Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay

Timo Gerkmann, *Member, IEEE*, and Richard C. Hendriks

***Abstract*—Recently, it has been proposed to estimate the noise power spectral density by means of minimum mean-square error (MMSE) optimal estimation. We show that the resulting estimator can be interpreted as a voice activity detector (VAD)-based noise power estimator, where the noise power is updated only when speech absence is signaled, compensated with a required bias compensation. We show that the bias compensation is unnecessary when we replace the VAD by a soft speech presence probability (SPP) with fixed priors. Choosing fixed priors also has the benefit of decoupling the noise power estimator from subsequent steps in a speech enhancement framework, such as the estimation of the speech power and the estimation of the clean speech. We show that the proposed speech presence probability (SPP) approach maintains the quick noise tracking performance of the bias compensated minimum mean-square error (MMSE)-based approach while exhibiting less overestimation of the spectral noise power and an even lower computational complexity.**

***Index Terms*—Noise power estimation, speech enhancement.**

I. Introduction

As DIGITAL speech communication devices, such as hearing aids or mobile telephones, have become more and more portable, usage of these applications in noisy environments occurs on a more frequent basis. Depending on the environment, the noise signal that corrupts the target speech signal can be quite nonstationary. These nonstationary noise corruptions can originate for example from a train that passes by at a train station or from passing cars and other people when communicating while walking along the street. The aim of speech enhancement algorithms is to reduce the additive noise without decreasing speech intelligibility. Most speech enhancement algorithms try to accomplish this by applying a gain function in a spectral domain, where the gain function is generally dependent on the noisy spectral coefficient, the spectral noise power and the spectral speech power. In [1], a gain function was proposed based on maximum-likelihood estimation, assuming a deterministic (unknown) model for the speech spectral coefficients and a complex Gaussian distribution for the noise spectral coefficients. While the speech model in [1] was assumed to be deterministic, it was proposed in [2] to model both the speech and noise spectral coefficients by complex Gaussian distributions and to estimate the speech spectral magnitude coefficients by minimizing the mean-square error (MSE) between the clean and estimated speech spectral magnitude. This was succeeded by the work presented in [3], where it was proposed to minimize the mean-square error MSE between the logarithms of the clean and estimated speech spectral magnitude, motivated by the idea that a mean-squared error between logarithms of magnitude spectra is perceptually more meaningful. Based on the observation that the observed distribution of speech spectral coefficients tends to be more super-Gaussian than Gaussian, see, e.g., [4] and [5], further improvements of the estimators presented in [2], [3] were obtained in [4]–[7], where it was proposed to derive Bayesian estimators under super-Gaussian distributions. However, all these methods have in common that they are a function of both the spectral noise power and the spectral speech power. The spectral noise and speech power are generally unknown and are to be estimated from the noisy data. Estimation of the spectral speech power can be done by employing the decision-directed approach [2], see [8]–[10] for detailed analyses, non-causal recursive *a priori* SNR estimators [11], or cepstral smoothing techniques [12].

In this paper, we focus on estimation of the spectral noise power. As the noise power may change rapidly over time, its estimate has to be updated as often as possible. Using an overestimate or an underestimate of the true, but unknown, spectral noise power will lead to an over-suppression or under-suppression of the noisy signal and might lead to a reduced intelligibility or an unnecessary amount of residual noise when employed in a speech enhancement framework. One way to estimate the spectral noise power is to exploit time instances where speech is absent. This requires detection of speech presence by means of a voice activity detector (VAD), see, e.g., [13], [14]. However, in nonstationary noise scenarios this detection is particularly difficult, as a sudden rise in the noise power may be misinterpreted as a speech onset. In addition, if the noise spectral power changes during speech presence, this change can only be detected with a delay.

To improve estimation of the spectral noise power, several approaches have been proposed during the last decade. Among the most established estimators are those based on minimum statistics (MS) [15]–[17]. In [15], the power spectrum of the noisy

Manuscript received May 30, 2011; revised August 18, 2011 and November 24, 2011; accepted November 29, 2011. Date of publication December 21, 2011; date of current version February 24, 2012. The research leading to these results was supported in part by the European Community's Seventh Framework Program under Grant Agreement PIAP-GA-2008-214699 AUDIS and in part by the Dutch Technology Foundation STW. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hui Jiang.

T. Gerkmann is with the Speech Signal Processing Group, Universität Oldenburg, 26111 Oldenburg, Germany (e-mail: timo.gerkmann@uni-oldenburg.de).

R. C. Hendriks is with the Signal and Information Processing Lab, Delft University of Technology, 2628 CD Delft, The Netherlands.

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/TASL.2011.2180896

signal is estimated on a frame-by-frame basis and observed over a time-span of about 1–3 seconds. In general, MS based spectral noise power estimators are based on the assumption that within the observed time-span, speech is absent during at least a small fraction of the total time-span. The spectral noise power is then obtained from the minimum of the estimated power spectrum of the noisy signal. However, if the noise power rises within the observed time-span, it will be underestimated or tracked with a certain delay. The worst case amount of delay generally depends on the length of the used time-span. The shorter the time-span, the shorter the maximum delay. However, decreasing the time-span increases the chance that speech is not absent within this observed time-span. The consequence of this is that the spectral noise power may be overestimated, as the estimator might track instances of the noisy spectral power instead of the noise spectral power. Thus, in [15, Sec. VI] mechanisms are proposed that allow for a tracking of rising noise powers also within the observed time-span. However, rising noise powers as caused, e.g., by passing trains, are often still tracked with a rather large delay of around one second. The local underestimation of the noise power is likely to result in annoying artifacts, like residual noise and *musical noise*, when the noise power estimate is applied in a speech enhancement framework.

The methods of Sohn and Sung [18], Cohen [16], and Rangachari and Loizou [17] are based on a recursive averaging of the noisy spectral power using the speech presence probability (SPP) which is obtained from the ratio of the likelihood functions of speech presence and speech absence. As opposed to the likelihood of speech absence, the likelihood of speech presence is parameterized by the *a priori* SNR. In case the *a priori* SNR is zero, the two likelihood functions overlap such that a distinction between speech presence and absence based on the likelihood ratio is not possible. In [18] and [16], the *a priori* signal-to-noise ratio (SNR) is estimated adaptively on a short time scale. As a consequence, in speech absence the adapted *a priori* SNR estimate is close to zero, and the two likelihood functions overlap such that, independent of the observation, the likelihood ratio is one. The resulting *a posteriori* SPP yields only the *a priori* SPP which, per definition, is also independent of the observation. Thus, without further modifications, a detection of speech absence in each time–frequency point is not possible. This problem is partly overcome in [18] by considering the joint likelihood function of an entire speech segment. Thus, the SPP estimate of [18] is frequency independent, such that the ability to track the noise power between speech spectral harmonics is lost. In [16], low values for the *a posteriori* SPP are enabled by an additional adaptation of the *a priori* SPPs with respect to the observation. However, as the methods in [16] and [17] are based on MS principles, they also show a delayed tracking of rising spectral noise powers similarly to [15].

The more recent contributions on the topic of spectral noise power estimation generally focus on tracking of the spectral noise power with a shorter delay, in order to improve noise reduction in environments with nonstationary noise. Some examples are the discrete Fourier transform (DFT)-subspace approach [19], or minimum mean-square error (MMSE)-based approaches [20], [21]. Although DFT-subspace-based approaches lead to quite some improvement for non-stationary noise sources compared to, e.g., MS-based spectral noise power estimators [22], computationally they are rather demanding. The MMSE-based algorithm [21] on the other hand is computationally much less demanding and at the same time robust to increasing noise levels as shown in a comparison presented in [22]. In the MMSE-based estimator [21], first a limited maximum-likelihood (ML) estimate of the *a priori* SNR is used to obtain an MMSE estimate of the noise periodogram. However, under the given *a priori* SNR estimate the resulting MMSE estimate exhibits a bias which can be computed analytically. However, in order to compensate for the bias, a second estimate of the *a priori* SNR is required, for which the decision-directed approach [2] is used.

In this work, we analyze the noise power estimator of [21], and show that under the given ML *a priori* SNR estimator the MMSE-based spectral noise power estimator can be interpreted as a VAD-based estimator. To improve the MMSE-based spectral noise power estimator we modify the original algorithm such that it evolves into a soft SPP instead of a hard SPP (i.e., VAD) based estimator, which automatically makes the estimator unbiased. The proposed estimator exhibits a computational complexity that is even lower than that of the MMSE-based approach [21] while maintaining its fast noise tracking performance without requiring a bias compensation.

As opposed to the SPP-based noise power estimators of [16] and [18], we use a fixed nonadaptive *a priori* SNR as a parameter of the likelihood of speech presence. This fixed *a priori* SNR represents the SNR that is typical in speech presence and prevents the likelihood functions from overlapping in speech absence. Thus, using the fixed *a priori* SNR enables the time–frequency dependent *a posteriori* SPP to yield values close to zero in speech absence without adapting the *a priori* SPP. Further, as opposed to [16], [17] the tracking delay remains small, as we do not use minimum statistics (MS) principles.

This work is organized as follows. After explaining the used notation and assumptions in Section II, we review the MMSE-based noise power estimator of [21], analyze its bias correction behavior and show that the estimator can be interpreted as a VAD-based noise power estimator in Section III. Then, in Section IV, we propose to replace the VAD implicitly used in [21] by a soft SPP with fixed priors. In Section V, we show that the proposed estimator results in similar or better results in nonstationary noise than competing algorithms, while exhibiting a lower computational complexity. While the basic idea of this work has been published in [23], in this paper we present a more detailed analysis, derivation and evaluation.

II. Signal Model

In this work, we consider a frame-by-frame processing of time-domain signals, where the windowed time-domain frames are transformed to the spectral domain by applying a DFT. Let the complex spectral speech and noise coefficients be given by $S_k(l)$ and $N_k(l)$, with k the frequency-bin index and l the time-frame index. We assume the speech and the noise signals to be additive in the short-time Fourier domain. The complex spectral noisy observation is thus given by $Y_k(l) = S_k(l) + N_k(l)$.

For notational convenience, the time-frame index l and the frequency index k will be left out, unless necessary for clarification. Random variables are denoted by capital letters, realizations by its corresponding lower case letters, and estimated quantities are denoted by a hat symbol, e.g., $\widehat{\xi}$ is an estimate of ξ. We assume that the speech and noise signals have zero mean and are independent so that $\mathrm{E}(|Y|^2) = \mathrm{E}(|S|^2) + \mathrm{E}(|N|^2)$, with $\mathrm{E}(\cdot)$ being the statistical expectation operator. The spectral speech and noise power are defined by $\mathrm{E}(|S|^2) = \sigma_{\mathrm{S}}^2$ and $\mathrm{E}(|N|^2) = \sigma_{\mathrm{N}}^2$, respectively. We then define the *a posteriori* SNR by $\gamma = |y|^2/\sigma_{\mathrm{N}}^2$ and the *a priori* SNR by $\xi = \sigma_{\mathrm{S}}^2/\sigma_{\mathrm{N}}^2$.

III. Review of MMSE-Based Noise Power Estimation

In order to guide the reader and help to appreciate later contributions in this paper, we first present in this section a review on the MMSE-based noise power estimator presented in [21]. To derive the MMSE-based noise power estimator it is assumed that the noise and speech spectral coefficients have a complex Gaussian distribution, i.e,

$$p_N(n) = \frac{1}{\sigma_{\mathrm{N}}^2\pi}\exp\left(-\frac{|n|^2}{\sigma_{\mathrm{N}}^2}\right) \tag{1}$$

$$p_S(s) = \frac{1}{\sigma_{\mathrm{S}}^2\pi}\exp\left(-\frac{|s|^2}{\sigma_{\mathrm{S}}^2}\right). \tag{2}$$

With these assumptions we obtain

$$p_Y(y) = \frac{1}{\sigma_{\mathrm{N}}^2(1+\xi)\pi}\exp\left(-\frac{|y|^2}{\sigma_{\mathrm{N}}^2(1+\xi)}\right). \tag{3}$$

For mathematical convenience we will use a polar notation for the complex spectral noise and noisy speech coefficients, that is, $N = De^{j\Delta}$ and $Y = Re^{j\Theta}$. Using this notation we can transform the distribution $p_N(n)$ of the spectral noise coefficients into polar coordinates, that is,

$$p_{D,\Delta}(d,\delta) = \frac{d}{\pi\sigma_N^2}\exp\left(\frac{d^2}{\sigma_N^2}\right). \tag{4}$$

Further, it follows from (2) in combination with the additivity and independence assumption of speech and noise that the distribution $p_{Y|D,\Delta}(y|d,\delta)$ is given by

$$p_{Y|D,\Delta}(y|d,\delta) = \frac{1}{\pi\sigma_S^2}\exp\left(\frac{2dr\cos(\delta-\theta)-r^2-d^2}{\sigma_S^2}\right). \tag{5}$$

The noise power estimators presented in [20] and [21] are based on an MMSE estimate of the noise periodogram, which can be obtained by computing the conditional expectation $\mathrm{E}(|N|^2|y)$. Using Bayes' rule, this can be written as

$$\mathrm{E}\left(|N|^2|y\right) = \frac{\int_0^{+\infty}\int_0^{2\pi} d^2 p_{Y|D,\Delta}(y|d,\delta)p_{D,\Delta}(d,\delta)\mathrm{d}\delta\mathrm{d}d}{\int_0^{+\infty}\int_0^{2\pi} p_{Y|D,\Delta}(y|d,\delta)p_{D,\Delta}(d,\delta)\mathrm{d}\delta\mathrm{d}d}. \tag{6}$$

Substituting (5) and (4) into (6) and using [24, Eqs. 8.431.5 and 6.643.2], we obtain

$$\mathrm{E}\left(|N|^2|y\right) = \left(\frac{1}{1+\widehat{\xi}}\right)^2|y|^2 + \frac{\widehat{\xi}}{1+\widehat{\xi}}\widehat{\sigma_{\mathrm{N}}^2} \tag{7}$$

where we have written $\mathrm{E}(|N|^2|y)$ as a function of the estimates $\widehat{\xi}$ and $\widehat{\sigma_{\mathrm{N}}^2}$ of the *a priori* SNR and spectral noise power, respectively, to explicitly show that these quantities have to be estimated in practice. In noise power estimation, it is a common assumption that the noise signal is more stationary than the speech signal [15]. Assuming a certain degree of correlation between the noise power present in neighboring signal segments, it is reasonable to use the spectral noise power estimate of the previous frame in (7), i.e., $\widehat{\sigma_{\mathrm{N}}^2} = \widehat{\sigma_{\mathrm{N}}^2}(l-1)$ as done in [21]. As speech spectral coefficients usually exhibit a larger degree of fluctuations between successive segments, estimation of the *a priori* SNR is difficult. In [21], it is proposed to employ a limited maximum-likelihood (ML) estimate for $\widehat{\xi}$ in (7), as briefly recapitulated in Section III-A, followed by a bias compensation, which will be discussed in Section III-B.

After estimating the noise periodogram via (7), the noise power spectral density is then obtained from (7) via recursive smoothing with $\alpha_{\mathrm{pow}} = 0.8$ [21]

$$\widehat{\sigma_{\mathrm{N}}^2}(l) = \alpha_{\mathrm{pow}}\widehat{\sigma_{\mathrm{N}}^2}(l-1) + (1-\alpha_{\mathrm{pow}})\mathrm{E}\left(|N|^2|y(l)\right). \tag{8}$$

Next, we show that the MMSE estimator (7) is biased when the estimated quantities $\widehat{\sigma_{\mathrm{N}}^2}$ and $\widehat{\sigma_{\mathrm{S}}^2}$ differ from the true quantities σ_{N}^2 and σ_{S}^2, respectively. Taking the expected value of (7) with respect to Y and stating the condition on the estimated quantities explicitly, we obtain

$$\mathrm{E}_Y\left(\mathrm{E}\left(|N|^2|Y,\widehat{\sigma_{\mathrm{N}}^2},\widehat{\sigma_{\mathrm{S}}^2}\right)\right) = \left(\frac{\widehat{\sigma_{\mathrm{N}}^2}}{\widehat{\sigma_{\mathrm{S}}^2}+\widehat{\sigma_{\mathrm{N}}^2}}\right)^2\left(\sigma_{\mathrm{S}}^2+\sigma_{\mathrm{N}}^2\right) + \frac{\widehat{\sigma_{\mathrm{S}}^2}}{\widehat{\sigma_{\mathrm{S}}^2}+\widehat{\sigma_{\mathrm{N}}^2}}\widehat{\sigma_{\mathrm{N}}^2} \tag{9}$$

where for this derivation we assume that $\widehat{\sigma_{\mathrm{S}}^2}$ and $\widehat{\sigma_{\mathrm{N}}^2}$ are not functions of Y, and we employ partial integration or use [24, Eq. 3.381.4]. When $\widehat{\sigma_{\mathrm{S}}^2} = \sigma_{\mathrm{S}}^2$ and $\widehat{\sigma_{\mathrm{N}}^2} = \sigma_{\mathrm{N}}^2$, we obtain from (9) that $\mathrm{E}_Y(\mathrm{E}(|N|^2|Y,\sigma_{\mathrm{N}}^2,\sigma_{\mathrm{S}}^2)) = \sigma_{\mathrm{N}}^2$, which means that the estimator in (7) is unbiased. However, as argued in [20], [21], the estimator (7) is biased when estimated quantities are used and $\widehat{\sigma_{\mathrm{S}}^2} \neq \sigma_{\mathrm{S}}^2$ and/or $\widehat{\sigma_{\mathrm{N}}^2} \neq \sigma_{\mathrm{N}}^2$, as then $\mathrm{E}_Y(\mathrm{E}(|N|^2|Y,\widehat{\sigma_{\mathrm{N}}^2},\widehat{\sigma_{\mathrm{S}}^2})) \neq \sigma_{\mathrm{N}}^2$.

A. Interpretation as a Voice Activity Detector

In this section, we show that the MMSE estimator can be interpreted as a VAD-based noise tracker when the *a priori* SNR is estimated by means of a limited ML estimate, as proposed in [21].

From (7), we see that MMSE solution results in a weighted sum of the noisy observation and the previous estimate of the spectral noise power $\widehat{\sigma_{\mathrm{N}}^2}$. The two weights are functions of the *a priori* SNR $\widehat{\xi}$ and gradually take values between zero and one, resulting in a *soft* decision between $|y|^2$ and $\widehat{\sigma_{\mathrm{N}}^2}$. However, in [21] it was proposed to use a limited maximum-likelihood (ML) estimate of the *a priori* SNR, which is obtained as

$$\widehat{\xi}(l) = \max\left(0, \widehat{\xi}^{\mathrm{ml}}(l)\right) = \max\left(0, \widehat{\gamma}(l)-1\right) \tag{10}$$

Bias & safety-net 都不需要，且 (11) VAD 可以被替换为 SPP

where $\widehat{\gamma}(l) = (|y(l)|^2)/(\widehat{\sigma_{\mathrm{N}}^2}(l-1))$. One reason to use this estimator for the *a priori* SNR is that it allows for the computation of an analytic expression for the bias as expressed by (9). Substituting (10) into (7) we see that this MMSE estimator can be seen as a VAD-based detector, i.e.,

$$E\left(|N(l)|^2\,|y(l)\right) = \begin{cases} \widehat{\sigma_{\mathrm{N}}^2}(l-1), & |y(l)|^2 \geq \widehat{\sigma_{\mathrm{N}}^2}(l-1) \\ |y(l)|^2, & |y(l)|^2 < \widehat{\sigma_{\mathrm{N}}^2}(l-1). \end{cases} \tag{11}$$

Notice that using the *a priori* SNR estimator from (10) we thus obtain a *hard* instead of a soft decision between the noisy observation and the estimate of the spectral noise power.

B. Bias Compensation

As argued in [20] and [21] the estimator (11) is biased when estimated quantities are used. Similar to [21] we derive this bias given that ξ is estimated using (10), but distinguish between the estimated $\widehat{\sigma_{\mathrm{N}}^2}$ and the true σ_{N}^2. Then, using [24, Eq. 3.381.1] we find for the bias

$$\begin{aligned} B^{-1} &= \mathrm{E}_Y\Big(\mathrm{E}\Big(|N|^2|Y,\widehat{\xi},\widehat{\sigma_{\mathrm{N}}^2}\Big)\Big)/\sigma_{\mathrm{N}}^2 \\ &= \frac{\sigma_{\mathrm{S}}^2+\sigma_{\mathrm{N}}^2}{\sigma_{\mathrm{N}}^2}\phi\left(2,\frac{\widehat{\sigma_{\mathrm{N}}^2}}{\sigma_{\mathrm{S}}^2+\sigma_{\mathrm{N}}^2}\right)+\exp\left(-\frac{\widehat{\sigma_{\mathrm{N}}^2}}{\sigma_{\mathrm{S}}^2+\sigma_{\mathrm{N}}^2}\right)\frac{\widehat{\sigma_{\mathrm{N}}^2}}{\sigma_{\mathrm{N}}^2} \end{aligned} \tag{12}$$

where $\phi(2,x) = \int_0^x \mathrm{e}^{-t}t\mathrm{d}t$ is the *incomplete gamma function* [24, Eq. 8.350.1]. Under the assumption that the spectral noise power is known, i.e., $\widehat{\sigma_{\mathrm{N}}^2} = \sigma_{\mathrm{N}}^2$, it is shown in [21] that the expectation of (11) is smaller than the true noise variance, i.e., an underestimation of σ_{N}^2 with $B > 1$, when σ_{S}^2 is small with respect to σ_{N}^2. The final estimate of the spectral noise power is then obtained in [21] as

$$\widehat{\sigma_{\mathrm{N}}^2} = \mathrm{E}\left(|N|^2|y\right)B.$$

Due to the nonstationarity of the spectral noise power across time, besides an underestimation, overestimations can occur as well, while the bias compensation in [21] can only account for an underestimation of the noise periodogram. Note that in principle (12) can also account for noise overestimation, as $B < 1$ for $\sigma_{\mathrm{N}}^2 < \widehat{\sigma_{\mathrm{N}}^2}$. However, strictly speaking, the parameters we need in order to estimate the bias in (12) are the same as we needed in the first place to compute (7).

C. Safety-Net

To overcome that the spectral noise power tracker stagnates when the noise level would make an abrupt step from one segment to the next, in [21] a so-called *safety-net* is employed. In this safety-net, the last 0.8 seconds of the noisy speech periodogram, i.e., $|y(l)|^2$, are stored. The final estimate of the spectral noise power is obtained by comparing the current noise power estimate to the minimum of the last 0.8 seconds of $|y(l)|^2$, as

$$\widehat{\sigma_{\mathrm{N}}^2}(l) \leftarrow \max\left(\widehat{\sigma_{\mathrm{N}}^2}(l), \min\left(|y(l-M+1)|^2,\ldots,|y(l)|^2\right)\right) \tag{13}$$

with M the number of time-frames in the period of 0.8 seconds.

Instead of first using a limited maximum-likelihood (ML) estimate for the *a priori* SNR that results in the VAD of (11), in this paper we argue that neither the bias compensation of Section III-B nor the safety-net of Section III-C is necessary if the hard decision of the VAD (11) is replaced by the soft decision of an SPP estimator.

IV. Unbiased Estimator Based on an SPP Estimate With Fixed Priors

In (11) of Section III-A, we have shown that the MMSE estimator (7) can be interpreted as a VAD-based spectral noise power estimator when the limited ML estimate $\max(0,\gamma(l)-1)$ is used to estimate the *a priori* SNR. In that case, the estimated noise term is only updated when $|y(l)|^2 < \widehat{\sigma_{\mathrm{N}}^2}(l-1)$. This is the reason that a bias compensation of (7) by (12) is necessary.

In this section, we propose to replace the hard decision of the VAD by a soft decision SPP with fixed priors, making a bias compensation unnecessary. Under speech presence uncertainty an MMSE estimator for the noise periodogram is given by

$$\mathrm{E}\left(|N|^2|y\right) = P(\mathcal{H}_0|y)\mathrm{E}\left(|N|^2|y,\mathcal{H}_0\right) + P(\mathcal{H}_1|y)\mathrm{E}\left(|N|^2|y,\mathcal{H}_1\right) \tag{14}$$

where $\mathcal{H}_0$ indicates speech absence, while $\mathcal{H}_1$ indicates speech presence.

A. Estimation of the Speech Presence Probability

SPP

Similar as for the derivation of (7), we assume that the speech and noise complex coefficients are Gaussian distributed. Using Bayes' theorem, for the *a posteriori* SPP we have

$$P(\mathcal{H}_1|y) = \frac{P(\mathcal{H}_1)p_{Y|\mathcal{H}_1}(y)}{P(\mathcal{H}_0)p_{Y|\mathcal{H}_0}(y) + P(\mathcal{H}_1)p_{Y|\mathcal{H}_1}(y)}. \tag{15}$$

Thus, to compute the *a posteriori* SPP we need models for the *a priori* probabilities $P(\mathcal{H}_1) = 1 - P(\mathcal{H}_0)$, as well as the likelihood functions for speech presence $p_{Y|\mathcal{H}_1}(y)$ and speech absence $p_{Y|\mathcal{H}_0}(y)$. Without an observation, we assume that it is equally likely that the time–frequency point under consideration contains speech or not. Hence, we choose uniform priors, i.e., $P(\mathcal{H}_1) = P(\mathcal{H}_0) = 0.5$, which can be considered a worst case assumption [1]. In contrast to [16], these fixed priors are independent of the observation.

The likelihood functions $p_{Y|\mathcal{H}_1}(y)$ and $p_{Y|\mathcal{H}_0}(y)$ in (15) indicate how well the observation y fits the modeling parameters for speech presence and speech absence, respectively. As in Section II, we assume the observation to be complex Gaussian distributed. We thus model the likelihood under speech absence by

$$p_{Y|\mathcal{H}_0}(y) = \frac{1}{\widehat{\sigma_{\mathrm{N}}^2}\pi}\exp\left(-\frac{|y|^2}{\widehat{\sigma_{\mathrm{N}}^2}}\right) \tag{16}$$

while we model the likelihood under speech presence by

$$p_{Y|\mathcal{H}_1}(y) = \frac{1}{\widehat{\sigma_{\mathrm{N}}^2}\,(1+\xi_{\mathcal{H}_1})\,\pi}\exp\left(-\frac{|y|^2}{\widehat{\sigma_{\mathrm{N}}^2}\,(1+\xi_{\mathcal{H}_1})}\right). \tag{17}$$

Notice that for the further derivation in this section, we make a distinction between the distribution p_Y in (3) and the distribution $p_{Y|\mathcal{H}_1}$ in (17). While in (3) ξ is the true local SNR, in (17) the *a priori* SNR $\xi_{\mathcal{H}_1}$ is a parameter of our model for speech presence. As such, it reflects the SNR that is *typical* if

speech *were* present [25], [26] and can also be interpreted as a long-term SNR rather than the short-term local SNR. In the radar or communication context, one would choose $\xi_{\mathcal{H}_1}$ in order to guarantee a specified performance in terms of false alarms or missed detections [1]. Similarly, in Section IV-D we will find the fixed optimal *a priori* SNR $10\log_{10}(\xi_{\mathcal{H}_1}) = 15$ dB by minimizing the total probability of error when the true *a priori* SNR ξ is uniformly distributed between $\xi_{\text{low}} = 0$ and $\xi_{\text{up}} = 100$, which corresponds to $10\log_{10}(\xi_{\text{low}}) = -\infty$ dB and $10\log_{10}\xi_{\text{up}} = 20$ dB. Choosing a fixed *a priori* SNR has the benefit of decoupling the noise power estimator from subsequent steps in a speech enhancement framework, such as the estimation of the speech power and the estimation of the clean speech. Further, note that the likelihoods of speech presence and absence (17), (16) differ only in the *a priori* SNR $\xi_{\mathcal{H}_1}$. Choosing an optimal fixed *a priori* SNR, guarantees that the two models for speech presence and speech absence differ, and thus enables *a posteriori* SPP estimates close to zero in speech absence. This is in contrast to [16], [18] where the *a priori* SNR is adaptively estimated. The adaptation in [16], [18] yields *a priori* SNR estimates close to zero in speech absence such that the likelihood functions (17), (16) are virtually the same and the *a posteriori* SPP yields only the prior, as $P(\mathcal{H}_1|y, \xi_{\mathcal{H}_1} = 0) = P(\mathcal{H}_1)$, independent of the observation y. Thus, without further modifications, speech absence can not be detected when the adapted $\xi_{\mathcal{H}_1}$ is zero. To overcome this undesired behavior, in [16] also the prior $P(\mathcal{H}_1)$ is adapted with respect to the observation. However, strictly speaking, this contradicts the definition of the *a priori* SPP.

Substituting (16) and (17) into (15) we obtain for the *a posteriori* SPP (e.g., [27])

$$P(\mathcal{H}_1|y) = \left(1 + \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}\,(1+\xi_{\mathcal{H}_1})\,\mathrm{e}^{-\frac{|y|^2}{\widehat{\sigma_{\mathrm{N}}^2}}\frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}}\right)^{-1} \tag{18}$$

where in this work, we assume $P(\mathcal{H}_1) = P(\mathcal{H}_0)$. As in (7) we employ the spectral noise power estimate of the previous frame, i.e., $\widehat{\sigma_N^2} = \widehat{\sigma_N^2}(l-1)$ in (16)–(18).

B. Derivation of $E(|N|^2|y, \mathcal{H}_0)$ and $E(|N|^2|y, \mathcal{H}_1)$

We can solve (18) for the *a posteriori* SNR $\widehat{\gamma} = |y|^2/\widehat{\sigma_N^2}$, and obtain a function of $\xi_{\mathcal{H}_1}$ and $P(\mathcal{H}_1|y)$, as

$$\widehat{\gamma} = \log\left(\frac{1+\xi_{\mathcal{H}_1}}{P(\mathcal{H}_1|y)^{-1}-1}\right)\frac{1+\xi_{\mathcal{H}_1}}{\xi_{\mathcal{H}_1}}. \tag{19}$$

Using the optimal *a priori* SNR $10\log_{10}(\xi_{\mathcal{H}_1}) = 15$ dB in (19), we see that already for speech presence probabilities $P(\mathcal{H}_1|y) > 0.075$ the *a posteriori* SNR satisfies $\widehat{\gamma} > 1$. Thus, when speech is present and $P(\mathcal{H}_1|y)$ is sufficiently large, we can rewrite the ML estimate of the *a priori* SNR from (10) as $\widehat{\xi}^{\text{ml}} = \widehat{\gamma} - 1$. The optimal estimator under speech presence can now be computed as

$$\mathrm{E}\left(|N|^2|y, \widehat{\xi}, \mathcal{H}_1\right) = \mathrm{E}\left(|N|^2|y, \widehat{\xi}^{\text{ml}} = \widehat{\gamma} - 1\right) = \widehat{\sigma_{\mathrm{N}}^2} \tag{20}$$

which, similar to (11), follows from substitution of $\widehat{\xi}^{\text{ml}} = \widehat{\gamma} - 1$ into (7). Under speech absence we have $Y = N$ and thus

$$\mathrm{E}\left(|N|^2|y, \mathcal{H}_0\right) = \mathrm{E}\left(|N|^2|n\right) = |n|^2 = |y|^2. \tag{21}$$

Thus, with (20) and (21), the MMSE estimator under speech presence uncertainty (14) turns into a soft weighting between the noisy observation $|y|^2$ and the previous noise power estimate $\widehat{\sigma_N^2}$ similar to (7):

$$\mathrm{E}\left(|N|^2|y\right) = P(\mathcal{H}_0|y)|y|^2 + P(\mathcal{H}_1|y)\widehat{\sigma_{\mathrm{N}}^2}. \tag{22}$$

Here $P(\mathcal{H}_0|y) = 1 - P(\mathcal{H}_1|y)$ and the spectral noise power estimate from the previous frame is employed, i.e., $\widehat{\sigma_N^2} = \widehat{\sigma_N^2}(l-1)$. The spectral noise power is then obtained by a recursive smoothing of $\mathrm{E}(|N|^2|y)$ as given in (8).

C. Avoiding Stagnation

If the noise power estimate $\widehat{\sigma_N^2}$ underestimates the true noise power σ_{N}^2, the *a posteriori* SPP in (18) will be overestimated. From (22) it follows that then the noise power will not be tracked as quickly as desired. In the extreme case, when $\widehat{\sigma_N^2}$ heavily underestimates the true noise power σ_{N}^2, the *a posteriori* SPP tends to one, $P(\mathcal{H}_1|y) = 1$. Then, the noise power will not be updated anymore, even though $|y|^2$ may be small with respect to the true, but unknown, noise power σ_{N}^2.

To avoid a stagnation of the noise power update due to an underestimated noise power, we check if the *a posteriori* SPP has been close to one for a long time. For this we propose the following memory and computationally efficient algorithm. First, we recursively smooth $P(\mathcal{H}_1|y)$ over time, as

$$\overline{\mathcal{P}}(l) = 0.9\,\overline{\mathcal{P}}(l-1) + 0.1\,P\left(\mathcal{H}_1|y(l)\right). \tag{23}$$

Then, if this smoothed quantity is larger than 0.99, we conclude that the update may have stagnated, and force the current *a posteriori* SPP estimate $P(\mathcal{H}_1|y(l))$ to be lower than 0.99, as

$$P\left(\mathcal{H}_1|y(l)\right) \leftarrow \begin{cases} \min\left(0.99, P\left(\mathcal{H}_1|y(l)\right)\right), & \overline{\mathcal{P}}(l) > 0.99 \\ P\left(\mathcal{H}_1|y(l)\right), & \text{else} \end{cases} \tag{24}$$

This procedure fits well into the framework and is more memory efficient than the safety-net of Section III-C as we do not need to store 0.8 seconds of data.

代替 safety-net

D. Finding the optimal $\xi_{\mathcal{H}_1}$

= 15dB 常数

We find a fixed optimal $\xi_{\mathcal{H}_1}$ by minimizing the total probability of error, given by $P_\epsilon = P(\mathcal{H}_1)P_{\mathrm{m}} + P(\mathcal{H}_0)P_{\mathrm{f}}$ [28, Ch. 2], averaged over *a priori* SNR values that are of interest for the considered application. Here P_{m} and P_{f} denote the missed-hit and the false-alarm rates, respectively. We define a missed hit as the probability that $P(\mathcal{H}_1|y)$ yields values lower than 0.5 even though speech is present, and a false-alarm as the probability that $P(\mathcal{H}_1|y)$ yields values larger than 0.5 even though speech is absent. Assuming we know the true spectral noise power, i.e.,

$\widehat{\sigma_N^2} = \sigma_N^2$, the point where $P(\mathcal{H}_1|y) = 0.5$ is referred to as $|y_{\text{int}}|$ and results from (18)

$$|y_{\text{int}}| = \sqrt{\sigma_N^2 \frac{1+\xi_{\mathcal{H}_1}}{\xi_{\mathcal{H}_1}} \log\left((1+\xi_{\mathcal{H}_1})\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}\right)}. \quad (25)$$

We now want to quantify the false-alarm and missed-hit rates, when the parameter for our speech presence model is given by $\xi_{\mathcal{H}_1}$. For this, we compute the errors for input data Y with several input SNR, when the speech absence and presence models are given by (16) and (17) resulting in (18) and (25). Note that the probability density function of the observed input data is given by (3) and is a function of the true SNR ξ, while the model for speech presence (16) is a function of $\xi_{\mathcal{H}_1}$ which represents the SNR that is typical if speech *were* present. In speech absence the local *a priori* SNR is zero and we have $p_{Y|\xi=0}(y) = p_N(n)$ given in (1).

False alarms occur when the input signal is noise only and the magnitude of the noisy spectral coefficients is larger than $|y_{\text{int}}|$. Using [24, Eq. 3.381.9] the probability of the false-alarm rate can be written as

$$\begin{aligned} P_{\text{f}}(\xi_{\mathcal{H}_1}) &= \int_{|y_{\text{int}}|}^{\infty}\int_{0}^{2\pi} p_N(re^{j\theta})r\,dr\,d\theta \\ &= \exp\left(-\frac{|y_{\text{int}}|^2}{\sigma_N^2}\right) \\ &= \left(\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}[1+\xi_{\mathcal{H}_1}]\right)^{-\frac{1+\xi_{\mathcal{H}_1}}{\xi_{\mathcal{H}_1}}}. \end{aligned} \quad (26)$$

Missed hits occur, when the input data is a speech-plus-noise signal and the magnitude of the noisy spectral coefficients is lower than $|y_{\text{int}}|$. The probability of the missed-hit rate is given by

$$\begin{aligned} P_{\text{m}}(\xi_{\mathcal{H}_1}, \xi) &= \int_{0}^{|y_{\text{int}}|}\int_{0}^{2\pi} p_Y(re^{j\theta})r\,dr\,d\theta \\ &= 1 - \exp\left(-\frac{|y_{\text{int}}|^2}{\sigma_N^2(1+\xi)}\right) \\ &= 1 - P_{\text{f}}(\xi_{\mathcal{H}_1})^{\frac{1}{1+\xi}}. \end{aligned} \quad (27)$$

We now find the optimal parameter $\xi_{\mathcal{H}_1}$ for the speech presence model when the true input SNR ξ is uniformly distributed between ξ_{low} and ξ_{up}. For this, we minimize the total probability of error for all ξ between ξ_{low} and ξ_{up}, as

$$\xi_{\mathcal{H}_1} = \arg\min_{\widetilde{\xi}_{\mathcal{H}_1}} \int_{\xi_{\text{low}}}^{\xi_{\text{up}}} \left(P(\mathcal{H}_0)P_{\text{f}}\left(\widetilde{\xi}_{\mathcal{H}_1}\right) + P(\mathcal{H}_1)P_{\text{m}}\left(\widetilde{\xi}_{\mathcal{H}_1}, \xi\right)\right) d\xi \quad (28)$$

where we denote the candidates for $\xi_{\mathcal{H}_1}$ as $\widetilde{\xi}_{\mathcal{H}_1}$, while ξ is the true, unknown SNR of the observed signal. Substituting $z = 1/(\xi+1)$, and setting $\xi_{\text{low}} = 0$, we obtain with [24, Eq. 2.325.2]

$$\begin{aligned} \xi_{\mathcal{H}_1} = \arg\min_{\widetilde{\xi}_{\mathcal{H}_1}} \Bigg(& P(\mathcal{H}_0)\xi_{\text{up}}P_{\text{f}}\left(\widetilde{\xi}_{\mathcal{H}_1}\right) + P(\mathcal{H}_1) \\ &\times\left(\xi_{\text{up}} + P_{\text{f}}\left(\widetilde{\xi}_{\mathcal{H}_1}\right) - \log\left(P_{\text{f}}\left(\widetilde{\xi}_{\mathcal{H}_1}\right)\right)\right. \\ &\times \text{Ei}\left(\log\left(P_{\text{f}}\left(\widetilde{\xi}_{\mathcal{H}_1}\right)\right)\right) \\ &- (1+\xi_{\text{up}})P_{\text{f}}\left(\widetilde{\xi}_{\mathcal{H}_1}\right)^{\frac{1}{1+\xi_{\text{up}}}} \\ &+ \log\left(P_{\text{f}}\left(\widetilde{\xi}_{\mathcal{H}_1}\right)\right) \\ &\left.\times\ \text{Ei}\left(\frac{\log\left(P_{\text{f}}\left(\widetilde{\xi}_{\mathcal{H}_1}\right)\right)}{1+\xi_{\text{up}}}\right)\right)\Bigg) \end{aligned} \quad (29)$$

with $\text{Ei}(x) = \int_{-\infty}^{x}(e^t/t)\,dt$ [24, Eq. 8.211.1]].

We choose a range for ξ_{low} to ξ_{up} that is realistic for the application under consideration. As we compute the integral in the linear domain, the influence of ξ_{low} is rather small, as long as $\xi_{\text{low}} \ll 1$, i.e., $10\log_{10}(\xi_{\text{low}}) \ll 0$ dB. We consider $10\log_{10}(\xi_{\text{low}}) = -\infty$ dB for the lower bound, and $10\log_{10}(\xi_{\text{up}}) = 20$ dB as an upper bound for a noise reduction application. Then, choosing the *a priori* SNR to be uniformly distributed between these values for ξ_{low} and ξ_{up}, we find the optimal choice for $\xi_{\mathcal{H}_1}$ to be $10\log_{10}(\xi_{\mathcal{H}_1}) = 15$ dB. In Fig. 1, the argument of (29) is plotted for several candidates $\widetilde{\xi}_{\mathcal{H}_1}$. Please note that $\xi_{\mathcal{H}_1}$ is computed offline, and we use the same $10\log_{10}(\xi_{\mathcal{H}_1}) = 15$ dB for all time–frequency points (k,l) throughout the algorithm.

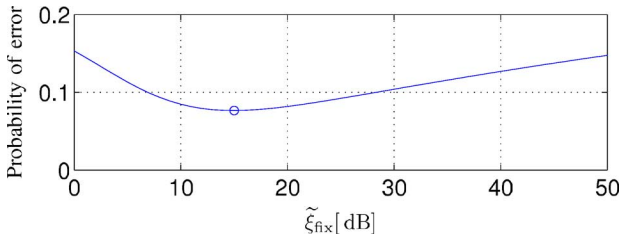


Fig. 1. The total probability of error given in (29) as a function of the candidates $\widetilde{\xi}_{\mathcal{H}_1}$ for $10\log_{10}(\xi_{\text{up}}) = 20$ dB. The minimum corresponds to $10\log_{10}(\xi_{\mathcal{H}_1}) = 15$ dB.

In Section V we show that the proposed approach based on an SPP estimate with fixed priors results in similar results as the estimator of [21], but neither requires a bias correction nor the safety-net of Section III-C.

V. Evaluation

In order to evaluate the proposed spectral noise power estimator, in this section we make a comparison to the MS approach [15] and the MMSE approach with bias compensation (MMSE-BC) proposed in [21]. In Section V-A, we evaluate the estimation accuracy of the competing noise power estimators, while in Section V-B we analyze their computational complexity. Sound examples and the code of the proposed estimator are available at http://www.speech.uni-oldenburg.de/57158.html.

A. Estimation Accuracy

We first compare the logarithmic error between the estimated spectral noise power and the reference spectral noise power. Then we employ the estimated spectral noise power to estimate the clean speech spectral coefficients. For the evaluation

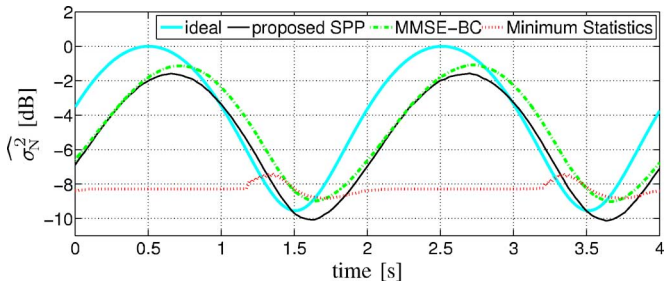


Fig. 2. Comparison of the results of the noise power estimators for modulated white Gaussian noise.

we use 320 sentences of the TIMIT database [29] and several synthetic and natural noise sources. The sampling rate is set at $f_s = 16$ kHz. We use a square-root Hann-window of length $N = 512$ for spectral analysis and synthesis, where successive segments overlap by 50%.

For the evaluation we use several synthetic as well as natural noise sources, that are, stationary white Gaussian noise, modulated white Gaussian noise, traffic noise, nonstationary vacuum cleaner noise, and babble noise. The modulated white Gaussian noise is a synthetic nonstationary noise source that we generate by a point-wise multiplication of the function

$$f(m) = 1 + 0.5\,\sin(2\pi m f_{\mathrm{mod}}/f_s) \tag{30}$$

with a white Gaussian noise sequence. Here m is the time-sample index, and we choose $f_{\mathrm{mod}} = 0.5$ Hz. The traffic noise is recorded next to a rather busy street, where many cars pass by. For the synthetic noise signals, i.e., the stationary white noise and the modulated white noise, the true noise power is known and is thus used for the evaluation. For the remaining nonstationary and thus non-ergodic noise sources the determination of the true spectral noise power is impossible, as only one realization of the random variable is available in each time–frequency point. In these cases we use the periodogram of the noise-only signal as an estimate of the true but unknown spectral noise power, i.e., $\sigma^2_{N,k}(l) \leftarrow |n_k(l)|^2$.

First, in Fig. 2, we compare the results of noise power estimation when the input signal consists only of a modulated white Gaussian noise signal. The true noise power is also given. We averaged the results over 60 seconds of data, i.e., 15 periods of 4 seconds length. It can be seen that all estimators can not fully follow the true noise power. For the considered example, the MS approach [15] has the worst tracking capability. Further, it can be seen that the MMSE-BC approach [21] has the tendency to overestimate the noise power when the noise power is decreasing.

As in [19], we compare the estimated noise power $\widehat{\sigma_N^2}$ to the reference σ_N^2 in terms of the log-error distortion measure. In contrast to what was proposed in [19] we separate the error measure into overestimation and underestimation, i.e.,

$$\mathrm{LogErr} = \mathrm{LogErr}_{\mathrm{ov}} + \mathrm{LogErr}_{\mathrm{un}} \tag{31}$$

where $\mathrm{LogErr}_{\mathrm{ov}}$ measures the contributions of an overestimation of the true noise power, as

$$\mathrm{LogErr}_{\mathrm{ov}} = \frac{10}{NL}\sum_{l=0}^{L-1}\sum_{k=0}^{N-1}\left|\min\left(0, \log_{10}\frac{\sigma^2_{N,k}(l)}{\widehat{\sigma^2_{N,k}(l)}}\right)\right| \tag{32}$$

while $\mathrm{LogErr}_{\mathrm{un}}$ measures the contributions of an underestimation of the true noise power, as

$$\mathrm{LogErr}_{\mathrm{un}} = \frac{10}{NL}\sum_{l=0}^{L-1}\sum_{k=0}^{N-1}\max\left(0, \log_{10}\frac{\sigma^2_{N,k}(l)}{\widehat{\sigma^2_{N,k}(l)}}\right). \tag{33}$$

Note that an overestimation of the true noise power as indicated by $\mathrm{LogErr}_{\mathrm{ov}}$ is likely to result in an attenuation of the speech signal in a speech enhancement framework and thus in speech distortions. On the other hand, an underestimation of the true noise power as indicated by $\mathrm{LogErr}_{\mathrm{un}}$ results in a reduced noise reduction and is likely to yield an increase of musical noise when the estimated noise power is employed in a speech enhancement framework [30].

We also employ the estimated noise power in a standard speech enhancement framework. For this we use the decision-directed estimation with a smoothing factor of 0.98 [2] to obtain an estimate of the *a priori* SNR. The estimated *a priori* SNR is then employed in a Wiener filter which is limited to be larger than -17 dB. We employ the Wiener filter, as this is the MMSE-optimal conditional estimator of the clean speech spectral coefficients given that the speech and noise spectral coefficients are complex Gaussian distributed, which fits the assumptions we have made to derive the noise power estimators. Still, for the sake of completeness, in Fig. 7 we also present the results we obtain when a state-of-the-art super-Gaussian estimator from [6] is used. For this filter, it is assumed that speech spectral magnitudes are generalized Gamma distributed with parameters $\gamma = 1$ and $\nu = 0.6$ in [6].

We measure the performance in terms of the segmental noise reduction and the segmental speech SNR as proposed by [5], as well as the segmental SNR improvement. For this, the resynthesized time-domain signals are segmented into non-overlapping segments of 10-ms length. Speech SNR (spSSNR), noise reduction (NR), and segmental SNR (SSNR) are only evaluated in signal segments that contain speech and are defined as follows:

$$\mathrm{spSSNR} = \frac{10}{|\mathbb{L}|}\sum_{l\in\mathbb{L}}\log_{10}\frac{\sum_{m=1}^{M} s_t^2(lM+m)}{\sum_{m=1}^{M}\left(s_t(lM+m)-\widetilde{s}_t(lM+m)\right)^2} \tag{34}$$

$$\mathrm{NR} = \frac{10}{|\mathbb{L}|}\sum_{l\in\mathbb{L}}\log_{10}\frac{\sum_{m=1}^{M} n_t^2(lM+m)}{\sum_{m=1}^{M}\widetilde{n}_t^2(lM+m)} \tag{35}$$

$$\mathrm{SSNR} = \frac{10}{|\mathbb{L}|}\sum_{l\in\mathbb{L}}\log_{10}\frac{\sum_{m=1}^{M} s_t^2(lM+m)}{\sum_{m=1}^{M}\left(\widehat{s}_t(lM+m)-s_t(lM+m)\right)^2} \tag{36}$$

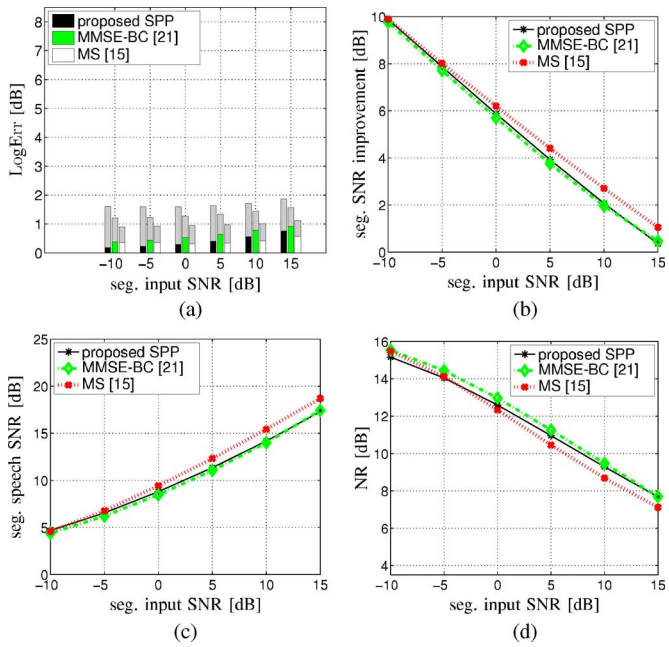


Fig. 3. Quality measures for stationary white Gaussian noise. The lower part of the bars in (a) represents the noise overestimation $\mathrm{LogErr}_{\mathrm{ov}}$, while the upper part represents the noise underestimation $\mathrm{LogErr}_{\mathrm{un}}$. The total height of the bars gives the LogErr. (a) Log estimation error. (b) Segmental SNR improvement. (c) Segmental speech SNR. (d) Segmental noise reduction.

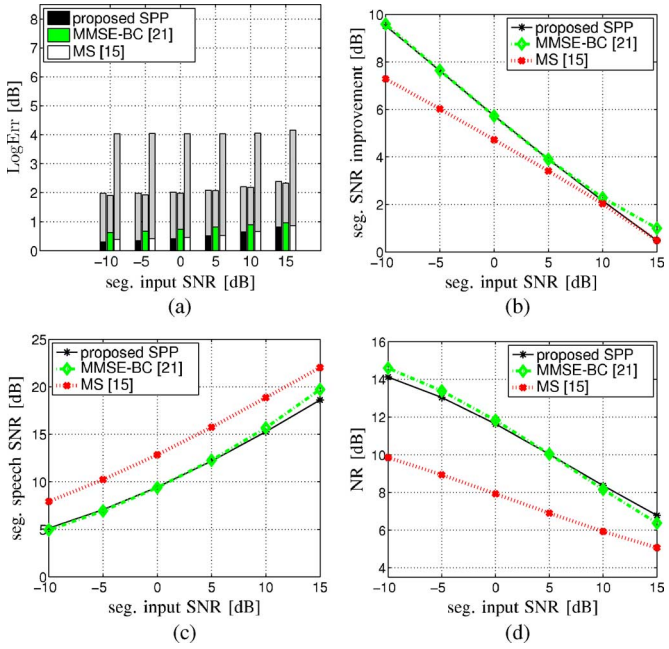


Fig. 4. Quality measures for modulated white Gaussian noise. As in Fig. 3, the bars in (a) indicate noise overestimation and underestimation. (a) Log estimation error. (b) Segmental SNR improvement. (c) Segmental speech SNR. (d) Segmental noise reduction.

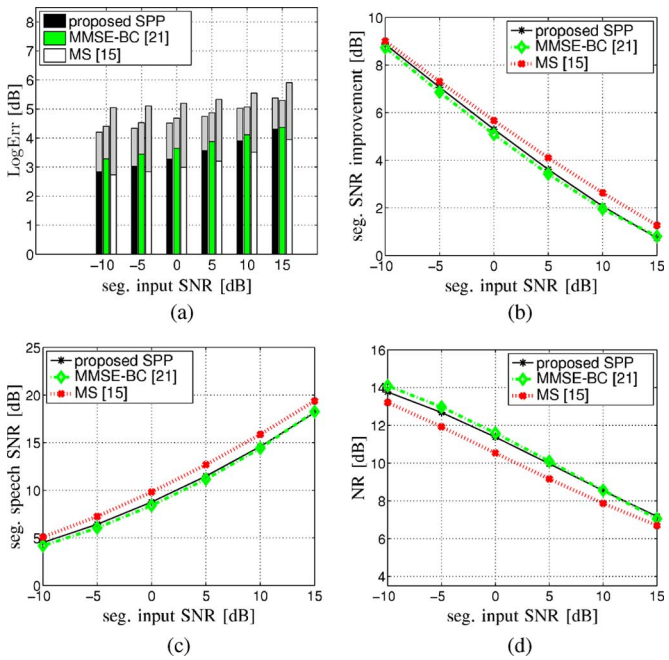


Fig. 5. Quality measures for traffic noise. As in Fig. 3 the bars in (a) indicate noise overestimation and underestimation. (a) Log estimation error. (b) Segmental SNR improvement. (c) Segmental speech SNR. (d) Segmental noise reduction.

where $M = 160$, m is the time-domain sample index, l is the segment index, $\mathbb{L}$ is the set of signal segments that contain speech. In (34)–(36), we assume that the delay introduced by the overlap-add is accounted for. We determine $\mathbb{L}$ by choosing all signal frames whose energy is larger than -45 dB with respect to the maximum frame energy in the considered TIMIT signal. Further, $s_{\mathrm{t}}(m)$ and $n_{\mathrm{t}}(m)$ are the speech and noise time-domain signal, and $\widehat{s}_{\mathrm{t}}(m)$ is the estimated clean-speech time-domain signal after applying the speech enhancement filter. The quantities $\widetilde{s}_{\mathrm{t}}(m)$ and $\widetilde{n}_{\mathrm{t}}(m)$ are obtained by applying the same speech enhancement filter coefficients that are applied to noisy speech also to the speech-only and the noise-only signals. The segmental speech SNR is a measure for speech distortions and becomes larger the lower the speech distortions are. The noise reduction NR indicates the relative noise reduction, while the segmental SNR takes into account both noise reduction and speech distortions. Note that for input SNRs below 0 dB, the segmental SNR can always be improved by nulling all coefficients. Thus, we suggest that it should always be read together with a measure for speech distortions.

The results of these evaluations are given in Figs. 3–6 for the Wiener filter and in Fig. 7 for the super-Gaussian filter. The measure $\mathrm{LogErr}_{\mathrm{ov}}$ is indicated by the lower bars and the measure $\mathrm{LogErr}_{\mathrm{un}}$ by the gray upper bars. The sum of both error measures, LogErr, is given by the total height of the bars. It can be seen that for the stationary white Gaussian noise signal (Fig. 3) the MS approach has the lowest LogErr error and the largest SNR improvement.

However, the results for the modulated white Gaussian noise signal in Fig. 4, clearly show that the MS approach is not able to track the noise spectral power with adequate speed and heavily underestimates the noise spectral power. This results in large values for $\mathrm{LogErr}_{\mathrm{un}}$ [Fig. 4(a)] and in a low noise reduction performance [Fig. 4(d)] that is likely to result in musical noise. At the same time, as the noise reduction is less aggressive, this also results in a larger speech SNR [Fig. 4(c)]. It can be seen however that for the modulated noise, the MS approach results in a poor tradeoff between noise reduction and speech distortion as it results in lower segmental SNR improvements [Fig. 4(b)].

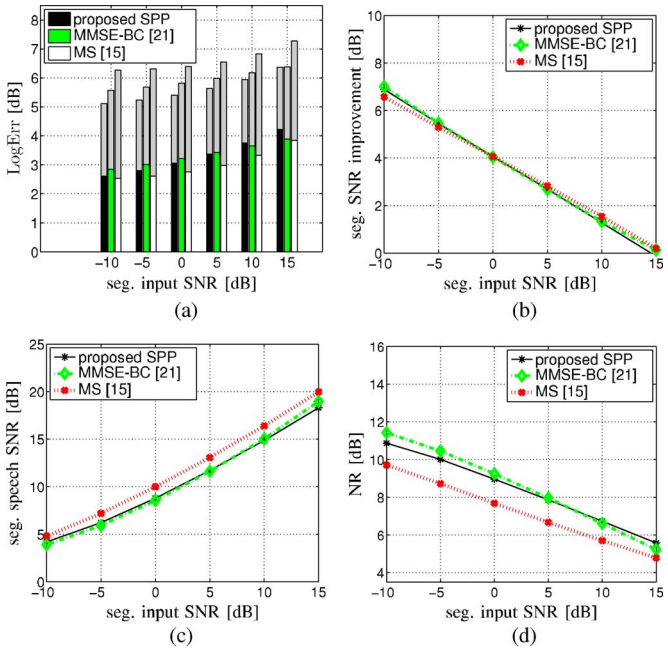


Fig. 6. Quality measures for babble noise. As in Fig. 3 the bars in (a) indicate noise overestimation and underestimation. (a) Log estimation error. (b) Segmental SNR improvement. (c) Segmental speech SNR. (d) Segmental noise reduction.

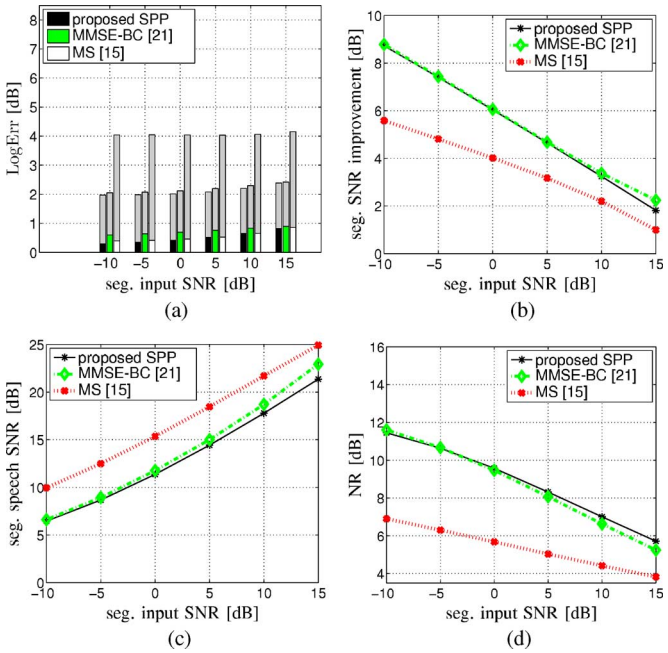


Fig. 7. Quality measures for modulated white Gaussian noise. In contrast to Fig. 4, we use here a super-Gaussian filter function from [6]. (a) Log estimation error. (b) Segmental SNR improvement. (c) Segmental speech SNR. (d) Segmental noise reduction.

Also for the natural nonstationary noise sources in Figs. 5 and 6, the MS approach yields the largest logarithmic estimation error LogErr, the lowest noise reduction and the largest speech SNR. However, for traffic noise [Fig. 5(b)], it performs somewhat better in terms of the segmental SNR than the proposed SPP approach and the MMSE estimator [21].

For almost all considered noise types, the bias-compensated MMSE estimator [21] exhibits an error $\mathrm{LogErr}_{\mathrm{ov}}$ that is generally somewhat larger than for the other reference methods, which is likely to result in more attenuation of speech components in a speech enhancement framework. The only exception is babble noise, where noise overestimation $\mathrm{LogErr}_{\mathrm{ov}}$ is rather similar for the proposed noise power estimators.

TABLE I
NORMALIZED PROCESSING TIME

	MS [15]	MMSE-BC$_1$ [21]	MMSE-BC$_2$ [21]	SPP
Proc.-time	4.8	3.3	4.5	1.0

In Fig. 7, we show the results for modulated white Gaussian noise, when the super-Gaussian filter function from [6] is used. The results of the LogErr do not explicitly depend on the chosen filter, as only the noise power estimator is evaluated. The results for the segmental SNR, segmental speech SNR and the segmental noise reduction are similar to the results in Fig. 4, in the sense that the MS approach yields the lowest noise reduction and the largest speech SNR, but also the lowest segmental SNR improvement. The proposed SPP and the bias-compensated MMSE estimator [21] yield rather similar results.

In general, the proposed low complexity approach based on SPP results in similar LogErr and SNR improvement as the MMSE-BC approach without requiring a bias compensation or the safety-net of Section III-C. The proposed SPP approach has the tendency to result in less noise overestimation, which is likely to result in less attenuation of speech components, but also results in slightly less noise reduction for low input SNRs.

B. Computational Complexity

In order to compare the different algorithms in terms of computational complexity, we computed the processing time of Matlab implementations of the three algorithms that are compared in this section. The processing times for each algorithm, normalized by the processing time of the proposed SPP approach, are given in Table I. Notice that the numbers given in table should be used as an indication. In general, they depend on implementational details and settings, e.g., sampling frequency and length of the fast Fourier transform (FFT). The numbers in Table I reflect all necessary processing steps to compute the spectral noise power, i.e., in order to highlight the complexity of the spectral noise power estimation algorithms, the DFT and inverse DFT necessary to transform a noisy signal frame to the DFT domain and to transform the reconstructed signal back to the time-domain are left out of this comparison on purpose.

The proposed SPP approach exhibits a computational complexity that is lower than the computational complexity of the MMSE-BC estimator [21] as the exponential function in (18) is the only special function that has to be computed online, while for the MMSE-BC approach, in addition to the exponential function, also the incomplete Gamma function in (12) has to be either computed or tabulated. At the same time, the proposed SPP approach is more memory efficient, as we do not need the safety-net of Section III-C. To demonstrate the influence of computing or tabulating the incomplete Gamma function on the computational complexity of the MMSE-BC estimator, Table I shows the relative computation time for the MMSE-BC estimator with tabulated and computed incomplete Gamma function, denoted by MMSE$-$BC$_1$ and MMSE$-$BC$_2$, respectively.

The computational complexity of the MS approach is somewhat higher than the computational complexity of both MMSE − BC_1 and MMSE − BC_2. Notice that the used implementation of the MS algorithm, as well as the implementation of the MMSE-BC algorithm are both somewhat more efficient than the implementations used in the comparison in [21]. This explains the relatively smaller difference in estimated computational complexity between MMSE-BC and MS in Table I compared to the table in [21].

VI. Conclusion

An important aspect of speech enhancement algorithms is the estimation of the spectral noise power. Recently, it was proposed to estimate this quantity by means of a minimum mean-square error (MMSE)-based estimator [21]. This method is of low computational complexity, while comparisons have shown [22] that spectral noise power estimation performance is improved compared to competing methods.

In this paper, we analyzed the MMSE-based estimator presented in [21] and further improved this method by presenting a modified version.

From the presented analysis of the original MMSE-based estimator we showed that this algorithm can be interpreted as a voice activity detector (VAD)-based noise power estimator, where the noise power is updated only when speech absence is signaled. This is due to the way in which the *a priori* signal-to-noise ratio (SNR) is computed. As a consequence, the method requires a bias compensation as was also originally proposed.

In the presented approach, we proposed to modify the MMSE-based estimator such that use is made of a soft speech presence probability (SPP) with fixed priors. As a result, the estimator becomes automatically unbiased and is of an even lower complexity than the reference MMSE-based approach.

From experimental results it followed that the presented soft SPP-based approach generally achieves similar performance as the original MMSE-based approach with the advantage that no bias compensation is necessary and the computational complexity is even lower.

References

[1] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[4] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[5] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Jan. 2005.

[6] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[7] I. Andrianakis and P. R. White, "Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors," *ELSEVIER Speech Commun.*, vol. 51, no. 1, pp. 1–14, Jan. 2009.

[8] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[9] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, May 1996, vol. 2, pp. 629–633.

[10] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.

[11] I. Cohen, "Speech enhancement using super-Gaussian speech models and noncausal *a priori* SNR estimation," *ELSEVIER Speech Commun.*, vol. 47, no. 3, pp. 336–350, 2005.

[12] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel *a priori* SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, Apr. 2008, pp. 4897–4900.

[13] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[14] K. W. Jang, D. K. Kim, and J.-H. Chang, "A uniformly most powerful test for statistical model-based voice activity detection," in *Proc. ISCA Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 2917–2920.

[15] 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.

[16] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[17] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *ELSEVIER Speech Commun.*, vol. 48, no. 2, pp. 220–231, 2006.

[18] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, May 1998, vol. 1, pp. 365–368.

[19] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.

[20] R. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 4421–4424.

[21] 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.*, Dallas, TX, Mar. 2010, pp. 4266–4269.

[22] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, May 2011, pp. 4640–4643.

[23] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust.*, New Paltz, NY, Oct. 2011, pp. 145–148.

[24] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals Series and Products*, 6th ed. San Diego, CA: Academic, 2000.

[25] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved *a posteriori* speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

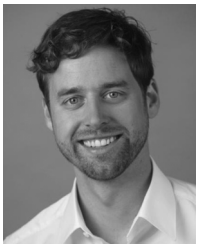
[26] T. Gerkmann, M. Krawczyk, and R. Martin, "Speech presence probability estimation based on temporal cepstrum smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, Mar. 2010, pp. 4254–4257.

[27] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *ELSEVIER Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

[28] H. L. Van Trees, *Detection, Estimation, and Modulation Theory. Part I.* New York: Wiley, 1968.

[29] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," National Inst. Standards Technol. (NIST), 1988.

[30] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Washington, DC, Apr. 1979, pp. 208–211.



Timo Gerkmann (M'10) studied electrical engineering at the universities of Bremen and Bochum, Germany, and received the Dipl.-Ing. degree and the Dr.-Ing. degree from the Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum, Bochum, Germany, in 2004 and 2010, respectively.

From January 2005 to July 2005, he was with Siemens Corporate Research, Princeton, NJ. In 2011, he was a Postdoctoral Researcher at the Sound and Image Processing Lab at the Royal Institute of Technology (KTH), Stockholm, Sweden. Since December 2011, he has headed the Speech Signal Processing Group at the Universität Oldenburg, Oldenburg, Germany. His main research interests are on speech enhancement algorithms, modeling of speech signals, and hearing aid processing.

Richard C. Hendriks received the B.Sc., M.Sc. (*cum laude*), and Ph.D. (*cum laude*) degrees in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2001, 2003, and 2008, respectively.

From 2003 to 2007, he was a Ph.D. Researcher at Delft University of Technology. From 2007 to 2010, he was a Postdoctoral Researcher at Delft University of Technology. Since 2010, he has been an Assistant Professor in the Multimedia Signal Processing Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology. In the autumn of 2005, he was a Visiting Researcher at the Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany. From March 2008 to 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.