \mathcal{L}\backslash \ell = m \text{ and } \forall i \quad (36)$$

$$\alpha^2(m,i) \geq 0, \forall i \quad (37)$$

$$\lambda(m,i) \geq 0, \forall i \quad (38)$$

$$\lambda(m,i)\alpha^2(m,i) = 0, \forall i \quad (39)$$

form necessary and sufficient conditions to solve it.

Notice that for $\rho = 0$ and $B = 1$, problem (30)–(33) is a similar problem as the one posed in [19].

Completing the derivation of the algorithm for $\rho \neq 0$, the solution of the KKT conditions (34)–(39) is given by (see also [19])

$$\alpha^2(m,i) = \frac{\max\left(\frac{\sigma_{\mathcal{W}}(m+n_\Delta,i)\sqrt{\gamma_i}\Delta}{\sqrt{\nu}} - \sigma_{\mathcal{W}}^2(m+n_\Delta,i), 0\right)}{\sigma_{\mathcal{S}}^2(m,i)(\Delta^2+\rho^2\left(1-a^{2N}\right))}\forall i \quad (40)$$

$$\alpha^2(\ell,i) = 0, \text{ for } \ell \in \mathcal{L}\backslash \ell = m \text{ and } \forall i \quad (41)$$

$$\sum_i \max\left(\frac{\sigma_{\mathcal{W}}(m+n_\Delta,i)\sqrt{\gamma_i}\Delta}{\sqrt{\nu}} - \sigma_{\mathcal{W}}^2(m+n_\Delta,i), 0\right) = r(\Delta^2+\rho^2\left(1-a^{2N}\right)), \quad (42)$$

where the existence and uniqueness of $\nu > 0$ in (42) are guaranteed by the derivation in the appendix. The algorithm for finding the optimal KKT point then consists in finding the unique root $\nu > 0$ of (42) using a root finding algorithm (*e.g.*, bisection method) and subsequently substituting the root in the optimal gains of (40). Also, note that the presented solution is identical to the one of [19] for $\rho = 0$ and $B = 1$, *i.e.*, for the case that no reverberation is present. The general behavior of this algorithm is that only critical bands which are relevant for intelligibility are amplified. Bands where the SNR is too low and where amplification within the energy constraint will not help to increase the intelligibility, will automatically be clipped to zero in order to save this energy for bands that can still contribute to the intelligibility. If late reverberation is present, *i.e.*, when $\rho > 0$, the total gain in (40) in all bands is decreased by a factor $(\Delta^2+\rho^2(1-a^{2N}))$. The decrease in all bands by the factor $(\Delta^2+\rho^2(1-a^{2N}))$ can be explained by the fact that amplifying speech will automatically increase the distortion introduced by the late reverberation. Due to this overall decrease in energy, more energy is left according to the energy constraint. This energy is automatically used to amplify these frequency bands that otherwise would be clipped to zero due to the lack of enough energy. This happens as ν decreases according to (42) when $\rho > 0$. A smaller ν in (40) then automatically results in less bands being clipped to zero.

This behavior is visualized in Fig. 2, where we show spectrograms of the original clean speech (Fig. 2(a)), the speech signal processed by Taal [19] (Fig. 2(b)) (*i.e.*, without taking reverberation into account), the speech signal processed by the proposed approach with $B = 3$ (Fig. 2(c)), and the noise signal (Fig. 2(d)), which consists of speech shaped noise. The SNR is set to 2 dB and the T60 of the room impulse response is set to 1 second.

In Fig. 2(b) we see that if reverberation is not taken into account, generally most energy is put at higher frequencies, while lower frequencies are discarded. This is due to the fact that the

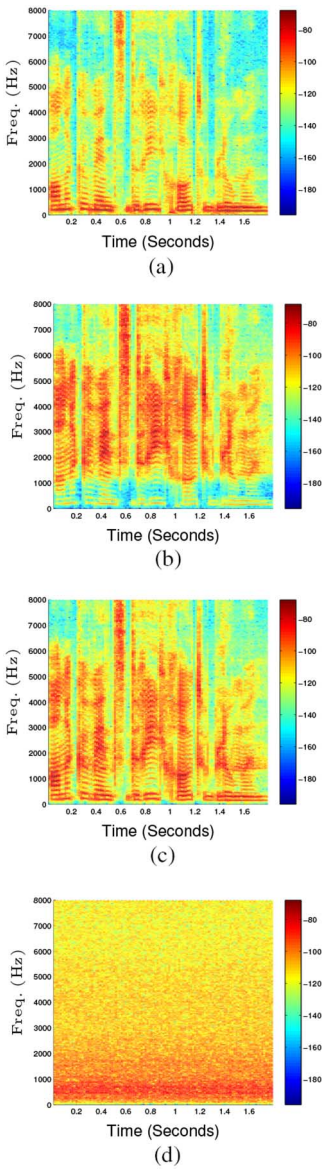

Fig. 2. Spectrogram of unprocessed signal, signal processed by Taal [19], the proposed approach, and the noise signal (speech shaped noise) (a) Unprocessed signal (b) Processed with Taal13 [19] (c) Processed with proposed approach (d) Noise signal.

presence of speech shaped noise will make the lower frequencies inaudible. Hence, to maximize the intelligibility, it is most effective to put energy in the higher frequencies. However, if the late reverberation is also taken into account, it becomes clear that amplifying speech will also introduce a certain amount of reverberation at the frequency band under consideration. The usefulness of amplifying speech saturates above a certain level, due to the introduction of late reverberation. In Fig. 2(c) we see therefore that the higher frequencies are also amplified, but to a less extent than in Fig. 2(b). The energy that is left is used to amplify the lower frequencies.

By increasing parameter B in the set $\mathcal{L}$, an effect similar as dynamic gain compression can be introduced. This is similar to the effect described in [30]. Hence, per time-frequency unit, B gains are available. One potentially unequal to zero, and $B-1$ gains that equal zero. Let $\sigma^2_{\mathcal{S}_p}(m,i)$ denote the variance of the processed speech in a time-frame m and critical band i. Combining the multiple gains per time-frequency unit by taking their average, the energy of the processed speech per time frame becomes $\sum_{\forall i}\sigma^2_{\mathcal{S}_p}(m,i)=\frac{r}{B}$. Setting $r=\sum_{\forall l\in\mathcal{L}}\sum_{\forall i}\sigma^2_{\mathcal{S}}(l,i)$, this implies that the energy per frame is set to the average energy over the last B time-frames. Dynamic range compression is known to have a positive effect on noisy speech, *e.g.*, [11], [22]. For reverberant speech a similar effect is to be expected as the resulting dynamic range compression will effectively work as a steady state suppressor where the energy in stationary high energy speech regions is somewhat reduced in favor of low energy transients.

V. Experimental Results

In this section we present experimental results on the presented intelligibility enhancement algorithm and compare to reference methods under several noisy and reverberant conditions. The comparisons are performed based on instrumental measures, as well as intelligibility listening tests. For the instrumental experiments we use more than 5 minutes of speech originating from the TIMIT database [40] sampled at 16 kHz. The speech signals are concatenated and degraded by stationary speech shaped noise at SNRs ranging from -10 dB up to 0 dB, where SNRs are measured between the original unprocessed speech as it would be played by the loudspeaker and the background noise (see Fig. 1 for a visualization of the experimental setup). The level of the speech is calibrated at 62.35 dB SPL, where 120 dB SPL indicates the maximum playback level. We assume that the recorded speech signal is noise free (i.e., either recorded in a noise-free environment, or, processed with a noise reduction algorithm). After processing the speech signal, leading to a signal s_p, the signal $x(n)$ is generated according to (1), where the convolution is performed in the DFT domain. In Sections V-A–V-C we demonstrate experimental results using generated room impulse responses, while in Section V-D experimental results are reported using measured room impulse responses.

Generation of room impulse responses is based on the Polack model [37] as in (13), where the exponential decay is given by $a=10^{-\frac{3}{T_{60}f_s}}$ and where we model the direct path by a delta impulse with height Δ and neglect early reflections. In all experiments in Sections V-A–V-C, Δ is set to $\Delta=1/d$ with $d=5$ the assumed distance between the loudspeaker and listener location. The pause τ between the direct path and the moment that the late reverberation starts is set to 50 ms based on values from literature [27]. The proposed approach is used on a frame-by-frame basis with 32 ms frames taken with 50% overlap and windowed with a square-root Hann window. The amplification factors $\alpha(m,i)$ are applied per critical band, maintaining the original phase of the clean speech signal. Some of the algorithms used in the experimental results depend on the noise PSD. To eliminate noise PSD estimation errors from the results, we use in all experiments an ideal voice activity detector and measure the noise PSD based on the noise-only signal (where any reverberation has decayed). Also, we assume the room dimensions and T60 reverberation time to be known, such that the diffuse room impulse response energy ρ can be computed.

To calculate ρ, we make use of the direct-to-reverberation ratio, which can be written as [41]

$$\frac{\Delta^2}{\rho^2} = \frac{A}{16\pi d^2}, \tag{43}$$

where d is the distance to the source, given by $\Delta = 1/d$, and A is the total absorption area. Given the T60 and the volume of the room, the total absorption area follows from Sabine's equation as (see e.g. [41])

$$A = \frac{24\log(10)V}{cT_{60}}, \tag{44}$$

with c the speed of sound and V the volume. Substituting (44) into (43), ρ^2 follows then as

$$\rho^2 = \frac{16\pi T_{60} c}{24\log(10)V}.$$

The volume is given by setting the room dimensions to approximately $10 \times 28 \times 4m^3$ ($L \times W \times H$).

The clean speech variance $\sigma_{\mathcal{S}}^2(m,i)$ for the proposed method is estimated by performing exponential smoothing across time with a smoothing constant $\beta = 0.996$, similar as in [16].

To measure the intelligibility improvements we use several instrumental speech intelligibility measures, among which the extended or short-time SII, denoted as ESII [32], $\mathrm{ASII_{ST}}$ [17] as in (6) and the speech based speech transmission index (STI) (sSTI) defined in [42] as the modified magnitude cross power spectrum based STI applied to running speech. The latter measure is chosen because it shows high correlation with the traditional STI for degradations with additive noise and reverberation [42]. Due to the very large frame sizes used in sSTI, this measure will show a relatively high variance. The $\mathrm{ASII_{ST}}$ measure is chosen as this is the measure that is being optimized in this paper, while the ESII is chosen as it is the measure that is approximated by $\mathrm{ASII_{ST}}$. Moreover, late reverberation can be argued to be a time-varying masker, and ESII is known to be a good predictor of intelligibility under time-varying noise maskers, which are uncorrelated with the target.

A. Influence of B

As a first experiment we investigate the influence of variable B in the set $\mathcal{L}$ on the various performance measures. To do so, we apply the algorithm outlined in Eqs. (40)-(42) for various values of B ranging from 1 up to 12 (corresponding to 32 ms up to 208 ms). The results are shown in Fig. 3 for T60 values of 1 second and 1.5 seconds, and speech shaped noise at input SNRs of -10 and 0 dB. From this we see that each performance measure shows some variability between the value of B and the performance. The optimal value of B depends on the experimental settings (T60 and SNR), but also on the used performance measure. Generally, performance is somewhat increased by increasing B from 1 up to 3 or 4. This is in line with the fact that Eq. (41) acts in practice as a dynamic range compressor similar as in [30]. When increasing B beyond $B = 4$, most performance measures show a slight decrease in performance. This can be explained by the fact that despite the beneficial dynamic compression, an increased B can also lead to speech sounds that get smeared out, introducing speech distortions.

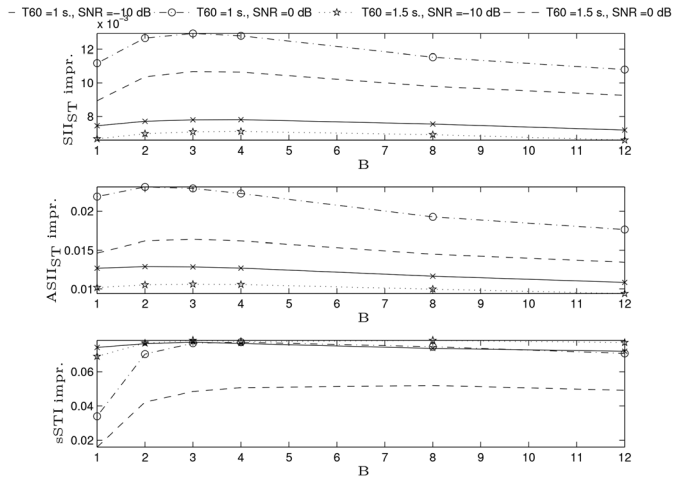

Fig. 3. Influence of variable B on various intelligibility measures for several settings of the T60 time and SNR.

Based on the experiments shown in Fig. 3 we use in the remaining experiments the values $B = 1$ and $B = 3$. We use $B = 1$ as for $\rho = 0$ this would lead as a special case to the algorithm proposed in [19].

B. Instrumental Comparison to Reference Methods

In this section we present experimental results with the proposed $\mathrm{ASII_{ST}}$ speech reinforcement algorithm and compare this to the ASII optimal algorithm published in [19], the SII optimal algorithm published in [16] and the steady state suppressor of Hodoshima *et al.* [29] referred to as Taal13, Sauert10, and Hodo06, respectively. Both reference algorithms Taal13 and Sauert10 do not explicitly take reverberation into account, while Hodo06 reduces overlap-masking that degrades speech intelligibility in reverberation. For all algorithms we make sure that they obey a constraint on the average energy per time frame. Similar to the proposed approach and Sauert10, we also use exponential smoothing to measure the speech variance $\sigma_{\mathcal{S}}^2(m,i)$ in reference method Taal13.

In the experiments as depicted in Figs. 4–6 we show a comparison in terms of intelligibility improvement over the unprocessed signal measured by the different instrumental intelligibility measures as a function of the T60 ranging from 0 seconds to 2 seconds. The results are depicted for input SNRs -10, -5 and 0 dB.

Compared to the three reference methods, the proposed method generally improves the predicted intelligibility with a maximum improvement in the order of a few percentage points. When there is no reverberation ($\rho = 0$), the proposed method and the method from [19] are in theory identical for $B = 1$. This is indeed also reflected in the experiments by the fact that for very small T60-values, the performance of these approaches coincide. With increasing T60, the improvement of the proposed method with $B = 1$ over Taal13 increases slightly as a function of T60. This is due to the fact that part of the reverberation is taken into account in determining the amplification per critical band. Generally, performance for the

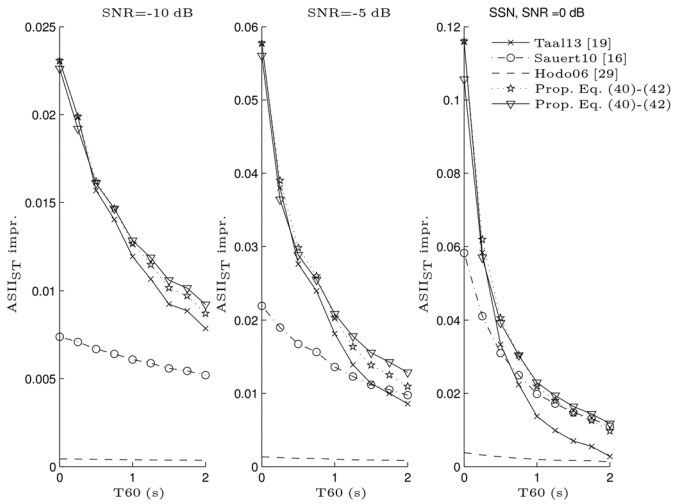

Fig. 4. Predicted intelligibility in terms of $\mathrm{ASII_{ST}}$ improvement for speech degraded by speech shaped noise.

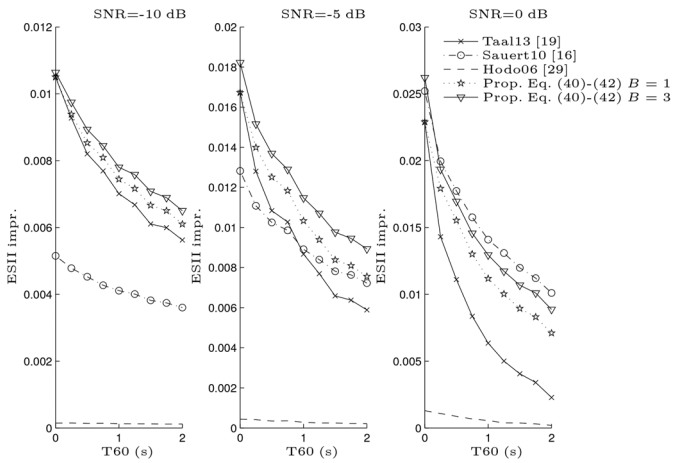

Fig. 5. Predicted intelligibility in terms of ESII improvement for speech degraded by speech shaped noise.

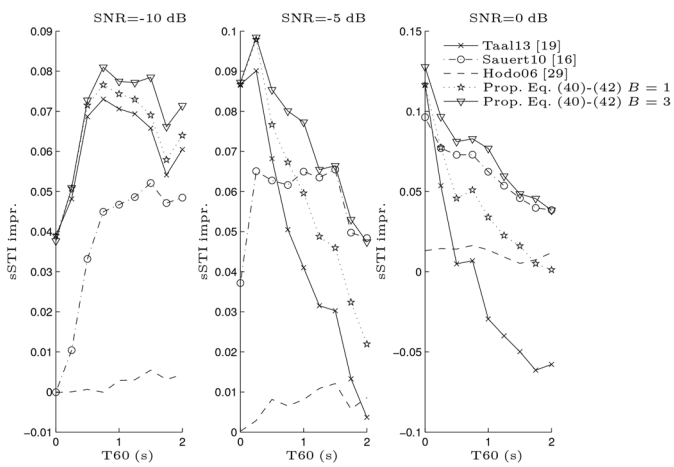

Fig. 6. Predicted intelligibility in terms of sSTI improvement for speech degraded by speech shaped noise.

proposed method with $B = 3$ improves over the proposed method with $B = 1$. This is in line with the evaluation performed in Section V-A.

For all three performance measures Hodo06 leads to the worst performance, *i.e.*, hardly any improvement over the unprocessed signal. Although this is very consistent across all performance measures and experimental settings, it could still be that the type of processing as performed by Hodo06 cannot very well be assessed by instrumental intelligibility measures. Intelligibility tests should give a definite answer. The algorithm Sauert10 presented in [16] optimizes the ESII measure under certain approximations. We therefore expect Sauert10 to perform well when measuring ESII. This is indeed visible for the higher SNRs, e.g. 0 dB in Fig. 5, where Sauert10 even improves over the proposed method for $B = 3$. For lower SNRs (*e.g.* -10 and -5 dB), Sauert10 performs somewhat worse than the proposed method in terms of ESII.

Although the results in Figs. 4–6 are carried out using only speech shaped noise, similar results have been obtained using other noise sources (among which babble noise), which have been left out here because of redundancy.

C. Intelligibility Listening Test

In this section, we compare the presented algorithm with $B = 3$ with the reference methods Taal13, Sauert10 and Hodo06 by means of an intelligibility test. In addition, we also test the intelligibility of the unprocessed signal.

The intelligibility test that was conducted is a closed Dutch speech-in-noise intelligibility test described in [43] that we use here as an intelligibility test for speech under noisy reverberant conditions. This intelligibility test consists of five-word sentences with a correct grammatical structure, similar to the one proposed by Hagerman in [44]. The sentences were sampled at a sampling frequency of 16 kHz. The possible words are arranged in an 10-by-5 matrix on a computer screen, such that the ith column contains exactly the 10 possible alternatives for the ith word. The listener selects via a graphical user interface for each test sentence one word from each column. After being processed by one of the aforementioned algorithms, the speech signals are convolved with a room impulse response that is generated according to the Polack model with a T60 time of 1 second and degraded by speech shaped noise at SNRs of -2, 0, 2 and 4 dB.

Eight native Dutch speaking subjects participated in the test. The order of presenting the different algorithms and the SNRs was randomized, with each sentence being used only once. With each test person, all processing conditions were repeated four times. The signals were presented diotically through head-phones (Sennheiser HD 600).

Fig. 7 shows the average intelligibility scores with standard errors of the mean. From these results we see that the proposed method improves under all conditions over all reference methods, as well as over the unprocessed signal. A t-test [45] with a significance level $\alpha = 0.05$ was performed to determine whether differences are statistical significant. From this t-test it follows that the proposed method with $B = 3$ is always significantly better than Hodo06 and the unprocessed signals. Compared to Sauert10, the proposed algorithm is significantly better for all SNRs, except for the 0 dB condition. Furthermore, compared to Taal13, the proposed method is significantly better for all SNRs, except at the SNR of 4 dB.

Generally, Hodo06 performs worst and sometimes degrades performance compared to the unprocessed condition (e.g., at the SNRs of 0 and 2 dB). However, the performance is less dramatic

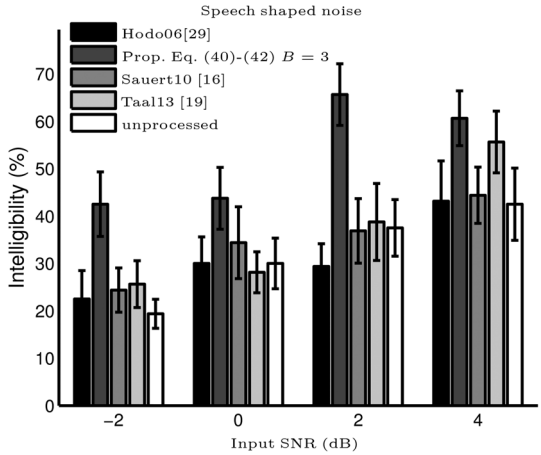

Fig. 7. Intelligibility listening test results.

as might appear from the instrumental measures. This probably indicates that the instrumental measures cannot exactly predict the consequences of the modifications introduced by Hodo06. Also, notice that the instrumental intelligibility scores are not mapped to an actual percentage of correct understood words, which partly explains the difference in the scores between Section V-C and V-B.

D. Evaluation Using Measured Impulse Responses

In this section we present an evaluation based on measured room impulse responses instead of generated room impulse responses as was done in the Section V-B and V-C. These measured room impulse responses do not exactly match the assumptions made in the derivations on the proposed approach and serve as an indication on how sensitive the proposed approach is with respect to a model mismatch. The used room impulse responses all originate from the Aachen impulse response (AIR) database described in [46]. From this database we use three impulse responses, namely, the impulse response measured in the Aula Carolina in Aachen at 3 m distance ($T_{60} = 3.3$ sec.), an impulse response measured in a stairway hall at 3 m distance ($T_{60} = 1.1$ sec.) and an impulse response measured in a lecture room at 4 m distance ($T_{60} = 0.82$ sec.).

The proposed algorithm depends on three impulse response related parameters, that are, ρ^2, Δ and a. Given a measured impulse response, the diffuse response energy ρ^2 can be calculated from the impulse response as $\rho^2 = \sum_{l=n_0}^{+\infty} |h(l)|^2 \frac{1}{f_s}$, with f_s the sampling frequency. The exponential decay can be calculated as $a = 10^{-\frac{3}{T_{60} f_s}}$ as defined in the Polack model [37], and damping Δ from loudspeaker to listener location can be estimated as $\Delta = \|s_p\|/\|s_{p,dir}\|$, where s_p is the (processed) signal as played by the loudspeaker, and $s_{s,dir}$ is the direct path signal as received at the listener location.

For evaluation we use the same measures as in the previous section, that are, ESII, ASII_{ST} sSTI, where the reference is chosen to be the direct path signal $s_{s,dir}$ as received at the listener location.

As a first evaluation we present in Fig. 8 a similar example as in Fig. 2, Section IV. Fig. 8 shows the original clean speech spectrogram (Fig. 8(a)), the spectrogram of the speech signal processed by Taal [19] (Fig. 8(b)) (without taking reverberation into account), the spectrogram of the speech signal processed

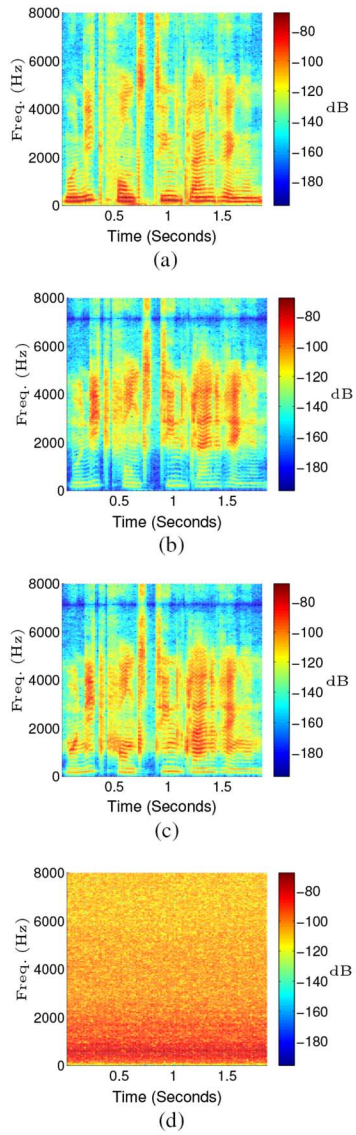

Fig. 8. Spectrogram of (a) unprocessed signal, | (b) signal processed by Taal [19], (c) the proposed approach, and (d) the noise signal (speech shaped noise), for an impulse response measured in the Aula Carolina in Aachen ($T_{60} = 3.3$ sec.).

by the proposed approach with $B = 3$ (Fig. 8(c)), and the noise signal (Fig. 8(d)), which consists of speech shaped noise. The SNR is in this example set to -5 dB and as measured impulse response we use the response of the Aula Carolina in Aachen.

In Fig. 8 we observe a similar behavior as in Fig. 2. We see that if reverberation is not taken into account (Fig. 8(b)) generally most energy is put at higher frequencies. However, taking reverberation into account, the efficiency of a channel saturates due to the generated reverberation. In this specific example, this will lead to less amplification of the higher frequency bands where this energy is in turn used at the lower frequency bands.

Finally, in Figs. 9–11 we show an instrumental evaluation using measured impulse responses originating from a stairway hall, the Aula Carolina in Aachen and a lecture room, respectively, as a function of SNR. These instrumental evaluations are in line with the evaluations from Section V-B that were carried out using generated impulse responses. In all cases, the proposed approach, Taal13 and Sauert10 improve over

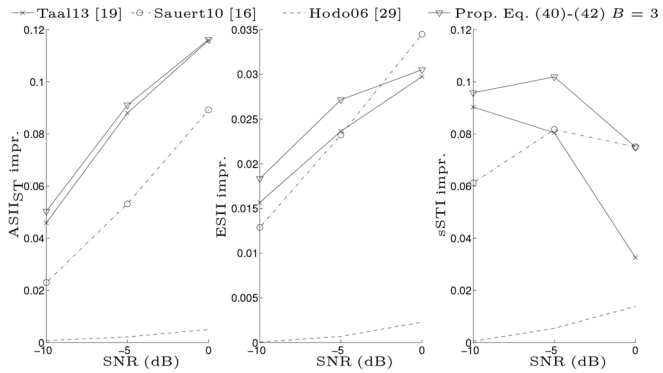

Fig. 9. Predicted intelligibility based on improvement in terms of $\mathrm{ASII_{ST}}$, ESII, and sSTI for speech degraded by speech shaped noise and reverberation generated using an impulse response measured in a stairway hall ($T_{60} = 1.1$ sec.).

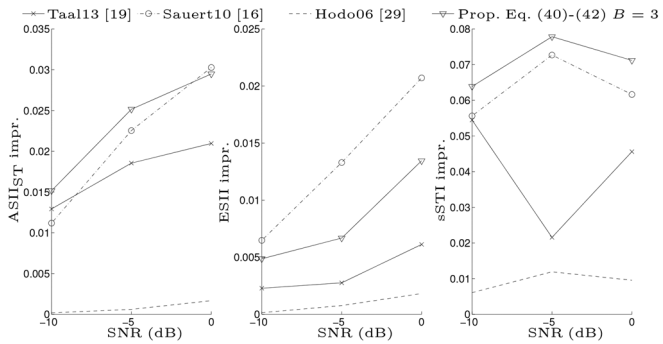

Fig. 10. Predicted intelligibility based on improvement in terms of $\mathrm{ASII_{ST}}$, ESII, and sSTI for speech degraded by speech shaped noise and reverberation generated using an impulse response measured in the Aula Carolina in Aachen ($T_{60} = 3.3$ sec.).

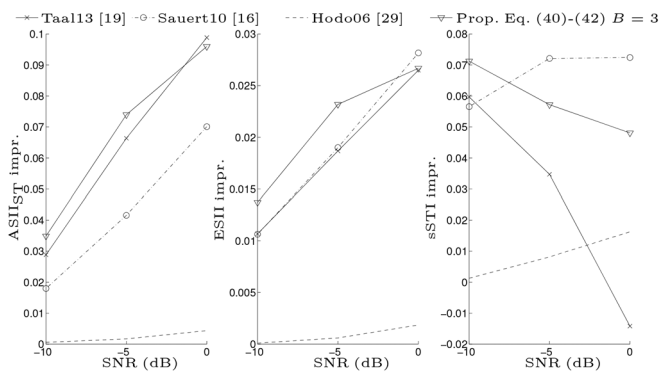

Fig. 11. Predicted intelligibility based on improvement in terms of $\mathrm{ASII_{ST}}$, ESII, and sSTI for speech degraded by speech shaped noise and reverberation generated using an impulse response measured in a lecture room ($T_{60} = 0.82$ sec.).

Hodo06. Similarly as for the experiments with generated impulse responses, the proposed approach generally improves over Taal13. With respect to Sauert10, the proposed approach shows typically improvements in terms of $\mathrm{ASII_{ST}}$. Similar as in Section V-B, Sauert10 exhibits sometimes slight improvements over the proposed approach at higher SNRs when measuring performance using ESII.

VI. Conclusions

In this paper we investigated how late reverberation and noise can be taken into account simultaneously when improving the speech intelligibility in the near-end speech enhancement application. To do so, we built further upon a recently proposed approximation of the speech intelligibility index (SII). This approximation facilitates constrained convex optimization. To be able to take reverberation into account, we locally optimized the approximated SII and made use of the Polack model to model the late reverberation. The optimization resulted in an algorithm that delivers amplification factors for each critical band and time frame. These amplification factors depend on the both the noise PSD as well as the late reverberation energy, and redistribute speech energy across frequency in order to increase intelligibility when exposed to noise and reverberation. Depending on the settings of the algorithm, it also performs dynamic range compression, which is known to be beneficial for intelligibility enhancement.

Instrumental intelligibility experiments, as well as intelligibility listening tests under noisy reverberant conditions showed that the proposed algorithm improves speech intelligibility, with maximum improvement over state-of-the-art reference algorithms in the order of 20 percent.

Appendix

In this appendix we proof the existence and uniqueness of $\nu > 0$ in (42). Without loss of generality, we set $\rho = 0$ and $\Delta = 1$ for notational convenience. We then get similar to (42)

$$\sum_i \max\left(\frac{\sigma_{\mathcal{W}_i}\sqrt{\gamma_i}}{\sqrt{\nu}} - \sigma^2_{\mathcal{W}_i}, 0\right) = r. \tag{45}$$

Indeed, it can be shown that (45) has a solution for $\nu > 0$, and that this solution is in fact unique. To see this, note that the function of ν at the left hand side of (45), $q : \mathbb{R}^+ \mapsto \mathbb{R}$ defined by

$$q(\nu) = \sum_i \max\left(\frac{\sigma_{\mathcal{W}_i}\sqrt{\gamma_i}}{\sqrt{\nu}} - \sigma^2_{\mathcal{W}_i}, 0\right), \tag{46}$$

is continuous and non-increasing. Furthermore, in a neighborhood of a root $\nu \in \mathcal{B}_\epsilon(\nu_0) = [\nu_0 - \varepsilon, \nu_0 + \epsilon]$, $q(\nu_0) = r$, with $\epsilon > 0$ small, at least one of the terms of the sum in (46) is strictly positive, since $r > 0$. For that term i_0, the maximum operator is in the active zone, *i.e.*, we have $\sigma_{\mathcal{W}_{i_0}}\sqrt{\gamma_{i_0}/\nu} - \sigma^2_{\mathcal{W}_{i_0}} > 0$. The strict monotonicity of the maximum operator in the active zone then implies that $q(\nu)$ is also strictly monotonous in $\mathcal{B}_\epsilon(\nu_0)$ (strictly decreasing). Joining this information with the fact that $q(0) = +\infty > r$ and $q(+\infty) = 0 < r$ guarantees the existence of a root by the intermediate value theorem and uniqueness by the strict monotone behavior.

Acknowledgment

The authors wish to thank the anonymous reviewers for their constructive comments which helped to improve the presentation of this work.

References

[1] P. Loizou, *Speech Enhancement Theory and Practice*. Boca Raton, FL, USA: CRC, 2007.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*. San Rafael, CA, USA: Morgan & Claypool, 2013.

[3] J. Benesty and M. M. Sondhi, , Y. Huang, Ed., *Springer Handbook of Speech Processing*. New York, NY, USA: Springer, 2008.

[4] M. Brandstein, , D. Ward, Ed., *Microphone arrays: Signal processing techniques and applications*. New York, NY, USA: Springer, 2001.

[5] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[6] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, Sep. 2007.

[7] K. Eneman *et al.*, "Evaluation of signal enhancement algorithms for hearing instruments," in *Proc. EURASIP Eur. Signal Process. Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *J. Acoust. Soc. Amer.*, vol. 130, no. 5, pp. 3013–3027, 2011.

[9] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, Jan. 2012.

[10] J. D. Griffiths, "Optimum linear filter for speech transmission," *J. Acoust. Soc. Amer.*, vol. 43, p. 81, 1968.

[11] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 277–282, Aug. 1976.

[12] C. Tantibundhit, J. R. Boston, C. C. Li, J. D. Durrant, S. Shaiman, K. Kovacyk, and A. El-Jaroudi, "New signal decomposition method based speech enhancement," *Signal Process.*, vol. 87, pp. 2607–2628, 2007.

[13] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, vol. 1, pp. 493–496.

[14] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, 2009, vol. 17, pp. 1844–1848.

[15] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proc. EURASIP Eur. Signal Process. Conf. (EUSIPCO)*, 2010, pp. 1919–1923.

[16] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachtagung Sprachkommun*. Berlin, Germany: VDE VERLAG GmbH, 2010.

[17] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 4061–4064.

[18] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. ISCA Interspeech*, 2012.

[19] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[20] V. Aubanel and M. Cooke, "Information-preserving temporal reallocation of speech in the presence of fluctuating maskers," in *Proc. ISCA Interspeech*, 2013, pp. 3592–3596.

[21] C. Valentini-Botinhao, J. Yamagishi, S. King, and R. Maia, "Intelligibility enhancement of hmm-generated speech in additive noise by modifying mel cepstral coefficients to increase the glimpse proportion," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 665–686, 2014.

[22] M. Cooke, C. Mayoe, and C. Valentini-Botinhao, "Intelligibility enhancing speech modifications: The hurricane challenge," in *Proc. ISCA Interspeech*, 2013.

[23] J. L. Hall and J. L. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *J. Acoust. Soc. Amer.*, vol. 127, pp. 280–285, 2010.

[24] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[25] *American National Standard Methods for the Calculation of the Speech Intelligibility index*, ansi S3.5-1997 ed., Amer. Nat. Stand. Inst..

[26] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1553–1564, Jul. 2012.

[27] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[28] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *ELSEVIER Speech Commun.*, vol. 45, no. 2, pp. 101–113, 2005.

[29] N. Hodoshima, T. Arai, A. Kusumot, and K. Kinoshita, "Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments," *J. Acoust. Soc. Amer.*, vol. 119, no. 6, pp. 4055–4064, 2006.

[30] J. B. Crespo and R. C. Hendriks, "Speech reinforcement in noisy reverberant environments using a perceptual distortion measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2014, pp. 910–914.

[31] J. Crespo and R. C. Hendriks, "Multizone speech reinforcement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 54–66, Jan. 2014.

[32] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[33] A. C. Neuman, M. Wroblewski, J. Hajicek, and A. Rubinstein, "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," *Ear Hear.*, vol. 31, no. 3, pp. 336–344, 2010.

[34] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[35] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process.*, 2010, pp. 4266–4269.

[36] J. S. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1746–1765, Sep. 2010.

[37] J. D. Polack, "La transmission de l'energie sonore dans les salles," Ph.D. dissertation, Universitè du Maine, Le Mans, France, 1988.

[38] W. Karush, "Minima of functions of several variables with inequalities as side constraints," M.S. thesis, Dept. of Math., Univ. of Chicago, Chicago, IL, USA, 1939H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proc. 2nd Berkeley Symposium*. Berkeley, CA, USA: Univ. of California Press, 1951, pp. 481–492.

[39] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proceedings of 2nd Berkeley Symposium*. Berkeley, CA, USA: Univ. of California Press, 1951, pp. 481–492.

[40] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," National Institute of Standards and Technology (NIST), 1988.

[41] M. J. Crocker, *Handbook of noise and vibration control*. New York, NY, USA: Wiley, 2007.

[42] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[43] R. Houben, J. Koopman, H. Luts, K. C. Wagener, A. van Wieringen, H. Verschuure, and W. A. Dreschler, "Development of a dutch matrix sentence test to assess speech intelligibility in noise," *Int. J. Audiol., Early Online*, vol. 127, pp. 1–4, 2014.

[44] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scand. Audiol.*, vol. 11, no. 2, pp. 79–87, 1982.

[45] D. J. Sheskin, *Parametric and Nonparametric Statistical Procedures*, 3rd ed. Boca Raton, FL, USA: Chapman & Hall/CRC, 2004.

[46] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. Digital Signal Process.*, 2009, pp. 1–5.

[47] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Speech reinforcement in noisy reverberant conditions under an approximation of the short-time SII," in *Proc. IEEE ICASSP*, 2015.

[48] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Comput. Speech Lang.*, vol. 28, no. 4, pp. 858–872, 2014.

Richard C. Hendriks obtained his M.Sc. and Ph.D. degrees (both cum laude) in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2003 and 2008, respectively. From 2003 till 2007, he was a Ph.D. Researcher at Delft University of Technology, Delft, The Netherlands. From 2007 till 2010, he was a Postdoctoral Researcher at Delft University of Technology. Since 2010, he has been an Assistant Professor in the Signal and Information Processing Lab of the faculty of Electrical Engineering, Mathematics and Computer Science at Delft University of Technology. In the autumn of 2005, he was a Visiting Researcher at the Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany. From March 2008 till March 2009, he was a Visiting Researcher at Oticon A/S, Copenhagen, Denmark. His main research interests are digital speech and audio processing, including single-channel and multi-channel acoustical noise reduction, speech enhancement, and intelligibility improvement.

João B. Crespo received his M.Sc. in electrical engineering from the Technical University of Lisbon, Portugal, in 2009. During the last year of his M.Sc., he was an exchange student at the Signal and Information Processing Lab (currently Circuits and Systems group) of the faculty of Electrical Engineering, Mathematics, and Computer Science at Delft University of Technology, The Netherlands. After enjoying working experience as a DSP Developer at ExSilent B.V., The Netherlands, he researched source-based listening enhancement at the Circuits and Systems group at Delft University of Technology. Currently, he is pursuing an M.Sc. in mathematics at VU University, Amsterdam.

Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is a Senior Researcher with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is also a Professor with the Section for Signal and Information Processing, Department of Electronic Systems, at Aalborg University. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.

Cees H. Taal received the B.S. and M.A. degrees in arts and technology from the Utrecht School of Arts, Utrecht, The Netherlands, in 2004 and the M.Sc. degree in computer science from the Delft University of Technology (DUT), Delft, The Netherlands, in 2007. From 2008 to 2012, he was a Ph.D. Researcher in the Multimedia Signal Processing Group, DUT, under the supervision of R. Heusdens and R. C. Hendriks in collaboration with Oticon A/S. From 2012 to 2013, he held Postdoc positions at the Sound and Image Processing Lab, Royal Institute of Technology (KTH), Stockholm, Sweden and the Leiden University Medical Center (LUMC), Leiden, the Netherlands. Since 2014, he has been with Philips Research, Eindhoven, The Netherlands, working on signal processing solutions applied to optical heart-rate and accelerometer signals as measured by wearable sensor technologies. His main research interests are in the field of audio, speech, and biomedical digital signal processing.