On Speech Enhancement Under PSD Uncertainty

Martin Krawczyk-Becker, *Member, IEEE*, and Timo Gerkmann, *Senior Member, IEEE*

***Abstract*— Many well-known and frequently employed Bayesian clean speech estimators have been derived under the assumption that the true power spectral densities (PSDs) of speech and noise are exactly known. In practice, however, only PSD estimates are available. Simply neglecting PSD estimation errors and handling the estimates as true values leads to speech estimation errors causing musical noise and undesired suppression of speech. In this paper, the uncertainty of the available speech PSD estimates is addressed. The main contributions are: (1) we summarize and examine ways to model and incorporate the uncertainty of PSD estimates for a more robust speech enhancement performance. (2) a novel nonlinear clean speech estimator is derived that takes into account prior knowledge about the absolute value of typical speech PSDs. (3) we show that the derived statistical framework provides uncertainty-aware counterparts to a number of well-known conventional clean speech estimators such as the Wiener filter and Ephraim and Malah's amplitude estimators. (4) we show how modern PSD estimators can be incorporated into the theoretical framework and propose to employ frequency dependent priors. Finally, the effects and benefits of considering the uncertainty of speech PSD estimates are analyzed, discussed, and evaluated via instrumental measures and a listening experiment.**

***Index Terms*—Speech enhancement, noise reduction, power spectral density, uncertainty**

I. Introduction

The enhancement of speech that has been corrupted by noise is a challenging and important field of research as it is an indispensable step to make communication devices like hearing aids or mobile phones work reliably also in adverse acoustic scenarios, i.e. on a busy street or in a crowded restaurant. Over the last decades, numerous speech enhancement approaches have been proposed. Here we concentrate on *single-channel* speech enhancement, which can be used in isolation if only a single microphone is available, but also as a post processing step after a multi-microphone preprocessing stage to further improve its performance. Among the most commonly used single-channel approaches are arguably Bayesian clean speech estimators working in the short-time discrete Fourier transform (STFT) domain. Besides the classical approaches, over the last years there has been an ever increasing interest in utilizing machine learning techniques like deep learning for speech enhancement, e.g. [1]. Nevertheless, Bayesian estimators remain relevant, as, for instance, they are fairly general and as opposed to deep neural networks do not rely on lengthy training [2]. It is further possible to combine Bayesian clean speech estimators with machine learning techniques, joining forces for an improved speech enhancement performance as shown in [3], [4]. Finding novel Bayesian estimators, which is what we do in this paper, hence also benefits state-of-the-art machine-learning based approaches.

The authors are with the Signal Processing Group, Department of Informatics, Universität Hamburg, 20148 Hamburg, Germany, e-mail: {martin.krawczyk-becker, timo.gerkmann}@uni-hamburg.de, web: http://www.inf.uni-hamburg.de/sp.

Well known examples of Bayesian estimators are the Wiener filter and Ephraim and Malah's short-time spectral amplitude estimator (STSA) [5]. For the derivation of such estimators, the PSDs are typically assumed to be deterministic and known. In practice, however, the true PSDs are not known and only estimates are available, which for instance are obtained from the noisy signal via maximum likelihood (ML) estimation, the decision-directed approach [5], or temporal cepstrum smoothing (TCS) [6]. Multiple approaches to increase the accuracy of speech PSD estimators have been proposed, for instance the iterative bias compensation mechanism [7] that aims at improving the performance of the decision-directed approach. But even in the noise free case the true speech PSDs can in principle not be determined as speech is a highly non-stationary and thus non-ergodic process [8]. The uncertainty in the speech PSDs has for instance been considered in [9] for the derivation of an improved speech PSD estimator based on a generalized autoregressive conditional heteroscedasticity (GARCH) model. In the GARCH model the true speech PSD itself is not handled as a deterministic parameter but modeled as an unobservable random variable. The resulting PSD estimator can be used as an alternative to, e.g., the well-known decision-directed approach [5]. In clean speech estimators, however, PSD estimates are commonly interpreted as true deterministic values, by which the uncertainty of the PSD estimates is completely neglected. PSD estimation errors thus directly propagate through to the final speech estimate, leading to distortions and/or a suboptimal noise reduction.

In this paper, which extends our conference paper [10], this problem is addressed and a new minimum mean square error (MMSE) optimal clean speech estimator is derived that explicitly takes into account the uncertainty of the available speech PSD estimates for an increased robustness. For conciseness and simplicity the uncertainty of the noise PSD is not addressed at this point. Nevertheless, many of the concepts presented here are also applicable to the noise PSD. Similar to [9], we explicitly assume that only an estimate of the speech PSD is given, while the true speech PSD is modeled as an unobservable random variable. Following this rationale, the speech prior becomes a scale mixture model [11], with the scale being the speech PSD. The challenge then lies in finding a suitable model of how the true PSD is distributed given its estimate. Well-known estimators, like the super-Gaussian estimator [12] and the estimator in [13] that is based on a multidimensional normal inverse Gaussian (MNIG) speech prior, arise as special cases. While the models used to obtain [12], [13] are chosen merely for their mathematical tractability, in this paper we use an interesting, recently proposed model [14] that follows from a strict Bayesian derivation. In [14], not

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2816241, IEEE/ACM Transactions on Audio, Speech, and Language Processing

only the true speech PSD but also its estimate are modeled as a random variable. The advantage of the formulation in [14] is twofold: First, for ML estimates of the speech PSD, a theoretically motivated relation between the true PSD and its estimate can be found. This relation also holds for smoothed ML estimates of the speech PSD, i.e. obtained via a temporal moving average on the spectrum [14]. While a simple moving average leads to undesired smearing of the speech PSD [5], more elaborate approaches like TCS [6], [15] have been shown to effectively reduce musical noise without smearing the speech, which improves the overall speech enhancement performance. Therefore, here we apply TCS [6], [15] and show how TCS can be integrated into the statistical model of [14]. Secondly, the model in [14] also provides a convenient and theoretically rigorous way to incorporate prior information about the true clean speech PSD, which for instance can be obtained off-line from a representative clean speech database.

Already in [14], the model of the distribution of the true speech PSD given only its estimate has been used to derive a clean speech estimator. We show that the proposed estimator and the one in [14] are two different solutions to the same problem, i.e. both start from the exact same problem formulation. The crucial difference lies in a critical assumption made in [14]. There, it is assumed that when the ML estimate of the speech PSD is given, the noisy observation does not provide additional information with respect to the true PSD. In this paper, we argue that this assumption is not true in general and show how this assumption can be avoided. Interestingly, we show that avoiding this assumption yields the fundamental difference that the resulting speech estimator is a nonlinear function of the noisy input – thus yielding a potentially more powerful estimator and building the bridge between uncertain speech PSDs and super-Gaussian estimators like in [12] as well as MNIG approaches [13].

After briefly introducing the notation and conventional Bayesian clean speech estimation without considering PSD uncertainty in Section II, a nonlinear estimator under speech PSD uncertainty is derived and compared to a linear alternative in Section III. In Section IV, a statistically rigorous model of the PSD uncertainty is presented, refined, and analyzed, before specific uncertainty-aware counterparts to well known clean speech estimators are provided in Section V. Finally, two uncertainty-aware estimators are analyzed in terms of their input-output characteristics (IOCs) in Section VI and evaluated with instrumental measures and a pairwise preference test in Section VII.

II. Conventional MMSE clean speech estimation

In the STFT domain, we denote the complex-valued spectral coefficients of the noisy signal at segment ℓ and frequency bin k 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 all processing steps are performed separately for each time-frequency point (ℓ, k), the indices are dropped for notational convenience. Both, S and V are modeled as zero-mean complex-valued random variables. The respective true PSDs are denoted by σ_{S}^2, σ_{V}^2, and $\sigma_{\mathrm{Y}}^2 = \sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{V}}^2$. To distinguish estimates from their true counterparts the hat symbol is introduced, e.g. $\widehat{\sigma_{\mathrm{S}}^2}$ is an estimate of σ_{S}^2.

$p\big(S\vert\widehat{\sigma_{\mathrm{S}}^2}\big)$	prior
$p\big(S\vert\sigma_{\mathrm{S}}^2\big)$	oracle prior
$p\big(\widehat{\sigma_{\mathrm{S}}^2}\vert\sigma_{\mathrm{S}}^2\big)$	hyperprior
$p\big(\sigma_{\mathrm{S}}^2\big)$	hyperhyperprior (HHP)
$p\big(\sigma_{\mathrm{S}}^2\vert\widehat{\sigma_{\mathrm{S}}^2}\big)$	PSD uncertainty model

Table I: Nomenclature for estimating clean speech coefficients S and functions $f(S)$ thereof via (3).

Conventional MMSE optimal estimators of the clean speech coefficients S — or functions $f(S)$ thereof — are conditioned not only on the noisy observation Y, but also on the PSDs of noise and speech

$$\begin{aligned}\widehat{f(S)} &= \mathrm{E}\big(f(S)\vert Y, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{V}}^2\big) \\ &= \int_S f(S) p\big(S\vert Y, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{V}}^2\big)\,\mathrm{d}S,\end{aligned} \tag{2}$$

with the speech posterior $p\big(S\vert Y, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{V}}^2\big)$. Common and well established choices of the function $f(S)$ are the complex coefficient S itself as for the Wiener filter or the spectral amplitude $|S|$ as for Ephraim and Malah's STSA [5]. Since σ_{V}^2 is treated as a known deterministic parameter, in the remainder of this paper the dependency on σ_{V}^2 is not stated explicitly, e.g. we write $p\big(S\vert Y, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{V}}^2\big) = p\big(S\vert Y, \sigma_{\mathrm{S}}^2\big)$ for notational convenience. As a result, every probability density function in the sequel is implicitly conditioned on the noise PSD σ_{V}^2.

III. Clean speech estimation under speech PSD uncertainty

In (2), the true PSDs of speech and noise are modeled as being perfectly known. If only an estimate $\widehat{\sigma_{\mathrm{S}}^2}$ of the true speech PSD σ_{S}^2 is available, analogously to (2), the MMSE optimal clean speech estimator is given by

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

Starting from (3), we first introduce the existing clean speech estimator [14]. We show that the approach in [14] relies on a rather restrictive simplification that constrains the estimator to a linear function of the noisy input Y. In the second part, we show that without this simplification a fundamentally different estimator is derived that, similar to super-Gaussian estimators like [12], is nonlinear in Y. Please note that the formulation in [14] is less general than the one in this paper as it only provides an estimator of the complex coefficients $f(S) = S$. For clarity, some quantities that are frequently encountered in the derivations are summarized in Table I.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2816241, IEEE/ACM Transactions on Audio, Speech, and Language Processing

A. Existing linear estimator [14]

To obtain the MMSE estimator in [14], we first apply Bayes' rule to the speech posterior in (3) in a fashion that gives

$$p\left(S|Y,\widehat{\sigma_{\mathrm{S}}^2}\right) = \frac{\int_0^\infty p\left(S,Y,\widehat{\sigma_{\mathrm{S}}^2},\sigma_{\mathrm{S}}^2\right)\mathrm{d}\sigma_{\mathrm{S}}^2}{p\left(Y,\widehat{\sigma_{\mathrm{S}}^2}\right)}$$
$$= \int_0^\infty p\left(S|Y,\widehat{\sigma_{\mathrm{S}}^2},\sigma_{\mathrm{S}}^2\right) p\left(\sigma_{\mathrm{S}}^2|Y,\widehat{\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\sigma_{\mathrm{S}}^2$$
$$\approx \int_0^\infty p\big(S|Y,\sigma_{\mathrm{S}}^2\big)\, p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\sigma_{\mathrm{S}}^2, \tag{4}$$

where in the numerator of the first line the joint distribution $p\left(S,Y,\widehat{\sigma_{\mathrm{S}}^2}\right)$ is expressed in terms of the marginal probability of $p\left(S,Y,\widehat{\sigma_{\mathrm{S}}^2},\sigma_{\mathrm{S}}^2\right)$. For the posterior $p\left(S|Y,\widehat{\sigma_{\mathrm{S}}^2},\sigma_{\mathrm{S}}^2\right)$, one can safely assume that once the true speech PSD σ_{S}^2 and Y are given, the estimate $\widehat{\sigma_{\mathrm{S}}^2}$ does not provide any additional information. The second simplification, i.e. $p\left(\sigma_{\mathrm{S}}^2|Y,\widehat{\sigma_{\mathrm{S}}^2}\right) \approx p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)$ is more restrictive. The argument in [14] is that the estimate $\widehat{\sigma_{\mathrm{S}}^2}$ is obtained from the noisy signal Y and contains all information about the true PSD σ_{S}^2 inherent in Y.

The simplified speech posterior (4) is plugged into (3) with $f(S) = S$ to obtain the clean speech estimator in [14]:

$$\widehat{S}_{[14]} = \int_0^\infty \int_S S p\big(S|Y,\sigma_{\mathrm{S}}^2\big)\ \mathrm{d}S\ p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\sigma_{\mathrm{S}}^2. \tag{5}$$

Note that the inner integral constitutes a conventional MMSE estimator given the true PSDs of speech and noise (2). When the noise prior and the speech prior given their true PSDs are Gaussian distributed, which is assumed in [14] as well as in this paper, this estimator is the conventional Wiener filter $\widehat{S}_W = \sigma_{\mathrm{S}}^2/\left(\sigma_{\mathrm{S}}^2+\sigma_{\mathrm{V}}^2\right)Y$. Accordingly, the clean speech estimator under PSD uncertainty (5) can be rewritten as [14]

$$\widehat{S}_{[14]} = Y\int_0^\infty \frac{\sigma_{\mathrm{S}}^2}{\sigma_{\mathrm{S}}^2+\sigma_{\mathrm{V}}^2}\, p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\sigma_{\mathrm{S}}^2 = Y\, G_{[14]}, \tag{6}$$

where the spectral gain $G_{[14]}$ under PSD uncertainty is a weighted mixture of Wiener filter gains. The Wiener filter itself is linear in Y, hence its spectral gain is independent of the noisy input Y. Due to the second simplification, i.e. $p\left(\sigma_{\mathrm{S}}^2|Y,\widehat{\sigma_{\mathrm{S}}^2}\right) \approx p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)$, also the weights in (6) are independent of Y. Consequently, also the resulting gain under PSD uncertainty $G_{[14]}$ is independent of Y and hence the estimator $\widehat{S}_{[14]}$ is linear in Y. However, linear estimators are often outperformed by nonlinear ones, a prominent example being the Wiener filter versus the nonlinear super-Gaussian estimator in [12]. Next, we show that when the simplification of [14] is avoided, even under the same statistical assumptions for speech and noise, the estimator under PSD uncertainty becomes a nonlinear function of the noisy input.

B. Proposed nonlinear estimator

For the derivation of the proposed nonlinear estimator under PSD uncertainty, we again start by reformulating the speech posterior in (3) via Bayes' rule, but in a different manner than in [14]:

$$p\left(S|Y,\widehat{\sigma_{\mathrm{S}}^2}\right) = \frac{\int_0^\infty p\left(S,Y,\widehat{\sigma_{\mathrm{S}}^2},\sigma_{\mathrm{S}}^2\right)\mathrm{d}\sigma_{\mathrm{S}}^2}{\int_S\int_0^\infty p\left(S,Y,\widehat{\sigma_{\mathrm{S}}^2},\sigma_{\mathrm{S}}^2\right)\mathrm{d}\sigma_{\mathrm{S}}^2\mathrm{d}S} =$$
$$\frac{\int_0^\infty p\left(Y|S,\sigma_{\mathrm{S}}^2,\widehat{\sigma_{\mathrm{S}}^2}\right) p\left(S|\sigma_{\mathrm{S}}^2,\widehat{\sigma_{\mathrm{S}}^2}\right) p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\sigma_{\mathrm{S}}^2}{\int_S p\left(Y|S,\sigma_{\mathrm{S}}^2,\widehat{\sigma_{\mathrm{S}}^2}\right)\int_0^\infty p\left(S|\sigma_{\mathrm{S}}^2,\widehat{\sigma_{\mathrm{S}}^2}\right) p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\sigma_{\mathrm{S}}^2\mathrm{d}S}$$
$$\approx \frac{p(Y|S)\int_0^\infty p\big(S|\sigma_{\mathrm{S}}^2\big)\, p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\sigma_{\mathrm{S}}^2}{\int_S p(Y|S)\int_0^\infty p(S|\sigma_{\mathrm{S}}^2)\, p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\sigma_{\mathrm{S}}^2\mathrm{d}S}, \tag{7}$$

where in the first line the denominator is expressed as the marginal distribution of $p\left(Y,S,\widehat{\sigma_{\mathrm{S}}^2},\sigma_{\mathrm{S}}^2\right)$ such that the exact same formulations and assumptions can be applied to the denominator as to the numerator. Note that the linear estimator in (6) and the proposed nonlinear estimator both rely on the uncertainty model $p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)$. However, in contrast to the linear estimator, where the simplification $p\left(\sigma_{\mathrm{S}}^2|Y,\widehat{\sigma_{\mathrm{S}}^2}\right) \approx p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)$ has to be made, here the uncertainty model results from the Bayesian reformulation of the speech posterior in (7). For mutually independent speech and noise, the likelihood $p\left(Y|S,\sigma_{\mathrm{S}}^2,\widehat{\sigma_{\mathrm{S}}^2}\right) \approx p(Y|S)$ is assumed to be the probability density function of the noise V shifted by S and thus neither depends on the true nor the estimated speech PSD. For Gaussian distributed noise we have, e.g. [5]:

$$p(Y|S) = \frac{1}{\pi\sigma_{\mathrm{V}}^2}\exp\left(-\frac{|Y-S|^2}{\sigma_{\mathrm{V}}^2}\right), \tag{8}$$

which is the same model used in conventional speech estimators that do not incorporate PSD uncertainty. We further assume that once the true speech PSD σ_{S}^2 is given, its estimate $\widehat{\sigma_{\mathrm{S}}^2}$ does not provide any additional information regarding S, leading to $p\left(S|\sigma_{\mathrm{S}}^2,\widehat{\sigma_{\mathrm{S}}^2}\right) \approx p\big(S|\sigma_{\mathrm{S}}^2\big)$. Since it is conditioned on the true speech PSD, which is not available in practice, we denote $p\big(S|\sigma_{\mathrm{S}}^2\big)$ as the *oracle* speech prior. The speech prior conditioned on the available speech PSD estimate is given by the integral over σ_{S}^2 in (7):

$$p\left(S|\widehat{\sigma_{\mathrm{S}}^2}\right) = \int_0^\infty p\big(S|\sigma_{\mathrm{S}}^2\big)\, p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\sigma_{\mathrm{S}}^2. \tag{9}$$

This kind of model, where the scale parameter σ_{S}^2 of the distribution of the desired quantity S is modeled as a random variable, is known as a scale mixture model. Equation (9) can be seen as an averaging of the oracle prior $p\big(S|\sigma_{\mathrm{S}}^2\big)$ over all possible values of the true σ_{S}^2 with a weighting based on the uncertainty model $p\left(\sigma_{\mathrm{S}}^2|\widehat{\sigma_{\mathrm{S}}^2}\right)$. Scale mixture models are commonly used in financial/economic prediction [16], but

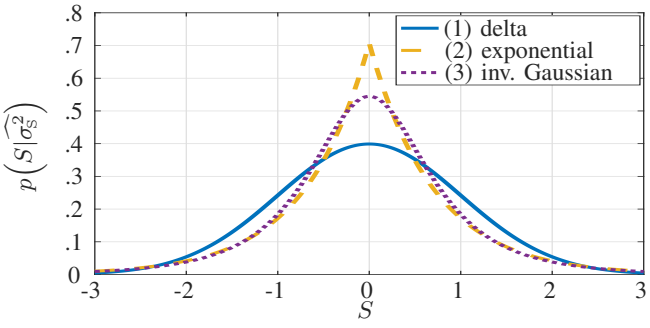

Figure 1. Speech prior $p\left(S|\widehat{\sigma_S^2}\right)$ for $\widehat{\sigma_S^2} = 1$ and different speech PSD uncertainty models $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$: (1) delta pulse at $\widehat{\sigma_S^2}$ leading to a Gaussian prior; (2) exponential distribution leading to a Laplace prior; (3) inverse Gaussian distribution with $\alpha_{[13]} = 0.9$ and $\delta_{[13]} = 0.9$ leading to a normal inverse Gaussian prior [13]. For simplicity, here we consider only real-valued speech coefficients S.

have also been used for speech PSD estimation [9] and speech enhancement [13].

IV. Models of speech PSD uncertainty $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$

The key to accurately incorporating the uncertainty of the speech PSD estimates is finding an adequate model of the speech PSD uncertainty $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$, which directly influences the form of the speech prior $p\left(S|\widehat{\sigma_S^2}\right)$ in (9). Specific examples are illustrated in Figure 1, where the resulting speech priors $p\left(S|\widehat{\sigma_S^2}\right)$ are presented for $\widehat{\sigma_S^2} = 1$. If the PSD estimates are assumed to be perfect, $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ is set to a Dirac impulse $\delta\left(\sigma_S^2 - \widehat{\sigma_S^2}\right)$. The speech prior (9) then becomes $p\left(S|\widehat{\sigma_S^2}\right) = p\left(S|\sigma_S^2 = \widehat{\sigma_S^2}\right)$, i.e. it follows the same model as the oracle speech prior, a Gaussian distribution in our case, but using the estimated PSD. In Figure 1, we consider two more choices of $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ that both lead to speech priors that have already been used for speech enhancement in the literature. First, an exponential distribution for $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$, which yields a Laplace speech prior $p\left(S|\widehat{\sigma_S^2}\right)$ [17]. Second, an inverse Gaussian distribution, which leads to a *normal inverse Gaussian* speech prior [16]. While the Laplace distribution has been used in [12], the normal inverse Gaussian distribution has been employed in [13] to derive clean speech estimators. These two models of the PSD uncertainty $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ provide mathematically tractable super-Gaussian speech priors $p\left(S|\widehat{\sigma_S^2}\right)$, which is argued to fit clean speech histograms better than a simple Gaussian distribution [12]. However, they are merely pragmatic choices and not necessarily represent theoretically justified models of the uncertainty of speech PSD estimates. For instance, given a reasonably accurate PSD estimate, there is little reason to believe that the true σ_S^2 follows an inverse Gaussian distribution.

A statistically rigorous model has recently been proposed in [14] and adopted here, which is solely based on assumptions about the employed speech PSD estimator as well as potential prior information about the true speech PSD. Similar to the examples in Figure 1, also this uncertainty model results in a super-Gaussian speech prior. But in contrast to the examples above, the exact shape of the speech prior may be different in each time frequency point based on the respective PSD estimate and the prior information about σ_S^2. We first outline the speech PSD uncertainty model proposed in [14] and then modify it for an improved speech enhancement performance.

A. A statistical model of the PSD uncertainty $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$

To find a statistically rigorous model of the PSD uncertainty $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ in (9), we reformulate $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ using Bayes' rule as in [14]

$$p\left(\sigma_S^2|\widehat{\sigma_S^2}\right) = \frac{p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)p\left(\sigma_S^2\right)}{p\left(\widehat{\sigma_S^2}\right)} \propto p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)p\left(\sigma_S^2\right). \quad (10)$$

In (10), we dropped the denominator since it cancels out when inserting $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ into (7) for the computation of the speech posterior. Thanks to the reformulation in (10), $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ is now split into two parts: the *hyperprior* $p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)$ which depends on the specific speech PSD estimator that is employed to obtain $\widehat{\sigma_S^2}$ and the *hyperhyperprior (HHP)* $p\left(\sigma_S^2\right)$, which allows to insert prior information about the true speech PSD. Please note that with the formulation in (10), both, the true and the estimated speech PSD are modeled as random variables.

B. Modeling the hyperprior $p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)$

Similar to [14], we use a χ^2 distribution to model the hyperprior $p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)$. We assume that the noisy observation Y given its PSD $\sigma_Y^2 = \sigma_S^2 + \sigma_V^2$ is zero-mean Gaussian distributed. Accordingly, $|Y|^2$ is exponentially distributed. An instantaneous ML estimate of the speech PSD is obtained via $\widehat{\sigma_S^2} = \max\left(|Y|^2 - \sigma_V^2, 0\right)$ [5]. With known σ_V^2 and neglecting the maximum operator, this ML estimate follows the same exponential distribution as $|Y|^2$, only shifted by σ_V^2. Considering the exponential distribution a special case of a χ^2 distribution with shape parameter $Q = 1$, we get [14]:

$$p\left(\widehat{\sigma_S^2}|\sigma_S^2\right) = \frac{(\widehat{\sigma_S^2} + \sigma_V^2)^{Q-1}Q^Q}{\Gamma(Q)} \frac{\exp\left(-Q\frac{\widehat{\sigma_S^2}+\sigma_V^2}{\sigma_S^2+\sigma_V^2}\right)}{(\sigma_S^2 + \sigma_V^2)^Q}. \quad (11)$$

Due to their strong temporal and spectral fluctuations, instantaneous PSD estimates are known to produce strong artifacts perceived as musical noise when directly used for clean speech estimation, e.g. [18]. A remedy to this problem is to smooth the instantaneous estimates. Smoothed χ^2 distributed random variables can be well modeled by a χ^2 distribution with an increased shape parameter Q [19], [20].

Intuitively, the reliability of a PSD estimate depends on the signal to noise ratio (SNR) and the amount of smoothing that is applied. This characteristic is covered by $p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)$ (11): the lower the noise PSD σ_V^2 and the higher Q, i.e. the more smoothing is applied, the more concentrated $p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)$

is around the true PSD σ_{S}^2. Please note that in practice the amount of smoothing that can be applied is limited by the non-stationarity of speech. At low SNR $\sigma_{\mathrm{S}}^2 \ll \sigma_{\mathrm{V}}^2$ speech PSD estimation is more challenging and inevitably less reliable. Accordingly, $p\left(\widehat{\sigma_{\mathrm{S}}^2}|\sigma_{\mathrm{S}}^2\right)$ becomes broader, accounting for the increased uncertainty in the PSD estimate.

In [14], the instantaneous PSD estimates are smoothed over time by means of a moving average in each frequency channel, which increases the shape parameter Q to the number of segments used for the moving average. This has been shown to be the ML estimate given a sequence of observations under the simplification that neighboring time frequency points are independent and identically distributed [5]. However, already Ephraim and Malah [5] 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. This is why a temporal moving average estimate of the speech PSD is rarely used in speech enhancement. To alleviate the smearing of speech, Ephraim and Malah proposed to use the decision-directed approach instead [5]. While the decision-directed approach effectively reduces musical noise and reduces the smearing of speech onsets, due to its nonlinearity it is not clear how the uncertainty of the resulting estimate, i.e., $p\left(\widehat{\sigma_{\mathrm{S}}^2}|\sigma_{\mathrm{S}}^2\right)$ can be modeled in a meaningful way. Thus, it is not well suited to be used in this setup.

Proposed improved estimation of the speech PSD: Fortunately, there is a sophisticated state-of-the-art alternative to the decision-directed approach based on temporal cepstrum smoothing (TCS) [6], [15], for which a meaningful model of the uncertainty can be obtained. We generally prefer TCS over the decision-directed approach, as it has been shown to produce less musical noise while providing a more natural sounding background noise [6]. In contrast to temporal smoothing in the spectral domain as in [14], TCS recursively smoothes the instantaneous PSD estimates over time in the cepstral domain. In the cepstrum, there are only few coefficients that contain speech related information, namely the lowest coefficients that represent the speech envelope and a single peak that corresponds to the fundamental period of voiced speech. To avoid distortions like the temporal smearing of speech onsets observed for the temporal moving average on the spectrum, the speech related cepstral coefficients are only slightly smoothed, while the remaining non-speech related coefficients are strongly smoothed. With this selective smoothing, undesired outliers in the final PSD estimate are strongly reduced while avoiding a smearing of the speech. It has been shown experimentally in [15] that the PSD estimates after TCS are well modeled by a χ^2 distribution. Furthermore, a direct relationship between the amount of smoothing in the cepstral domain and the resulting shape parameter Q of the PSD estimate in the spectral domain has been established. Details on how Q is calculated can be found in [15, Sec. IV], where $\widetilde{\mu}_{[15]}$ corresponds to Q after TCS.

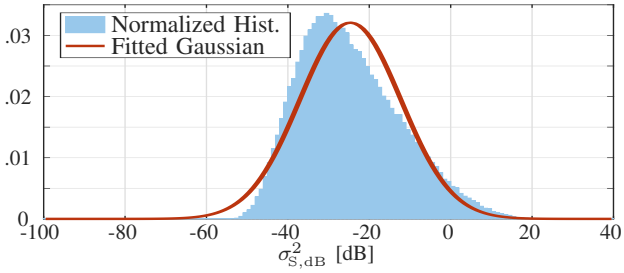

Figure 2. Histogram of $|S|^2$ in dB over an hour of TIMIT utterances for the frequency band at 1 kHz together with a ML fitted Gaussian distribution (mean $\mu_{\sigma_{\mathrm{S}}^2} \approx -24$ dB, standard deviation $\phi_{\sigma_{\mathrm{S}}^2} \approx 12.5$ dB).

C. Modeling the hyperhyperprior $p\left(\sigma_{\mathrm{S}}^2\right)$

As proposed in [14], the second part of (10), i.e. the HHP $p\left(\sigma_{\mathrm{S}}^2\right)$, allows to bring in prior information about the true PSDs of the desired speech sound. If no information about the true speech PSD is available, $p\left(\sigma_{\mathrm{S}}^2\right)$ could for instance be set to a uniform distribution, the boundaries set in accordance with the limitations of the recording setup. A promising way to include prior information in $p\left(\sigma_{\mathrm{S}}^2\right)$ has been proposed in [14]. Due to its non-stationarity, even for clean speech the true σ_{S}^2 is not available. However, the periodogram is available as an unbiased, yet variant, estimator of σ_{S}^2. Thus, a long-term histogram of $|S|^2$ is used to model the HHP. For this, a subset of the TIMIT training set is used that has been excluded from the evaluation in Section VII. This is depicted in Figure 2, where the histogram of $10\log_{10}\left(|S|^2\right)$ for 1 hour of gender balanced utterances from the TIMIT training set is shown. We excluded speech absence regions by considering only time-frequency points for which $|S|^2$ is at most 60 dB below the maximum $|S|^2$ of the respective utterance. As in [14], a Gaussian distribution with sample mean $\mu_{\sigma_{\mathrm{S}}^2}$ and standard deviation $\phi_{\sigma_{\mathrm{S}}^2}$ is fitted to the histogram. This Gaussian distribution is then taken as a model of the distribution of $\sigma_{\mathrm{S,dB}}^2 = 10\log_{10}\left(\sigma_{\mathrm{S}}^2\right)$. A Gaussian distribution of the logarithmic $\sigma_{\mathrm{S,dB}}^2$ corresponds to a log-normal distribution in the linear domain, which yields the proposed model of $p\left(\sigma_{\mathrm{S}}^2\right)$, [14]

$$p\left(\sigma_{\mathrm{S}}^2\right) = \frac{10}{\ln(10)\,\sigma_{\mathrm{S}}^2}\frac{1}{\sqrt{2\pi}\phi_{\sigma_{\mathrm{S}}^2}}\exp\left(-\frac{\left(\sigma_{\mathrm{S,dB}}^2 - \mu_{\sigma_{\mathrm{S}}^2}\right)^2}{2\phi_{\sigma_{\mathrm{S}}^2}^2}\right) \quad (12)$$

for $\sigma_{\mathrm{S}}^2 \geq 0$ and $p\left(\sigma_{\mathrm{S}}^2\right) = 0$ otherwise. 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 $\mu_{\sigma_{\mathrm{S}}^2}$.

Proposed frequency dependent HHP: In [14], the same $\mu_{\sigma_{\mathrm{S}}^2}$ and $\phi_{\sigma_{\mathrm{S}}^2}$ have been used in every frequency band. However, the long term envelope of speech is not flat and thus a frequency independent $p\left(\sigma_{\mathrm{S}}^2\right)$ will only fit the observed histograms in some frequency bands, while it might substantially deviate in others. Therefore, here we set $\mu_{\sigma_{\mathrm{S}}^2}$ and $\phi_{\sigma_{\mathrm{S}}^2}$ in (12) frequency dependently to achieve a better fit of $p\left(\sigma_{\mathrm{S}}^2\right)$ in each frequency band. For this, we compute the histograms from which we obtain $\mu_{\sigma_{\mathrm{S}}^2}$ and $\phi_{\sigma_{\mathrm{S}}^2}$ separately for each STFT band. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2816241, IEEE/ACM Transactions on Audio, Speech, and Language Processing

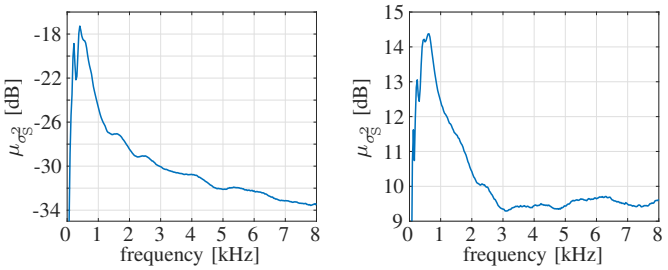

Figure 3. Mean $\mu_{\sigma_S^2}$ and standard deviation $\phi_{\sigma_S^2}$ of $\sigma^2_{\mathrm{S,dB}}$ over frequency.

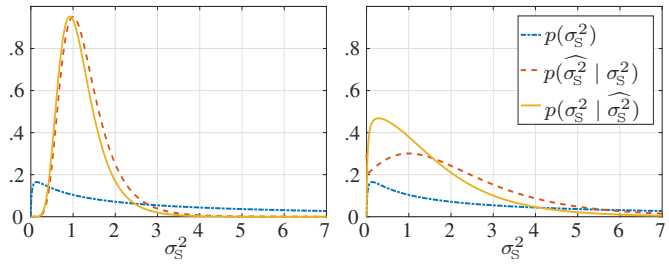

Figure 4. PSD uncertainty model $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ together with hyperprior $p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)$ and HHP $p\left(\sigma_S^2\right)$ for $\widehat{\sigma^2_{\mathrm{S,dB}}} = 0$ dB, $Q = 10$, $\mu_{\sigma_S^2} = 10$ dB, and $\phi_{\sigma_S^2} = 9$ dB. Left: estimated SNR $\widehat{\sigma_S^2}/\sigma_V^2$ of 5 dB. Right: estimated SNR $\widehat{\sigma_S^2}/\sigma_V^2$ of -5 dB. The stronger the noise, the less reliable the PSD estimate $\widehat{\sigma_S^2}$ is and the influence of the HHP $p\left(\sigma_S^2\right)$ on $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ increases.

frequency dependent $\mu_{\sigma_S^2}$ and $\phi_{\sigma_S^2}$ are depicted in Figure 3.

As already stated in [14], the form of the histogram and thus $p\left(\sigma_S^2\right)$ depends on the speech material and also on the employed STFT setup.

D. Interplay between hyperprior and HHP

With the models of the hyperprior $p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)$ and the HHP $p\left(\sigma_S^2\right)$ at hand, the PSD uncertainty model $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ can be obtained via (10). Two examples are plotted over the true speech PSD σ_S^2 in Figure 4. In both cases we have a PSD estimate of $\widehat{\sigma^2_{\mathrm{S,dB}}} = 0$ dB, i.e. $\widehat{\sigma_S^2} = 1$, while the expected value of the true PSD $\sigma^2_{\mathrm{S,dB}}$ is $\mu_{\sigma_S^2} = 10$ dB. Furthermore, $\phi_{\sigma_S^2}$ is set to 9 dB and Q is set to 10. The two plots differ only in the estimated SNR $\widehat{\sigma_S^2}/\sigma_V^2$. At the left of Figure 4, the estimated SNR is 5 dB, while at the right of Figure 4 the SNR is -5 dB. As stated in Section IV-B, the higher the estimated SNR, the more reliable the estimate $\widehat{\sigma_S^2}$ is, which is also reflected in the strong concentration of the hyperprior $p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)$ around $\widehat{\sigma_S^2} = 1$. Accordingly, the PSD uncertainty model $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ follows the hyperprior rather closely, while the less reliable prior information in the HHP $p\left(\sigma_S^2\right)$ only has a minor influence. For low estimated SNRs, the speech PSD estimate is less reliable and the influence of the HHP $p\left(\sigma_S^2\right)$ on the uncertainty model $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ is dominant, as illustrated in the right plot of Figure 4. Hence, the uncertainty model $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ provides a sensible compromise between the PSD estimate and the prior information about the true σ_S^2, putting more emphasize on whatever is deemed more reliable.

Summarizing this section, the employed model of the PSD uncertainty $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$ (10) lets us conveniently incorporate, both, prior information about the true speech PSD via the HHP $p\left(\sigma_S^2\right)$ (12) and the uncertainty of the employed speech PSD estimator via the hyperprior $p\left(\widehat{\sigma_S^2}|\sigma_S^2\right)$ (11). We further proposed a new way to incorporate the uncertainty of a state-of-the art speech PSD estimator based on TCS into this framework and made the HHP $p\left(\sigma_S^2\right)$ frequency dependent.

V. Uncertainty-aware counterparts to well-known conventional estimators

Inserting (12) and (11) into (10), we now have a model of the speech PSD uncertainty $p\left(\sigma_S^2|\widehat{\sigma_S^2}\right)$. With this model at hand, a clean speech estimator can be obtained using the speech posterior (7) in (3). The estimator can be realized, e.g., by solving the integrals in (3) and (7) numerically. Alternatively, we can insert (7) in (3) and change the order of the integrals:

$$\widehat{f(S)} = \frac{\int_0^\infty \int_S f(S) p\left(Y|S,\sigma_V^2\right) p\left(S|\sigma_S^2\right) \mathrm{d}S \; p\left(\sigma_S^2|\widehat{\sigma_S^2}\right) \mathrm{d}\sigma_S^2}{\int_0^\infty \int_S p\left(Y|S,\sigma_V^2\right) p\left(S|\sigma_S^2\right) \mathrm{d}S \; p\left(\sigma_S^2|\widehat{\sigma_S^2}\right) \mathrm{d}\sigma_S^2}, \tag{13}$$

where the inner integrals over S in the numerator and the denominator only depend on the true PSD σ_S^2. These integrals have been solved analytically for various $f(S)$ and $p\left(S|\sigma_S^2\right)$ as part of the derivation of some well established conventional estimators. For $f(S) = S$, (13) leads to

$$\widehat{S} = \frac{\int_0^\infty \frac{\sigma_S^2}{\left(\sigma_S^2+\sigma_V^2\right)^2} \mathrm{e}^{\nu} \; p\left(\sigma_S^2|\widehat{\sigma_S^2}\right) \mathrm{d}\sigma_S^2}{\int_0^\infty \frac{1}{\sigma_S^2+\sigma_V^2} \mathrm{e}^{\nu} \; p\left(\sigma_S^2|\widehat{\sigma_S^2}\right) \mathrm{d}\sigma_S^2} \, Y, \tag{14}$$

which is the counterpart to the Wiener filter under speech PSD uncertainty. Here we introduce $\nu = \frac{\sigma_S^2}{\sigma_S^2+\sigma_V^2} \frac{|Y|^2}{\sigma_V^2}$ for a concise notation. For $f(S) = |S|$, the counterpart to the STSA [5] under speech PSD uncertainty is obtained

$$\widehat{|S|} = \frac{\sqrt{\pi}}{2} \frac{\int_0^\infty \frac{1}{\sigma_S^2+\sigma_V^2} \frac{\sqrt{\nu}}{\gamma} \Phi\left(\frac{3}{2}, 1; \nu\right) \; p\left(\sigma_S^2|\widehat{\sigma_S^2}\right) \mathrm{d}\sigma_S^2}{\int_0^\infty \frac{1}{\sigma_S^2+\sigma_V^2} \mathrm{e}^{\nu} \; p\left(\sigma_S^2|\widehat{\sigma_S^2}\right) \mathrm{d}\sigma_S^2} \, |Y|, \tag{15}$$

with the a posteriori SNR $\gamma = \frac{|Y|^2}{\sigma_V^2}$ and the confluent hypergeometric function $\Phi(\cdot,\cdot;\cdot)$. For $f(S) = \log(|S|)$ we get the counterpart to the log-spectral amplitude estimator (LSA) [21] under speech PSD uncertainty:

$$\widehat{|S|} = \frac{\int_0^\infty \frac{\sigma_S^2}{\left(\sigma_S^2+\sigma_V^2\right)^2} \mathrm{e}^{\nu+\frac{1}{2}\mathrm{expint}(\nu)} \; p\left(\sigma_S^2|\widehat{\sigma_S^2}\right) \mathrm{d}\sigma_S^2}{\int_0^\infty \frac{1}{\sigma_S^2+\sigma_V^2} \mathrm{e}^{\nu} \; p\left(\sigma_S^2|\widehat{\sigma_S^2}\right) \mathrm{d}\sigma_S^2} \, |Y|, \tag{16}$$

where $\mathrm{expint}(\nu)$ denotes the exponential integral function. With (14), (15), and (16) only the outer integral over σ_S^2 needs

Algorithm 1 Proposed speech enhancement under speech PSD uncertainty

Input: noisy speech Y, HHP parameters $\mu_{\sigma_\mathrm{S}^2}$, $\phi_{\sigma_\mathrm{S}^2}$
Output: clean speech estimate $\widehat{S}$
1: **for** each segment **and** each frequency band **do**
2: estimate the noise PSD σ_V^2, e.g. via [22]
3: estimate the speech PSD $\widehat{\sigma_\mathrm{S}^2}$, e.g. via TCS [6], [15]
4: compute the form parameter Q, e.g. for TCS via [15]
5: compute $\widehat{S}$ via (14), (15), or (16) using (10) – (12)
6: **end for**

to be solved numerically. Without analytical solutions to the integrals, the complexity is significantly higher than that of the conventional counterparts. To reduce the complexity at the cost of an increased memory consumption, the relation between Y and $\widehat{S}$ can also be tabulated. In contrast to conventional estimators, which typically only require a two-dimensional look-up table, here the table has 5 dimensions, i.e. Y, $\widehat{\sigma_\mathrm{S}^2}$, σ_V^2, Q, and the frequency dependent parameters of the HHP $\mu_{\sigma_\mathrm{S}^2}$, $\phi_{\sigma_\mathrm{S}^2}$. To conclude this section, the different steps that are necessary to obtain the final clean speech estimate $\widehat{S}$ are compactly summarized by a few lines of pseudo-code in Algorithm 1.

VI. Input-Output Characteristic

In this section, we compare the proposed nonlinear estimator to the alternative linear approach [14] in terms of their input-output characteristic (IOCs) [23], which are presented in Figure 5, and show how the prior information encoded in $p\big(\sigma_\mathrm{S}^2\big)$ can benefit the clean speech estimation. The IOC of an estimator presents the amplitude of the clean speech estimate $\widehat{S}$ as a function of the respective noisy input amplitude $|Y|$. The lower the curve, the more suppression is applied by the respective estimator. To make the analysis less dependent on an absolute scaling, the input and the output are both normalized by σ_V. The noise PSD is $\sigma_\mathrm{V}^2 = 1$, the speech PSD estimate is $\widehat{\sigma_\mathrm{S}^2} = 1$, and $Q = 10$ in (11). The difference between the three plots in Figure 5 lies only in the choice of $p\big(\sigma_\mathrm{S}^2\big)$, i.e. the available prior knowledge about the true speech PSD σ_S^2. For all three plots, the true speech PSD σ_S^2 is assumed to follow a log-normal distribution (12) with a standard deviation of $\phi_{\sigma_\mathrm{S}^2} = 3$ dB. Only its mean $\mu_{\sigma_\mathrm{S}^2}$ differs from plot to plot, i.e how much the PSD estimate $\widehat{\sigma_\mathrm{S}^2}$ deviates from the expected value of the true PSD.

In Figure 5, the IOCs of the proposed approach are presented together with the linear estimator proposed in [14]. As a reference we present the IOC of the conventional Wiener filter, denoted by "Wiener", which relies only on the PSD estimate $\widehat{\sigma_\mathrm{S}^2}$ and completely neglects the prior information about the true PSD σ_S^2. To allow for a sensible comparison, we set $f(S) = S$ in (3) for our approach, such that all algorithms are MMSE estimators of S. Hence, the proposed estimator is implemented according to (14). Preliminary analyses showed that the general effects of incorporating PSD uncertainty that we present for the complex estimator similarly apply to the amplitude and log-amplitude estimators in (15) and (16).

At the top of Figure 5, we have $\widehat{\sigma_{\mathrm{S,dB}}^2} = \mu_{\sigma_\mathrm{S}^2}$, i.e. in the dB-domain the PSD estimate is exactly the expected value of the true PSD. Since the prior information and the PSD estimate coincide, the influence on the IOCs is small and [14] closely follows the conventional Wiener filter. We can also see that the Wiener filter and [14] are linear estimators, while the proposed approach is nonlinear and applies less suppression to large inputs. Note that this is a typical behavior also known for estimators based on super-Gaussian speech priors. This nonlinear IOC is known to better protect speech at the cost of a slightly increased tendency to produce musical noise compared to their Gaussian counterparts. In the middle, the PSD estimate is 10 dB lower than $\mu_{\sigma_\mathrm{S}^2}$, 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. This is taken into account by the uncertainty-aware approaches, which apply less suppression than the conventional Wiener filter "Wiener" that solely relies on the PSD estimate. Both estimators trade-off the PSD estimate and the prior information about $p\big(\sigma_\mathrm{S}^2\big)$ according to the statistical models they are based on as detailed in Section IV-D.

Finally, at the bottom of Figure 5, $\mu_{\sigma_\mathrm{S}^2}$ is 10 dB lower than the actual estimate. Analogously to the previous discussion, the uncertainty aware estimators now apply more attenuation, since it is less likely that relevant speech energy is present in the input Y than suggested by the estimate $\widehat{\sigma_\mathrm{S}^2}$.

VII. Evaluation

The algorithms are evaluated on 128 sentences taken from the test set of the TIMIT [24] database, degraded by pink noise, white noise modulated with a frequency of 0.5 Hz, traffic noise, factory noise, and babble noise at various SNRs, of which the last two noise types are taken from [25]. This totals 3200 files that are used for the evaluation. 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. We evaluate noise reduction (NR) and segmental speech SNR (SSNR) [26] to separately assess the amount of noise reduction and speech distortions that are introduced, which indicate how aggressive a clean speech estimator is. Finally, we also present wideband Perceptual Evaluation of Speech Quality (WB-PESQ) [27] scores, which have been shown to correlate with the overall quality of spectrally enhanced speech [28]. For an improved visualization we do not present absolute WB-PESQ values, but the improvement over the unprocessed noisy signal. The linear estimator under PSD uncertainty in (6) [14] and the proposed nonlinear estimator for $f(S) = S$ (14) are compared to the conventional Wiener filter. We choose $f(S) = S$ to allow for a fair comparison of the three approaches, which then are all estimators of the complex coefficients S.

To allow for a more general investigation of the effects of incorporating PSD estimation uncertainty, the algorithms are evaluated for two different speech PSD estimators of different

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2816241, IEEE/ACM Transactions on Audio, Speech, and Language Processing

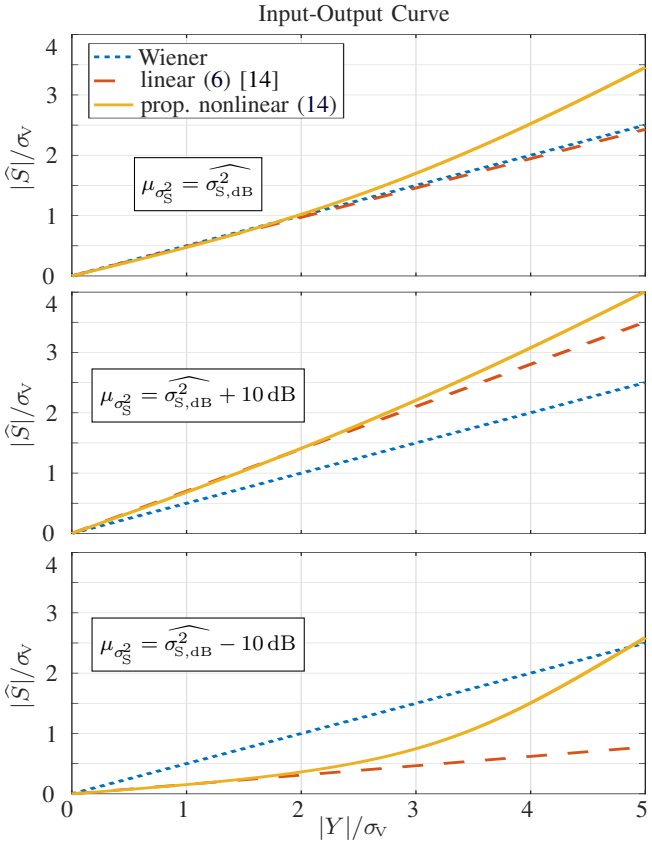

Figure 5. Input-output characteristics for $\sigma_V^2 = 1$, $\widehat{\sigma_S^2} = 1$, $Q = 10$. The three plots vary only in the prior information about the true PSD, i.e. $p(\sigma_S^2)$ (12). The mean of the true $\sigma^2_{S,dB}$ is $\mu_{\sigma_S^2} = \widehat{\sigma^2_{S,dB}}$ (top), $\mu_{\sigma_S^2} = \widehat{\sigma^2_{S,dB}} + 10$ dB (middle), and $\mu_{\sigma_S^2} = \widehat{\sigma^2_{S,dB}} - 10$ dB (bottom), while its standard deviation is always $\phi_{\sigma_S^2} = 3$ dB.

quality. First, the strongly fluctuating instantaneous PSD estimates are simply averaged over neighboring segments directly in the spectral domain. Secondly, we apply elaborate TCS [6], [15], which is known to provide estimates that allow for high quality speech estimation. Finally, the proposed estimator under PSD uncertainty is compared to the conventional Wiener filter by means of a pairwise preference listening test.

A. Moving average speech PSD estimation

In this section, the speech PSD estimates are obtained via a moving average of instantaneous PSD estimates directly in the spectral domain over $Q = 5$ neighboring segments, which corresponds to a time window of 96 ms. The results in Figure 6, which have been averaged over all five noise types, reveal that both, the linear and the proposed nonlinear speech estimator under PSD uncertainty outperform the conventional Wiener filter in terms of WB-PESQ. The benefits are most pronounced at low SNRs, where an improvement of more than 0.2 MOS in WB-PESQ is achieved. Indeed, informal listening reveals that for this simple PSD estimator, the Wiener filter shows strong and annoying musical noise. This is substantially reduced by considering the PSD uncertainty, which is also reflected in the higher NR. The reason for this is that in low SNR regions the speech PSD estimation is especially challenging and the uncertainty-aware approaches therefore rely more on the prior information in form of the HHP $p(\sigma_S^2)$ rather than the fluctuating estimates, as shown in the left panel of Figure 4.

The performance of the two uncertainty-aware approaches is rather similar in terms of WB-PESQ for this simple PSD estimator. However, judging from informal listening, the proposed nonlinear estimator better preserves the speech component while being somewhat more prone to producing musical noise. This trade-off is characteristic for nonlinear, e.g. super-Gaussian, estimators, see e.g. [29].

B. TCS-based speech PSD estimation

Here we employ the more elaborate TCS, which greatly reduces random fluctuations in the PSD estimates while avoiding undesired temporal smearing of speech. Accordingly, the conventional Wiener filter performs consistently better in terms of signal quality with substantially less musical noise as compared to using the PSD estimates from the previous section. This is also reflected in the higher WB-PESQ scores and increased NR in Figure 7.

With the improved PSD estimates, the linear estimator [14] does not improve over the Wiener filter anymore. Similar to using the simple PSD estimator of the previous section, [14] is more aggressive than the Wiener filter. However, now that the Wiener filter itself produces far less musical noise, the higher noise reduction does not outweigh the reduced SSNR anymore. In contrast, the proposed nonlinear estimator still yields improvements in the predicted speech quality over the Wiener filter. It better protects speech components, indicated by the higher SSNR at the cost of only a slight decrease in noise reduction. The relative improvement in WB-PESQ over the conventional Wiener filter is however smaller than for the simple PSD estimator in Figure 6. Intuitively, the more reliable the PSD estimation, the smaller the detrimental effect of neglecting its uncertainty. Consequently, the potential benefit of uncertainty-aware approaches is larger when only poor PSD estimates are available.

Nevertheless, even for elaborate PSD estimators, the overall performance can still be improved by more accurate prior information in terms of the HHP $p(\sigma_S^2)$. Currently, $p(\sigma_S^2)$ is trained offline and fixed, but it could potentially also be trained separately for different phonemes instead. This would open up new and interesting possibilities to combine the generality of traditional Bayesian clean speech estimation with the power of modern machine learning based approaches for phoneme or speech recognition.

C. Listening Experiment

To verify the predictions made by the instrumental measures, we conducted a listening experiment. 13 self-reported normal hearing listeners participated in a pairwise preference test, where the proposed nonlinear estimator under PSD uncertainty (14) was compared to its conventional counterpart, the Wiener filter. Eight utterances from the evaluation set of the previous section were presented to each listener via headphones in a quite office room, 4 at an SNR of -5 dB and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2816241, IEEE/ACM Transactions on Audio, Speech, and Language Processing

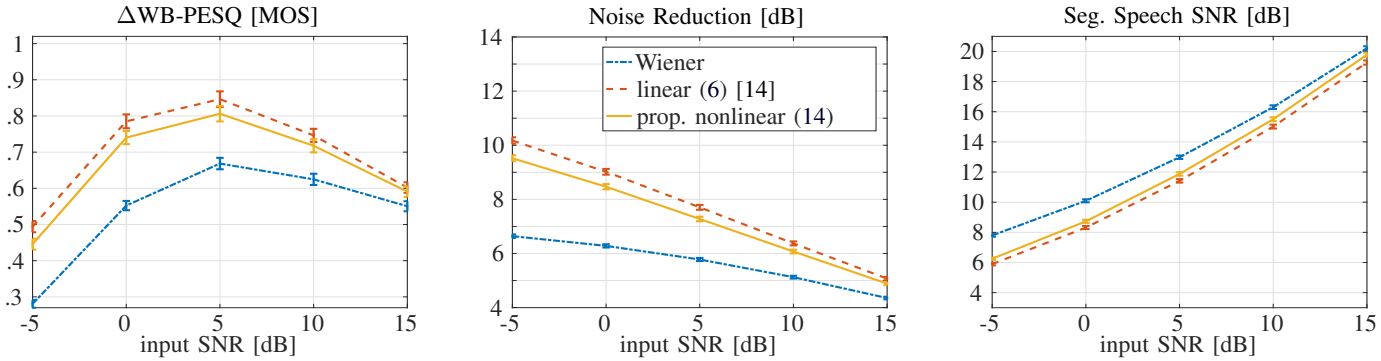

Figure 6. Average WB-PESQ improvement, noise reduction (NR), and segmental speech SNR (SSNR) together with the respective 95 % confidence intervals. The speech PSD estimate has been obtained via a moving average smoothing in the spectral domain.

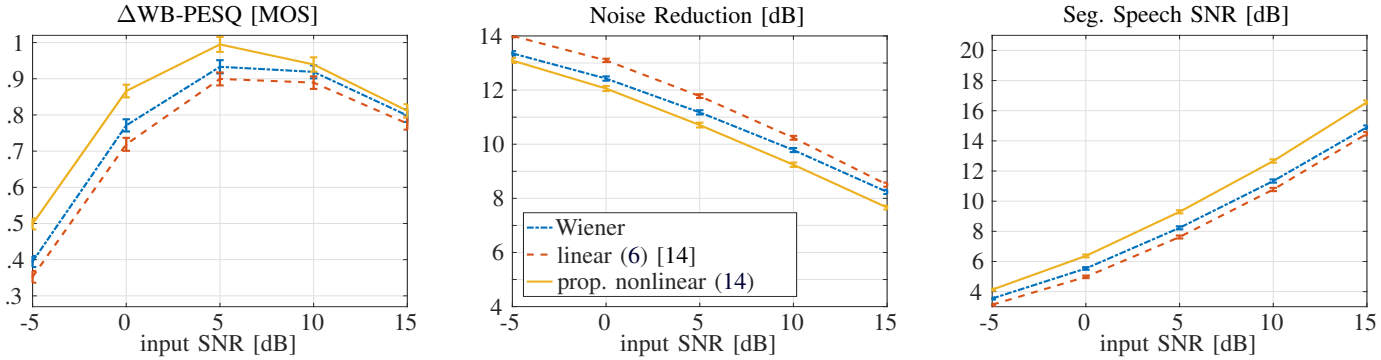

Figure 7. As Figure 6, but with the speech PSD estimate obtained via temporal cepstrum smoothing (TCS).

4 at an SNR of 5 dB. The speech PSD was again estimated via TCS. To isolate the effects of speech PSD modeling from those of suboptimal noise PSD estimation, here we used stationary pink noise with a known PSD. For each example, the listeners were asked to judge which of the two presented signals A) offers the higher speech quality, B) contains less noise, C) offers the higher noise quality, and D) they overall prefer (OP). Each category was rated separately, accumulating to 32 pairwise comparisons per listener. The results are presented in Figure 8 in terms of the average preference for the proposed estimator (14). The higher the value, the more often the proposed estimator under PSD uncertainty was preferred over the Wiener filter. The noise reduction and noise quality were rated similarly, with a preference for the Wiener filter at -5 dB, which, however, is not statistically significant ($p > 0.05$) according to a two-sided binomial test. At the same time, the majority of listeners preferred the proposed estimator in terms of speech quality and overall, with a statistically significant preference for both at -5 dB SNR and for speech quality at 5 dB SNR. The results are in line with the instrumental measures in Figure 7 and underline the potential of incorporating speech PSD uncertainty into speech enhancement.

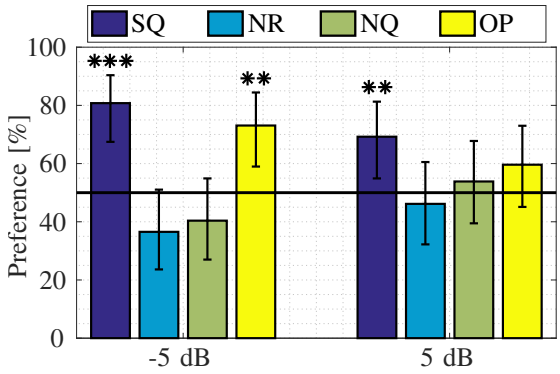

Figure 8. Results of the listening experiment. Averaged preference for the proposed nonlinear estimator under PSD uncertainty over the Wiener filter in terms of speech quality (SQ), noise reduction (NR), noise quality (NQ), and overall (OP) together with the 95 % confidence intervals. The asterisks mark statistically significant preferences, with ∗∗ for $p < .01$ and ∗∗∗ for $p < .001$.

VIII. Conclusions

Speech PSD estimation is both an important and an error-prone part of speech enhancement algorithms. In this paper, we showed that incorporating the uncertainty of speech PSD estimates may increase the robustness of clean speech estimators. While for the estimator in [14] it has been assumed that the noisy input Y does not yield additional information on σ_S^2 when the ML estimate $\widehat{\sigma_S^2}$ is given, we showed that avoiding this assumption yields a fundamentally different estimator which is *nonlinear* with respect to the noisy input. Furthermore, the new estimator provides counterparts to several well-known clean speech estimators such as the Wiener filter and Ephraim and Malah's amplitude estimators under PSD uncertainty. Finally, we showed how a modern PSD estimator based on temporal cepstrum smoothing can be integrated into the PSD uncertainty framework, which improved its overall performance substantially.

REFERENCES

[1] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2604–2612, May 2016.

[2] M. Kolbæk, Z. H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[3] A. Chinaev, J. Heymann, L. Drude, and R. Haeb-Umbach, "Noise-presence-probability-based noise PSD estimation by using DNNs," in *Speech Communication; 12. ITG Symposium*, Paderborn, Germany, Oct. 2016.

[4] R. Rehr and T. Gerkmann, "On the importance of super-Gaussian speech priors for machine-learning based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 357–366, Feb. 2018.

[5] 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. 32, no. 6, pp. 1109–1121, Dec. 1984.

[6] C. Breithaupt, T. Gerkmann, and R. 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.

[7] M. Djendi and P. Scalart, "Reducing over- and under-estimation of the a priori SNR in speech enhancement techniques," *Digital Signal Processing*, vol. 32, pp. 124 – 136, Sep. 2014.

[8] T. Gerkmann and R. 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.

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

[10] M. Krawczyk-Becker and T. Gerkmann, "Nonlinear speech enhancement under speech PSD uncertainty," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018.

[11] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 1, pp. 99–102, 1974.

[12] 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.

[13] R. C. Hendriks and R. Martin, "MAP estimators for speech enhancement under Normal and Rayleigh inverse Gaussian distributions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 918–927, Mar. 2007.

[14] 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.

[15] T. Gerkmann and R. 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.

[16] O. E. Barndorff-Nielsen, "Normal inverse Gaussian distributions and stochastic volatility modelling," *Scand. J. Statist.*, vol. 24, no. 1, pp. 1–13, Mar. 1997.

[17] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook*, 3rd ed., R. C. Dorf, Ed. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group, 2006.

[18] 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.

[19] R. Martin and T. Lotter, "Optimal recursive smoothing of non-stationary periodograms," in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Darmstadt, Germany, Sep. 2001, pp. 167–170.

[20] 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.

[21] Y. Ephraim and D. 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.

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

[23] J. E. Porter and S. 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.

[24] 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," Gaithersburg, MD, USA, 1993.

[25] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *ELSEVIER Speech Commun.*, vol. 12, pp. 247–251, Jul. 1993.

[26] 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.

[27] ITU-T, "Perceptual evaluation of speech quality (PESQ)," *ITU-T Recommendation P.862*, 2001.

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

[29] 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.

Martin Krawczyk-Becker (S'15—M'17) received the Dipl.-Ing. degree in electrical engineering and information technology from the Ruhr-Universität Bochum, Bochum, Germany in 2011. From January 2010 to July 2010, he was with Siemens Corporate Research, Princeton, NJ, USA. In 2016 he received his Dr.-Ing. degree from the Universität Oldenburg, Oldenburg, Germany. Martin Krawczyk-Becker is currently a postdoctoral researcher at the University of Hamburg, Germany.

Timo Gerkmann (S'08—M'10—SM'15) studied Electrical Engineering and Information Sciences at the universities of Bremen and Bochum, Germany. He received his Dipl.-Ing. degree in 2004 and his Dr.-Ing. degree in 2010 both in Electrical Engineering and Information Sciences from the Ruhr-Universität Bochum, Bochum, Germany. In 2005, he spent six months with Siemens Corporate Research in Princeton, NJ, USA. During 2010 to 2011 Dr. Gerkmann was a postdoctoral researcher at the Sound and Image Processing Lab at the Royal Institute of Technology (KTH), Stockholm, Sweden. From 2011 to 2015 he was a professor for Speech Signal Processing at the Universität Oldenburg, Oldenburg, Germany. During 2015 to 2016 he was a Principal Scientist for Audio & Acoustics at Technicolor Research & Innovation in Hanover, Germany. Since 2016 he is a professor for Signal Processing at the University of Hamburg, Germany. His research interests are on signal processing algorithms for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human-machine interfaces. Timo Gerkmann is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing.