# NONLINEAR SPEECH ENHANCEMENT UNDER SPEECH PSD UNCERTAINTY

*Martin Krawczyk-Becker, Timo Gerkmann*

Signal Processing Group, Department of Informatics, Universität Hamburg, Germany
{martin.krawczyk-becker, timo.gerkmann}@uni-hamburg.de

## ABSTRACT

Most Bayesian clean speech estimators, like the Wiener filter or Ephraim and Malah's amplitude estimators, are derived under the assumption that the true power spectral density (PSD) of speech is known. In practice, however, only estimates are available. When the PSD estimation errors are neglected, they propagate through to the final speech estimate, resulting in undesired artifacts such as musical noise and speech distortions. To increase the robustness to PSD estimation errors, recently a linear estimator has been proposed that explicitly takes into account the uncertainty of the available speech PSD estimate. In this paper, we show that in the derivation of this estimator a limiting statistical assumption is made, and that avoiding this assumption leads to a novel, potentially more powerful nonlinear estimator under PSD uncertainty. In combination with a sophisticated speech PSD estimator, the proposed approach achieves a higher predicted speech quality than the linear alternative and its conventional counterpart, the Wiener filter.

***Index Terms—*** speech enhancement, uncertainty, power spectral density

## 1. INTRODUCTION

To reduce the detrimental effect of acoustic noise on the performance of, e.g., hearing aids and mobile phones, sophisticated speech enhancement algorithms are employed. In this paper, the focus is specifically on single-channel Bayesian clean speech estimators that work in the short-time discrete Fourier transform (STFT) domain. Such algorithms are either used directly on a single microphone signal or as a post processing step after a multi-microphone preprocessing stage. Well known examples are the Wiener filter and Ephraim and Malah's amplitude estimators [1, 2]. Over the years, numerous advanced estimators have been derived, e.g. by assuming different distributions for the spectral speech coefficients [3, 4] and/or different optimization criteria [5, 6, 7]. See e.g. [8] for an overview. The vast majority of these estimators are derived under the assumption that the PSD of speech is deterministic and known. However, the true PSD is typically not known and can only be estimated, e.g. via maximum likelihood (ML) estimation, the decision-directed approach [1], or temporal cepstrum smoothing (TCS) [9]. Even in the noise free case, determining the true speech PSDs is in principle not possible as speech is a highly non-stationary and thus non-ergodic process [10]. This has for instance been considered in [11] to derive a speech PSD estimator based on a generalized autoregressive conditional heteroscedasticity (GARCH) model. In clean speech estimators, however, PSD estimates are commonly interpreted as the true PSD. This effectively neglects the uncertainty of the PSD estimates, resulting in suboptimal noise suppression, speech distortions, and musical noise. Please note that in general, the same considerations also apply to the noise PSD. However, for conciseness, here we concentrate on the uncertainty about the speech PSD.

Recently, in [12] a clean speech estimator has been proposed that explicitly models the uncertainty about the speech PSD. On the one hand, the model establishes a theoretically motivated relation between the true PSD and its ML estimate, which also holds for smoothed ML estimates. On the other hand, the model also provides a convenient and theoretically rigorous way to incorporate information about the true clean speech PSD, which for instance can be obtained off-line from a representative clean speech database.

In this paper, a novel minimum mean square error (MMSE) optimal clean speech estimator under PSD uncertainty is derived. We model the true speech PSD as an unobservable random variable of which only an imperfect estimate is available, using the same PSD uncertainty model as in [12]. The major difference between the estimator in [12] and the proposed one results from a restrictive assumption made in [12], i.e. that the speech PSD estimate provides all information about the true PSD that is contained in the noisy observation. We argue that this assumption is not true in general and therefore present a way to avoid it. As a result and in contrast to [12], the proposed clean speech estimator is a *nonlinear* function of the noisy input and thus potentially more powerful. Furthermore, in [12] instantaneous PSD estimates are smoothed directly in the spectral domain via a moving average filter to reduce random outliers that cause musical noise. However, this simple smoothing is known to also smear the PSD estimate at speech on- and offsets [1]. More elaborate approaches like TCS [9, 13] have been shown to effectively reduce musical noise without smearing the speech, which improves the overall speech enhancement performance. Therefore, here we apply TCS instead and show how it can be integrated into the statistical model of [12].

After briefly introducing the signal model in Section 2, the proposed estimator is derived in Section 3 and the uncertainty model is outlined in Section 4. We then compare the proposed estimator to the one in [12] and its conventional, uncertainty unaware counterpart by means of their input-output characteristics (IOCs) in Section 5, followed by an instrumental evaluation in Section 6.

## 2. SIGNAL MODEL AND NOTATION

The complex-valued STFT coefficients of the noisy signal at segment $\ell$ and frequency bin $k$ are denoted as

$$Y_{k,\ell} = S_{k,\ell} + V_{k,\ell}, \tag{1}$$

with mutually independent spectral coefficients of speech $S_{k,\ell}$ and additive noise $V_{k,\ell}$. Since each time-frequency point $(\ell, k)$ is processed separately, the indices are dropped for notational convenience in the sequel. Given their true PSDs $\sigma_{\mathrm{S}}^2$ and $\sigma_{\mathrm{V}}^2$, both, $S$ and $V$ are modeled as zero-mean complex-valued Gaussian distributed random variables. To distinguish estimates from their true counterparts we use the hat symbol, e.g. $\widehat{\sigma_{\mathrm{S}}^2}$ is an estimate of $\sigma_{\mathrm{S}}^2$.

## 3. SPEECH ESTIMATION UNDER PSD UNCERTAINTY

Conventional MMSE optimal estimators of the clean speech coefficients $S$ are derived via

$$\widehat{S} = \mathrm{E}\big(S \mid Y, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{V}}^2\big) = \int_S S p\big(S \mid Y, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{V}}^2\big)\,\mathrm{d}S, \tag{2}$$

i.e. the true PSDs of speech and noise are assumed to be known. If only an estimate $\widehat{\sigma_{\mathrm{S}}^2}$ of the true speech PSD $\sigma_{\mathrm{S}}^2$ is available, analogously to (2), the MMSE optimal clean speech estimator is given by

$$\widehat{S} = \mathrm{E}\Big(S \mid Y, \widehat{\sigma_{\mathrm{S}}^2}, \sigma_{\mathrm{V}}^2\Big) = \int_S S\, p\Big(S \mid Y, \widehat{\sigma_{\mathrm{S}}^2}, \sigma_{\mathrm{V}}^2\Big)\,\mathrm{d}S. \tag{3}$$

To keep the notation concise, in the remainder of this paper we do not state the dependency on $\sigma_{\mathrm{V}}^2$ explicitly, but implicitly assume that each probability density function (PDF) is conditioned on $\sigma_{\mathrm{V}}^2$, e.g. we use $p\Big(S \mid Y, \widehat{\sigma_{\mathrm{S}}^2}\Big)$ to denote $p\Big(S \mid Y, \widehat{\sigma_{\mathrm{S}}^2}, \sigma_{\mathrm{V}}^2\Big)$. Using Bayes' rule we can reformulate the speech posterior of (3) as:

$$p\Big(S \mid Y, \widehat{\sigma_{\mathrm{S}}^2}\Big) = \frac{\int_0^\infty p\Big(S, Y, \widehat{\sigma_{\mathrm{S}}^2}, \sigma_{\mathrm{S}}^2\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2}{p\Big(Y, \widehat{\sigma_{\mathrm{S}}^2}\Big)}, \tag{4}$$

where for the numerator the joint distribution $p\Big(S, Y, \widehat{\sigma_{\mathrm{S}}^2}\Big)$ is expressed in terms of the marginal probability of $p\Big(S, Y, \widehat{\sigma_{\mathrm{S}}^2}, \sigma_{\mathrm{S}}^2\Big)$ to facilitate the upcoming derivations.

### 3.1. Existing linear estimator [12]

An interesting MMSE optimal clean speech estimator under speech PSD uncertainty has recently been proposed in [12]. The estimator can be derived by applying Bayes' rule to (4) such that we have

$$\begin{aligned} p\Big(S \mid Y, \widehat{\sigma_{\mathrm{S}}^2}\Big) &= \int_0^\infty p\Big(S \mid Y, \widehat{\sigma_{\mathrm{S}}^2}, \sigma_{\mathrm{S}}^2\Big)\, p\Big(\sigma_{\mathrm{S}}^2 \mid Y, \widehat{\sigma_{\mathrm{S}}^2}\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2 \\ &\approx \int_0^\infty p\big(S \mid Y, \sigma_{\mathrm{S}}^2\big)\, p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2, \end{aligned} \tag{5}$$

where the denominator of (4) is canceled out. The first simplification, $p\big(S \mid Y, \sigma_{\mathrm{S}}^2\big) \approx p\Big(S \mid Y, \widehat{\sigma_{\mathrm{S}}^2}, \sigma_{\mathrm{S}}^2\Big)$, implies that when $Y$ and the true PSDs are known, the estimate $\widehat{\sigma_{\mathrm{S}}^2}$ does not provide any additional information regarding $S$. The second simplification, $p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big) \approx p\Big(\sigma_{\mathrm{S}}^2 \mid Y, \widehat{\sigma_{\mathrm{S}}^2}\Big)$, is more restrictive. In [12] it is argued that the estimate $\widehat{\sigma_{\mathrm{S}}^2}$, which is computed from $Y$, contains all information about the true $\sigma_{\mathrm{S}}^2$ inherent in $Y$. Here we argue that this assumption does not hold in general and show how it can be avoided in the next section. A model for $p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)$ is outlined in Section 4.

To obtain the clean speech estimator in [12], the simplified speech posterior (5) is plugged into (3) and the integral over $S$ is solved, leading to

$$\widehat{S}_{[12]} = Y \int_0^\infty \frac{\sigma_{\mathrm{S}}^2}{\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{V}}^2}\, p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2 = Y\, G_{[12]}. \tag{6}$$

The spectral gain $G_{[12]}$ under PSD uncertainty is a weighted mixture of Wiener filter gains and independent of the noisy input $Y$, making the estimator linear in $Y$. Now we show that without the simplification of [12], even under the same statistical assumptions for speech and noise, the estimator under PSD uncertainty becomes a nonlinear function of the noisy input $Y$.

### 3.2. Proposed nonlinear estimator

To derive the proposed nonlinear estimator under PSD uncertainty, the speech posterior in (4) is again reformulated via Bayes' rule, but in a different manner:

$$\begin{aligned} &p\Big(S \mid Y, \widehat{\sigma_{\mathrm{S}}^2}\Big) = \\ &\frac{\int_0^\infty p\Big(Y \mid S, \sigma_{\mathrm{S}}^2, \widehat{\sigma_{\mathrm{S}}^2}\Big)\, p\Big(S \mid \sigma_{\mathrm{S}}^2, \widehat{\sigma_{\mathrm{S}}^2}\Big)\, p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2}{\int_S \int_0^\infty p\Big(Y \mid S, \sigma_{\mathrm{S}}^2, \widehat{\sigma_{\mathrm{S}}^2}\Big)\, p\Big(S \mid \sigma_{\mathrm{S}}^2, \widehat{\sigma_{\mathrm{S}}^2}\Big)\, p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2\mathrm{d}S} \\ &\approx \frac{\int_0^\infty p(Y \mid S)\, p\big(S \mid \sigma_{\mathrm{S}}^2\big)\, p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2}{\int_0^\infty \int_S p(Y \mid S)\, p\big(S \mid \sigma_{\mathrm{S}}^2\big)\,\mathrm{d}S p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2}. \end{aligned} \tag{7}$$

For mutually independent speech and noise, the likelihood $p(Y \mid S) \approx p\Big(Y \mid S, \sigma_{\mathrm{S}}^2, \widehat{\sigma_{\mathrm{S}}^2}\Big)$ is assumed to be the PDF of the noise $V$ shifted by $S$, see e.g. [1]. Consequently, it neither depends on the true nor the estimated speech PSD. It is further assumed that when the true speech PSD $\sigma_{\mathrm{S}}^2$ is given, its estimate $\widehat{\sigma_{\mathrm{S}}^2}$ does not provide any additional information regarding $S$, which results in $p\big(S \mid \sigma_{\mathrm{S}}^2\big) \approx p\Big(S \mid \sigma_{\mathrm{S}}^2, \widehat{\sigma_{\mathrm{S}}^2}\Big)$. Note that the denominator $p\Big(Y, \widehat{\sigma_{\mathrm{S}}^2}\Big)$ in (4) is expressed as the marginal distribution of $p\Big(Y, \widehat{\sigma_{\mathrm{S}}^2}, S, \sigma_{\mathrm{S}}^2\Big)$ in (7). As a consequence, numerator and denominator are the same, except for the integral over $S$, such that we conveniently need the same models in the numerator as in the denominator.

To implement the proposed estimator, we insert (7) into (3) and change the order of the integrals in the numerator as we already did in the denominator of (7). The inner integrals over $S$ then only depend on the true PSD $\sigma_{\mathrm{S}}^2$ and can be solved, see e.g. [14, IV.C] for details. This leads to the final estimator

$$\widehat{S} = \frac{\int_0^\infty \frac{\sigma_{\mathrm{S}}^2}{\left(\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{V}}^2\right)^2}\, \mathrm{e}^{\nu}\, p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2}{\int_0^\infty \frac{1}{\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{V}}^2}\, \mathrm{e}^{\nu}\, p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)\,\mathrm{d}\sigma_{\mathrm{S}}^2}\, Y, \tag{8}$$

which is the counterpart to the Wiener filter under speech PSD uncertainty. Here we introduce $\nu = \frac{\sigma_{\mathrm{S}}^2}{\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{V}}^2} \frac{|Y|^2}{\sigma_{\mathrm{V}}^2}$ for a concise notation. Thus, in contrast to the estimator in (6), the proposed estimator (8) is a nonlinear function of the noisy input Y. Similar to [12], the remaining integrals over $\sigma_{\mathrm{S}}^2$ are solved numerically.

## 4. A MODEL OF SPEECH PSD UNCERTAINTY $p\Big(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\Big)$

A statistically rigorous model for the uncertainty of the speech PSD has recently been proposed in [12]. The model is based on assumptions about the employed speech PSD estimator and potential a priori

information about the true speech PSD. To derive this model for the PSD uncertainty $p\left(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\right)$, Bayes' rule is applied [12]:

$$p\left(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\right) = \frac{p\left(\widehat{\sigma_{\mathrm{S}}^2} \mid \sigma_{\mathrm{S}}^2\right) p\left(\sigma_{\mathrm{S}}^2\right)}{p\left(\widehat{\sigma_{\mathrm{S}}^2}\right)} \propto p\left(\widehat{\sigma_{\mathrm{S}}^2} \mid \sigma_{\mathrm{S}}^2\right) p\left(\sigma_{\mathrm{S}}^2\right), \quad (9)$$

where we dropped $p\left(\widehat{\sigma_{\mathrm{S}}^2}\right)$ as it cancels out when $p\left(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\right)$ is plugged into (8). In (9), the uncertainty model $p\left(\sigma_{\mathrm{S}}^2 \mid \widehat{\sigma_{\mathrm{S}}^2}\right)$ is split into two parts. First, the *hyper prior* $p\left(\widehat{\sigma_{\mathrm{S}}^2} \mid \sigma_{\mathrm{S}}^2\right)$ that depends on the specific speech PSD estimator that is used to obtain $\widehat{\sigma_{\mathrm{S}}^2}$. Second, the *hyperhyper prior* $p\left(\sigma_{\mathrm{S}}^2\right)$ that allows to insert information about the true speech PSD. With the formulation in (9), both, the true and the estimated speech PSD are modeled as random variables.

Similar to [12], we use a $\chi^2$ distribution to model the hyper prior $p\left(\widehat{\sigma_{\mathrm{S}}^2} \mid \sigma_{\mathrm{S}}^2\right)$. In [12], the final PSD estimate is obtained by averaging instantaneous PSD estimates, i.e. $\widehat{\sigma_{\mathrm{S}}^2} = \max\left(|Y|^2 - \sigma_{\mathrm{V}}^2, 0\right)$ [1], over $Q$ neighboring frames directly in the spectrum to reduce outliers that would cause musical noise. The smoothed estimate is modeled to be $\chi^2$ distributed with shape parameter $Q$ [12]. However, already Ephraim and Malah [1] found that the simple moving average filter, while effectively reducing undesired outliers for sufficiently long filters, also smears sudden PSD changes, e.g. at speech onsets and offsets. Here we propose to use a more elaborate approach, namely TCS [9, 13]. With this quefrency selective smoothing, undesired outliers in the final PSD estimate are strongly reduced while avoiding a smearing of the speech. The resulting PSD estimate is again $\chi^2$ distributed, with shape parameter $Q$ obtained according to [13, Sec. IV]. Please note that thanks to the $\chi^2$ distribution in [12], the hyper prior $p\left(\widehat{\sigma_{\mathrm{S}}^2} \mid \sigma_{\mathrm{S}}^2\right)$ becomes wider towards low SNRs, nicely incorporating the effect that speech PSD estimation becomes increasingly challenging and error-prone at low SNRs.

The hyperhyper prior $p\left(\sigma_{\mathrm{S}}^2\right)$ in (9) allows to bring in a priori information about the true PSDs of the desired speech sound. To find a model for the hyperhyper prior, first a normal distribution with mean $\mu_{\sigma_{\mathrm{S}}^2}$ and standard deviation $\phi_{\sigma_{\mathrm{S}}^2}$ is fitted to a histogram of $10\log|S|^2$. Here, the histogram is computed over a subset of the TIMIT training set using the same STFT setup as in the evaluation. The fitted distribution serves as a model for $p\left(\sigma_{\mathrm{S,dB}}^2\right)$, where $\sigma_{\mathrm{S,dB}}^2$ is the speech PSD in dB, i.e. $\sigma_{\mathrm{S,dB}}^2 = 10\log\sigma_{\mathrm{S}}^2$. Transforming this into the linear domain then yields a log-normal distribution for the hyperhyper prior $p\left(\sigma_{\mathrm{S}}^2\right)$. In contrast to [12], here we exclude speech absence regions by considering only time-frequency points for which $|S|^2$ is at most 60 dB below the maximum $|S|^2$. For this setup, the mean and the standard deviation in the dB domain are $\mu_{\sigma_{\mathrm{S}}^2} \approx -29$ dB and $\phi_{\sigma_{\mathrm{S}}^2} \approx 11$ dB. In this context, the standard deviation $\phi_{\sigma_{\mathrm{S}}^2}$ is considered a measure of uncertainty in the expected value $\mu_{\sigma_{\mathrm{S}}^2}$: the larger $\phi_{\sigma_{\mathrm{S}}^2}$, the more likely it is that the unknown true PSD $\sigma_{\mathrm{S,dB}}^2$ differs substantially from its available expected value $\mu_{\sigma_{\mathrm{S}}^2}$.

## 5. INPUT-OUTPUT CHARACTERISTIC

In Figure 1, we compare the proposed nonlinear estimator to the alternative linear approach [12] in terms of their IOCs [15] and show how the information in the hyperhyper prior $p\left(\sigma_{\mathrm{S}}^2\right)$ from the previous section can benefit the clean speech estimation. The IOC of an estimator presents the magnitude of the speech estimate $\widehat{S}$ as a function

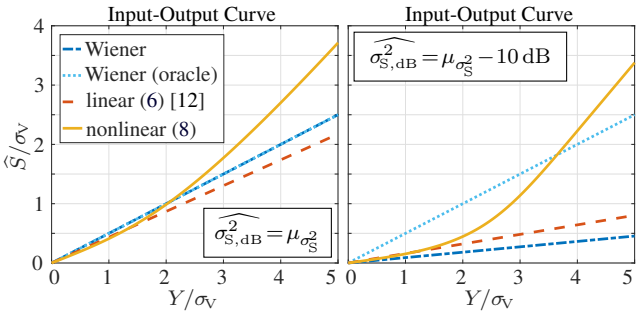


**Fig. 1**. IOCs for $\sigma_{\mathrm{V,dB}}^2 = \mu_{\sigma_{\mathrm{S}}^2} = -29\,\mathrm{dB}$, $\phi_{\sigma_{\mathrm{S}}^2} = 11\,\mathrm{dB}$, $Q = 10$. Only the speech PSD estimate $\widehat{\sigma_{\mathrm{S,dB}}^2}$ differs between the plots.

of the respective noisy input magnitude $|Y|$. Both, the input and the estimate, are normalized by $\sigma_{\mathrm{V}}$ to make the analysis less dependent on an absolute scaling. The lower the curve, the more suppression is applied by the respective estimator. As references, we also provide the IOCs of two versions of the Wiener filter: First, the conventional Wiener filter that uses only the PSD estimate $\widehat{\sigma_{\mathrm{S}}^2}$, which we denote as "Wiener". Second, the Wiener filter using the mean $\mu_{\sigma_{\mathrm{S}}^2}$ of $p\left(\sigma_{\mathrm{S,dB}}^2\right)$ instead of $\widehat{\sigma_{\mathrm{S}}^2}$, which we denote as "Wiener (oracle)". The two references can be interpreted as two extreme cases of the linear estimator [12]. While "Wiener" assumes that the estimate $\widehat{\sigma_{\mathrm{S}}^2}$ is exact, "Wiener (oracle)" assumes that $\mu_{\sigma_{\mathrm{S}}^2}$ is exact, both completely dismissing the uncertainty in the respective quantity. Note that we chose the Wiener filter as a reference so that all approaches are estimators of the complex clean speech coefficients $S$ and besides the PSD uncertainty model rely on the exact same statistical assumptions.

The two plots in Figure 1 differ only in how far the speech PSD estimate $\widehat{\sigma_{\mathrm{S,dB}}^2}$ deviates from the mean $\mu_{\sigma_{\mathrm{S}}^2}$ of the hyperhyper prior $p\left(\sigma_{\mathrm{S,dB}}^2\right)$ in the dB-domain. At the left, the PSD estimate $\widehat{\sigma_{\mathrm{S,dB}}^2}$ is exactly $\mu_{\sigma_{\mathrm{S}}^2}$, meaning that the estimate is likely to be close to the true speech PSD. Accordingly, the lines for "Wiener" and "Wiener (oracle)" overlap in Figure 1 (left). Furthermore, also the estimator [12] follows the Wiener filter. The difference between "linear (6) [12]" and "Wiener" is due to the remaining uncertainty in $\widehat{\sigma_{\mathrm{S,dB}}^2}$ and $\mu_{\sigma_{\mathrm{S}}^2}$. While the Wiener filter and [12] are linear estimators, the proposed approach is nonlinear and applies less suppression to large inputs. Note that this is a typical behavior that is also observed for estimators that are based on super-Gaussian speech priors like [4, 6].

At the right of Figure 1, the PSD estimate is 10 dB below $\mu_{\sigma_{\mathrm{S}}^2}$, meaning that the true speech PSD is likely to be higher than the estimate $\widehat{\sigma_{\mathrm{S}}^2}$. Thus, it is more likely that the input contains relevant speech energy and it would be beneficial to apply less suppression. In this situation the full potential of considering the PSD uncertainty is on display. Since the noise PSD is the same as before but the speech PSD estimate is lower, the conventional Wiener filter applies more suppression than in the first plot. In contrast, "Wiener (oracle)", which only relies on $\mu_{\sigma_{\mathrm{S}}^2}$ applies far less suppression. The two uncertainty-aware estimators trade-off the PSD estimate and the information about $p\left(\sigma_{\mathrm{S}}^2\right)$ according to the uncertainty model in (9). This theoretically justified compromise is controlled by the uncertainty of $\widehat{\sigma_{\mathrm{S}}^2}$ and $\mu_{\sigma_{\mathrm{S}}^2}$, which is modeled by the hyper prior $p\left(\widehat{\sigma_{\mathrm{S}}^2} \mid \sigma_{\mathrm{S}}^2\right)$ and the hyperhyper prior $p\left(\sigma_{\mathrm{S}}^2\right)$, respectively. The

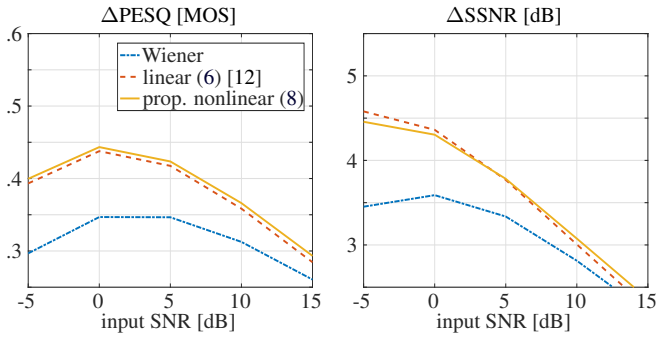


**Fig. 2**. Improvement in PESQ and SSNR over the noisy signal when the speech PSD is estimated via temporal spectrum smoothing.

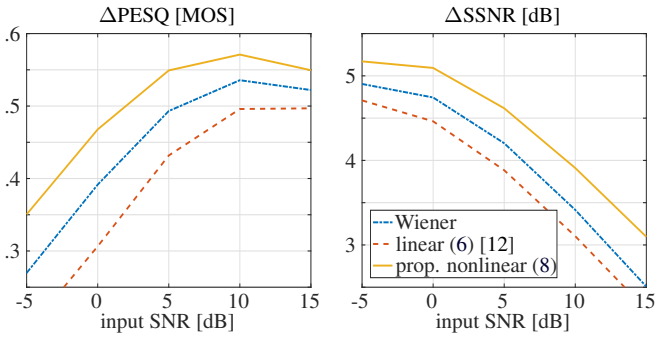


**Fig. 3**. Improvement in PESQ and SSNR over the noisy signal when the speech PSD is estimated via temporal cepstrum smoothing.

linear estimator [12] accordingly provides a compromise between "Wiener" and "Wiener (oracle)". The proposed estimator, due to its nonlinearity provides a completely new IOC and applies less attenuation than [12] and even "Wiener (oracle)" for large arguments.

## 6. EVALUATION

The estimators are evaluated on 128 gender-balanced sentences from the test set of the TIMIT [16] database, degraded by white noise, white noise modulated with a frequency of 0.5 Hz, street noise, and speech shaped noise at various SNRs. The STFT is computed with a segment length of 32 ms, an overlap of 50 %, and a square-root Hann window for analysis and synthesis. The maximum attenuation in each time-frequency point is set to -15 dB to avoid undesired artifacts and speech distortions. The noise PSD $\sigma_{\mathrm{V}}^2$ is estimated via [17]. We evaluate segmental SNR (SSNR) and 'Perceptual Evaluation of Speech Quality' (PESQ), which has been shown to correlate with the overall quality of spectrally enhanced speech [18]. For a better visualization we present the improvement over the unprocessed noisy signal instead of absolute values and average the results over all noise types.

The algorithms are evaluated for two speech PSD estimators of different quality. In the first part, similar to [12], the instantaneous PSD estimates are simply averaged over $Q = 10$ neighboring segments directly in the spectrum, corresponding to a time window of 176 ms. As a more sophisticated estimator, we then employ TCS [9, 13], which has been shown to provide estimates that allow for high quality speech estimation.

### 6.1. Speech PSD estimation via moving average in the spectrum

Figure 2 shows that both, the linear and the proposed nonlinear speech estimator under PSD uncertainty outperform the Wiener filter reference in terms of PESQ and SSNR. The largest improvements are achieved at low SNRs, with improvements of about 0.1 PESQ and 1 dB in SSNR. Based on informal listening, in combination with this simple PSD estimator the Wiener filter shows strong and annoying musical noise. When using the speech estimators under PSD uncertainty, however, musical noise is substantially reduced. As stated in Section 4 and the end of Section 5, speech PSD estimates are deemed less reliable in low SNRs by the uncertainty model and thus the influence of the hyperhyper prior $p(\sigma_{\mathrm{S}}^2)$, increases. Accordingly, at low SNRs where musical noise is most prominent, the uncertainty-aware estimators rely more on the hyperhyper prior rather than the fluctuating error-prone PSD estimates, effectively reducing musical noise. The two uncertainty-aware approaches yield virtually the same PESQ and SSNR improvements. Informal listening however reveals that the proposed nonlinear estimator better preserves speech while slightly more musical noise remains. This trade-off is characteristic for nonlinear, e.g. super-Gaussian, estimators, see e.g. [4].

### 6.2. TCS-based speech PSD estimation

Compared to the PSD estimator from the previous section, TCS greatly reduces random fluctuations in the PSD estimates while avoiding temporal smearing of speech. As a result, the Wiener filter achieves a higher signal quality with less musical noise, which is also indicated by larger PESQ and SSNR improvements in Figure 3.

The linear estimator [12] is the most aggressive of the three approaches, for TCS as well as for the simple PSD estimator in the previous section. The main benefit of [12] over the Wiener filter in the last section came from its improved suppression of musical noise, which outweighed the accompanying speech distortions. Since with TCS musical noise is less of a problem, the improvement diminishes and [12] does not improve over the Wiener filter anymore. In contrast, the proposed nonlinear estimator yields a higher predicted speech quality than the Wiener filter even for sophisticated TCS-based PSD estimates. First, it benefits from the same PSD uncertainty model as [12]. Second, due to its nonlinear behavior, with IOCs resembling those of super-Gaussian estimators in Figure 1, it protects speech components better than the linear estimator. The improvement relative to the Wiener filter is nevertheless smaller than for the simple PSD estimator in Figure 2. With a more reliable PSD estimator, the detrimental effect of neglecting its uncertainty and thus the benefit of uncertainty-aware speech estimators reduces.

## 7. CONCLUSIONS

In practice, the true PSD of speech is unknown and only estimates are available. To increase the robustness of speech enhancement frameworks to PSD estimation errors, recently a clean speech estimator has been proposed that explicitly takes into account the speech PSD uncertainty [12]. Here we avoid the restrictive assumption of [12] that the noisy input $Y$ does not yield additional information on $\sigma_{\mathrm{S}}^2$ when its ML estimate $\widehat{\sigma_{\mathrm{S}}^2}$ is given. The derivation then yields a fundamentally different estimator which is a nonlinear function of the noisy input. In contrast to the linear estimator in [12], the proposed nonlinear approach improves the predicted speech quality even when a sophisticated PSD estimator is employed.

## 8. REFERENCES


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

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

[3] Rainer 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, Sept. 2005.

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

[5] Chang Huai You, Soo Ngee Koh, and Susanto Rahardja, "$\beta$-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, July 2005.

[6] Colin Breithaupt, Martin Krawczyk, and Rainer Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4037–4040.

[7] Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "Log-spectral magnitude MMSE estimators under super-Gaussian densities," in *ISCA Interspeech*, Brighton, UK, Sept. 2009, pp. 1319–1322.

[8] Richard C. Hendriks, Timo Gerkmann, and Jesper Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-art*, Morgan & Claypool, Colorado, USA, Feb. 2013.

[9] Colin Breithaupt, Timo Gerkmann, and Rainer Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4897–4900.

[10] Timo Gerkmann and Rainer Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

[11] Israel Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *ELSEVIER Signal Process.*, vol. 86, no. 4, pp. 698–709, Apr. 2006.

[12] G. Enzner and P. Thüne, "Robust MMSE filtering for single-microphone speech enhancement," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 4009–4013.

[13] Timo Gerkmann and Rainer Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[14] Martin Krawczyk-Becker and Timo Gerkmann, "On MMSE-based estimation of spectral speech coefficients under phase-uncertainty," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 12, pp. 2251–2262, Dec. 2016.

[15] Jack E. Porter and Steven F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, San Diego, CA, USA, Mar. 1984, pp. 18A.2.1–18A.2.4.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[17] Timo Gerkmann and Richard C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

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