ROBUST MMSE FILTERING FOR SINGLE-MICROPHONE SPEECH ENHANCEMENT

Gerald Enzner and Philipp Thüne

Adaptive Systems Laboratory, Ruhr-Universität Bochum, D-44780 Bochum, Germany
Email: {gerald.enzner,philipp.thuene}@rub.de

ABSTRACT

MMSE filtering of signals contaminated with additive noise is addressed with explicit uncertainty of the second-order target signal statistics. The unfortunate lack of stationarity of speech, and hence the phenomenon of musical noise in speech enhancement, is an ideal problem for the proposed approach. Specifically, we complement the established short-time power-spectral subtraction for speech power estimation with a prior of the momentary speech-power level. The MMSE estimator for Gaussian speech amplitudes is then derived under these circumstances. The potential for the enhancement of noisy speech is briefly demonstrated by SNR and PESQ analysis.

Index Terms— optimum filtering, speech enhancement

1. INTRODUCTION & RELATION TO PRIOR WORK

The status of optimum filtering of noisy signals has come a long way from the celebrated Wiener filter, which is the linear minimum mean-square error (MMSE) filter [1, 2, 3]. It relies on second-order statistical *a priori* information, i.e., the power spectral densities (PSDs) of target and noise signals. More general insight into MMSE filtering is provided in concurrent statistical signal processing literature, e.g., [4, 5], which yields the Wiener filter as a "nonlinear" MMSE estimator based on Gaussian random processes.

With respect to the field of speech enhancement, alternative cost functions and spectral weighting rules were deployed in form of the short-time spectral amplitude estimator (MMSE-STSA) [6], the log-spectral amplitude estimator (MMSE-LSA) [7], the psychoacoustically-motivated speech enhancement rules [8, 9], the more recent family of Bayesian extensions of previous techniques [10, 11, 12], and eventually the super-Gaussian speech modeling approach [13, 14] in MMSE filtering.

Another dimension in speech enhancement, which is largely independent of the optimum filtering rule under consideration, is to support the construction of the filter with an estimate of the noise PSD. This includes estimators based on voice-activity detection (VAD) [15, 16], minimum statistics [17], minimum-controlled recursive averaging [18], or speech presence probability [19]. We shall for the remainder of this paper assume that one of these methods can provide us with the noise PSD of sufficient accuracy.

The simplest way of qualifying the dimension of our paper is to see it as a fresh alternative for achieving the suppression of the "musical noise" phenomenon. The latter is essentially due to the inaccurate short-time estimation of the time-varying and frequency-dependent speech PSD from a noisy speech signal. Other authors have approached this important dimension by local averaging of spectral magnitudes [20], noise over-subtraction and spectral flooring [21], or additional cepstral smoothing [22]. We do, instead, frame the ubiquitous short-time spectral subtraction technique [20] into a Bayesian framework in order to unite both the optimum filtering and the musical-noise suppression in the MMSE sense.

2. CLASSICAL MMSE-FILTERING OF NOISY SPEECH

2.1. Signal Model and Signal Enhancement Objective

Fig. 1 depicts the top-level architecture of the optimum filtering problem. Speech $s(k)$ is recorded or observed in the presence of additive noise $n(k)$. Consecutive segments of the noisy signal $y(k)$ at discrete time k are translated into the short-time Fourier domain via discrete Fourier transform (DFT). Here it is assumed that the time-frequency domain spectrum $Y(\Omega_\mu,\kappa)$ is decorrelated both along discrete frame-time κ and discrete frequency Ω_μ. While some authors have recently considered and successfully exploited a residual correlation of the time-frequency representation of noisy signals across multiple time-frequency bins [23, 24, 25], we will here carry on with the conventional assumption of decorrelation. The noisy DFT-domain representation $Y(\Omega_\mu,\kappa) = S(\Omega_\mu,\kappa) + N(\Omega_\mu,\kappa)$ is then weighted by individual weights $H(\Omega_\mu,\kappa)$, such that optimal reconstructions $\widehat{S}(\Omega_\mu,\kappa)$ and $\widehat{s}(k)$ of the input speech are obtained.

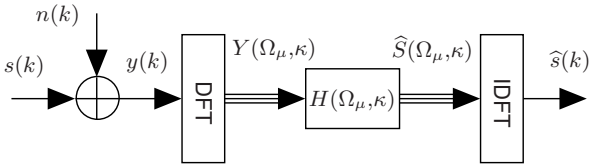

Fig. 1. Additive noise model and spectral enhancement workflow.

2.2. Minimum Mean-Square Error Estimation

In most of the cases in speech enhancement, the signals $s(k)$ and $n(k)$ are initially modeled as stationary random processes with power spectral densities $\Phi_s(\Omega_\mu)$ and $\Phi_n(\Omega_\mu)$, respectively [5]. The actual enhancement by $H(\Omega_\mu,\kappa)$, however, takes place "instantaneously" (i.e., per time-frequency bin) with possibly independent local speech and noise statistics. The most popular optimization criterion then is the minimization of the mean-square error between, say, the desired output $S(\Omega_\mu,\kappa)$ in the frequency domain and the actual output $\widehat{S}(\Omega_\mu,\kappa)$, i.e.,

$$J_{s,y} = \int\int |S - \widehat{S}\,|^2\, p(S,Y)\,\mathrm{d}S\mathrm{d}Y \qquad (1)$$

where time and frequency indices were dropped to denote random variables rather than signals. The concept of the mean square-error $J_{s,y}$ is the averaging of the square error $|S-\widehat{S}|^2$ across all possible random events S and Y with joint probability density $p(S,Y)$. This conventional concept is spelled out so precisely here, because it will be subject to generalization below. Right now, we further declare the status quo in terms of the factorization of $p(S,Y) = p(S|Y)\,p(Y)$

via the conditional probability $p(S|Y)$ and the evidence $p(Y)$, such that a minimization of the inner integral

$$J_s = \int \left| S - \widehat{S} \right|^2 p(S|Y)\,\mathrm{d}S\ . \tag{2}$$

achieves the global minimization of $J_{s,y} = \int J_s\, p(Y)\,\mathrm{d}Y$. In (2), Y thus assumes the particular qualification of an arbitrary and given observation $Y = Y(\Omega_\mu,\kappa)$ and is effectively treated as a deterministic parameter. Continuing the derivation from (2), the complex "Wirtinger" derivative [26, 3] with respect to $\widehat{S}$ is formed and equated to zero, i.e., $\partial J_s/\partial\widehat{S}^* = 0$, obtaining

$$\widehat{S}_s \;=\; \int S\, p(S|Y)\,\mathrm{d}S \;=\; \mathcal{E}_{p(s|y)}\{S|Y\} \tag{3}$$

as the well known representation of MMSE estimation in terms of the posterior mean of S, given a particular Y. With an assumption of joint Gaussianity of $S(\Omega_\mu,\kappa)$ and $N(\Omega_\mu,\kappa)$, the expectation can be resolved into the celebrated Wiener filter [5] in its anti-causal frequency-domain representation [2],

$$\widehat{S}_s \;=\; H_w\, Y \tag{4}$$

and herein

$$H_w(\Phi_s) \;=\; \frac{\Phi_s}{\Phi_y} \;=\; \frac{\Phi_s}{\Phi_s+\Phi_n} \;=\; \frac{\mathrm{SNR}}{1+\mathrm{SNR}}\ , \tag{5}$$

where $\Phi_s = \mathcal{E}\{|S(\Omega_\mu,\kappa)|^2\}$, $\Phi_n = \mathcal{E}\{|N(\Omega_\mu,\kappa)|^2\}$, and $\mathrm{SNR} = \Phi_s/\Phi_n$ are possibly time- and frequency-dependent.

For the remainder of this paper, we shall assume the availability of a good estimator of the stationary or slowly time-varying noise PSD Φ_n, thus not facing relevant uncertainty of that noise PSD. The straightforward and yet typical way to resolve the uncertainty of the speech PSD, i.e., on the basis of the available data, is the "spectral subtraction". Based on the additive relationship $\Phi_y = \Phi_s + \Phi_n$ of speech and noise PSD, an ad-hoc speech PSD is determined via short-time estimation of the noisy-speech PSD Φ_y, i.e.,

$$\begin{aligned}\widehat{\Phi}_s(\Omega_\mu,\kappa) \;&=\; \widehat{\Phi}_y(\Omega_\mu,\kappa) - \Phi_n(\Omega_\mu)\\ &=\; \frac{1}{Q}\sum_{i=\kappa-Q+1}^{\kappa} |Y(\Omega_\mu,i)|^2 - \Phi_n(\Omega_\mu)\ ,\end{aligned} \tag{6}$$

where Q describes the period of short-time stationarity. Further utilization of the short-time estimation $\widehat{\Phi}_s(\Omega_\mu,\kappa)$ in the Wiener filter (5) partly yields the desired noise reduction, but also leaves considerable "musical noise" in the output $\widehat{s}(k)$. Since the imperfection of the short-time spectral subtraction is responsible for the musical noise phenomena, several authors aimed to refine the local speech-level or SNR estimation by enhanced local averaging [6, 27, 28, 29] or deeper modeling of speech statistics [30, 31]. In what follows we will pursue a different perspective in that we leave the ad-hoc speech PSD estimator (6) and its inaccuracy to what it is, yet asking the precise MMSE estimator $H(\Omega_\mu,\kappa)$ under these circumstances.

3. ROBUST MMSE FILTERING APPROACH

We will work along the structure of the previous MMSE "status quo" section to extend the methodologies with explicit modeling of the uncertainty of the speech second-order statistics $\Phi_s(\Omega_\mu,\kappa)$.

3.1. Extended Minimum Mean-Square Error Estimation

Looking back at (1), we now form a generalized mean-square error cost function by averaging $|S-\widehat{S}|^2$ across all possible random events S, Y, Φ_s, and $\widehat{\Phi}_s$ with joint probability density $p(S,Y,\Phi_s,\widehat{\Phi}_s)$, i.e.,

$$J_{s,y,\Phi_s,\widehat{\Phi}_s} = \iiiint \left|S-\widehat{S}\right|^2 p(S,Y,\Phi_s,\widehat{\Phi}_s)\,\mathrm{d}S\mathrm{d}Y\mathrm{d}\Phi_s\mathrm{d}\widehat{\Phi}_s. \tag{7}$$

In this way, we imply uncertainty of the actual speech PSD Φ_s, while taking the availability of our ad-hoc estimate $\widehat{\Phi}_s$ in (6) into account in the form of a noisy representation of Φ_s. Naturally this extension will then require statistical modeling of Φ_s and $\widehat{\Phi}_s$ in order to evaluate the integral. We shall come to this aspect in Sec. 4.

From (7), we can proceed along the principles of the previous section by factorizing the joint density $p(S,Y,\Phi_s,\widehat{\Phi}_s) = p(S,\Phi_s|Y,\widehat{\Phi}_s)\,p(Y,\widehat{\Phi}_s)$ via the posterior distribution of the unobservable quantities S and Φ_s and the evidence distribution of available quantities Y and $\widehat{\Phi}_s$. Minimization of the inner integral

$$J_{s,\Phi_s} = \iint \left|S-\widehat{S}\right|^2 p(S,\Phi_s|Y,\widehat{\Phi}_s)\,\mathrm{d}S\mathrm{d}\Phi_s \tag{8}$$

will then again achieve global minimization of the cost $J_{s,y,\Phi_s,\widehat{\Phi}_s} = \iint J_{s,\Phi_s}\, p(Y,\widehat{\Phi}_s)\,\mathrm{d}Y\mathrm{d}\widehat{\Phi}_s$. In (8), both Y and $\widehat{\Phi}_s$ now qualify as our arbitrary and given observations $Y = Y(\Omega_\mu,\kappa)$ and $\widehat{\Phi}_s = \widehat{\Phi}_s(\Omega_\mu,\kappa)$, respectively, and are thus treated in what follows as deterministic input data. In the next step, the complex derivative with respect to the sought quantity $\widehat{S}$ is again formed and equated to zero, i.e., $\partial J_{s,\Phi_s}/\partial\widehat{S}^* = 0$, thus obtaining a counterpart of (3),

$$\begin{aligned}\widehat{S}_{s,\Phi_s} \;&=\; \iint S\, p(S,\Phi_s|Y,\widehat{\Phi}_s)\,\mathrm{d}S\,\mathrm{d}\Phi_s\\ &=\; \mathcal{E}_{p(s,\Phi_s|y,\widehat{\Phi}_s)}\{S|Y,\widehat{\Phi}_s\}\ ,\end{aligned} \tag{9}$$

which is recognized as the conditional-mean estimate of the speech S across the joint probability of S and Φ_s, given Y and $\widehat{\Phi}_s$.

We continue from (9) by simply factorizing the joint probability $p(S,\Phi_s|Y,\widehat{\Phi}_s) = p(S|Y,\Phi_s,\widehat{\Phi}_s)\,p(\Phi_s|Y,\widehat{\Phi}_s)$ via Bayes' rule. Then $p(S|Y,\Phi_s,\widehat{\Phi}_s) = p(S|Y,\Phi_s)$, since $\widehat{\Phi}_s$ bears additional unrelated values $Y = Y(\Omega_\mu,i)$, $i<\kappa$, while $p(\Phi_s|Y,\widehat{\Phi}_s) = p(\Phi_s|\widehat{\Phi}_s)$, since $Y = Y(\Omega_\mu,\kappa)$ is already exploited in $\widehat{\Phi}_s$. Hence

$$\begin{aligned}\widehat{S}_{s,\Phi_s} \;&=\; \iint S\, p(S|Y,\Phi_s)\,\mathrm{d}S\, p(\Phi_s|\widehat{\Phi}_s)\,\mathrm{d}\Phi_s\\ &=\; \mathcal{E}_{p(\Phi_s|\widehat{\Phi}_s)}\big\{\,\mathcal{E}_{p(s|y,\Phi_s)}\{S|Y,\Phi_s\}\;\big|\;\widehat{\Phi}_s\,\big\}\\ &=\; \mathcal{E}_{p(\Phi_s|\widehat{\Phi}_s)}\big\{\,\widehat{S}_s\;\big|\;\widehat{\Phi}_s\,\big\}\end{aligned} \tag{10}$$

and we recognize the inner expectation in (10) exactly as our former posterior-mean estimate (3), now subject to variation and further expectation in the speech PSD Φ_s. Still considering jointly Gaussian complex speech and noise amplitudes, we thus recycle the Wiener filter $H_w = H_w(\Phi_s)$ as a function of the random variable Φ_s,

$$\begin{aligned}\widehat{S}_{s,\Phi_s} \;&=\; \mathcal{E}_{p(\Phi_s|\widehat{\Phi}_s)}\big\{\,H_w\,Y\;\big|\;\widehat{\Phi}_s\,\big\}\\ &=\; Y\;\mathcal{E}_{p(\Phi_s|\widehat{\Phi}_s)}\big\{\,H_w\;\big|\;\widehat{\Phi}_s\,\big\}\\ &=\; Y\int H_w(\Phi_s)\,p(\Phi_s|\widehat{\Phi}_s)\,\mathrm{d}\Phi_s\\ &=\; Y\widehat{H}_{\Phi_s}\ ,\end{aligned} \tag{11}$$

such that the observation Y can be successfully isolated from the integration. The remaining integral for $\widehat{H}_{\Phi_s}$ obviously takes the form of a mean-square error estimator of the function $H(\Phi_s)$ of the random variable Φ_s, given a particular value $\widehat{\Phi}_s$ of our ad-hoc estimator of the speech PSD. This result nicely dictates the need to determine in the next step of our derivation the posterior distribution $p(\Phi_s|\widehat{\Phi}_s)$ of the actual speech PSD according to Bayes' rule,

$$p(\Phi_s|\widehat{\Phi}_s) = \frac{p(\widehat{\Phi}_s|\Phi_s)\,p(\Phi_s)}{\int p(\widehat{\Phi}_s|\Phi_s)\,p(\Phi_s)\,\mathrm{d}\Phi_s}\,, \tag{12}$$

which in turn requires statistical modeling of the relationship of Φ_s and $\widehat{\Phi}_s$ via the likelihood $p(\widehat{\Phi}_s|\Phi_s)$ and a prior $p(\Phi_s)$.

4. BAYESIAN APPROACH TO THE SPEECH PSD

We shall approach the elements of (12) step by step in order to prepare the evaluation of the robust MMSE estimator in (11). We hence enter a statistical modeling perspective one layer below the conventional statistical modeling of speech and noise amplitudes.

4.1. Likelihood $p(\widehat{\Phi}_s|\Phi_s)$

Looking back at definition (6), the probability density of the sum-of-squares, or more accurately speaking, our mean-of-squares of complex Gaussian random variables $Y(\Omega_\mu,\kappa)$ is χ^2-distributed [5],

$$p(\widehat{\Phi}_y) = \frac{\widehat{\Phi}_y^{Q-1}\exp\Big(-Q\frac{\widehat{\Phi}_y}{\Phi_y}\Big)}{\Big(\frac{\Phi_y}{Q}\Big)^Q\Gamma(Q)} \tag{13}$$

when $\widehat{\Phi}_y \geq 0$, and zero otherwise. $\Gamma(n{+}1){=}n!$, $n\in\mathbb{N}$, denotes the discrete Gamma function. Then applying the shift $\widehat{\Phi}_s = \widehat{\Phi}_y - \Phi_n$ as shown by (6), the probability density is also shifted. Thus

$$p(\widehat{\Phi}_s|\Phi_s) = \frac{(\widehat{\Phi}_s+\Phi_n)^{Q-1}\exp\Big(-Q\frac{\widehat{\Phi}_s+\Phi_n}{\Phi_s+\Phi_n}\Big)}{\Big(\frac{\Phi_s+\Phi_n}{Q}\Big)^Q\Gamma(Q)} \tag{14}$$

when $\widehat{\Phi}_s \geq -\Phi_n$. Note $\Phi_n \geq 0$. And we substituted $\Phi_y = \Phi_s + \Phi_n$ to cast the previous expression into the sought likelihood.

4.2. Prior $p(\Phi_s)$

Our likelihood has been formulated rigorously according to the well-defined data observation model. Since a convenient data generation model for the speech spectral power Φ_s is not available, we here conduct a clean speech histogram measurement by

- considering speech data $s(k)$ in the order of 1 minute at 16 kHz sampling from the TIMIT database [32],
- running DFT spectral analysis $S(\Omega_\mu,\kappa)$ on Hamming-windowed segments of $s(k)$, cf. Fig. 1, where each segment is 512 samples with 75% overlap,
- and finally assessing a histogram of the spectral speech level $10\log_{10}(|S(\Omega_\mu,\kappa)|^2)$ across all time and frequency bins, where each spectral amplitude $|S(\Omega_\mu,\kappa)|^2$ is meant to be a particular realization of our random variable Φ_s.

Fig. 2 depicts the resulting speech histogram on the logarithmic scale $\Phi_{s,\mathrm{dB}} = 10\log_{10}(\Phi_s)$. We acknowledge that such a histogram can vary with recording quality, sampling frequency, DFT size or window overlap. Furthermore it should be noted that its placement on the $\Phi_{s,\mathrm{dB}}$-axis is somewhat arbitrary in that it depends on the arithmetic data type and amplitude range of signal $s(k)$. The histogram thus needs to be found individually for the application at hand.

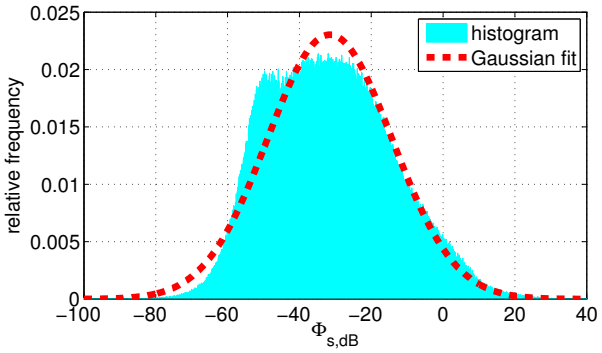

Fig. 2. Spectral speech power histogram for 1 min of 16 kHz data.

At this point, we can however report that the "Gaussian" shape or similar characteristics have also been reproduced with a range of parameters in the vicinity of the typical parameters used for the histogram. In line with other authors [29], we thus recommend to fit this prior speech level to a Gaussian distribution

$$p_{\mathrm{dB}}(\Phi_{s,\mathrm{dB}}) \;=\; \frac{1}{2\pi\sigma^2_{\Phi_{s,\mathrm{dB}}}}\exp\Big(-\frac{(\Phi_{s,\mathrm{dB}}-\mu_{\Phi_{s,\mathrm{dB}}})^2}{2\sigma^2_{\Phi_{s,\mathrm{dB}}}}\Big) \tag{15}$$

with sample mean and variance determined from the histogram. For the sake of numerical compatibility with our likelihood, the "linear" counterpart of (15) is easily devised as

$$p(\Phi_s) \;=\; p_{\mathrm{dB}}(\Phi_{s,\mathrm{dB}})\,\frac{\partial\Phi_{s,\mathrm{dB}}}{\partial\Phi_s} \;=\; \frac{10\;p_{\mathrm{dB}}\big(10\log_{10}(\Phi_s)\big)}{\ln(10)\Phi_s}\,. \tag{16}$$

4.3. Speech Posterior and Robust MMSE Filter

The transformation of the log-normal distribution in (16) is however not found in the family of conjugate priors of the χ^2-likelihood. As a result, and conventional in Bayesian inference, we thus have to revert to numerical integration to firstly obtain the evidence $p(\widehat{\Phi}_s) = \int p(\widehat{\Phi}_s|\Phi_s)\,p(\Phi_s)\,\mathrm{d}\Phi_s$ for the speech posterior (12) and secondly the optimal weights $\widehat{H}_{\Phi_s}(\widehat{\Phi}_s) = \int H_w(\Phi_s)\,p(\Phi_s|\widehat{\Phi}_s)\,\mathrm{d}\Phi_s$ in (11).

By then explicitly defining an ad-hoc Wiener filter (WF)

$$\widehat{H}_w(\widehat{\Phi}_s) = \widehat{\Phi}_s/(\widehat{\Phi}_s+\Phi_n) \tag{17}$$

in line with (5), this one-to-one relationship of $\widehat{\Phi}_s$ and $\widehat{H}_w$ can be exploited to render the resulting weights $\widehat{H}_{\Phi_s}$ conveniently as a function of the ad-hoc weights $\widehat{H}_w$ in Fig. 3. Here it turns out that our uncertainty-aware (or "robust") MMSE filter $\widehat{H}_{\Phi_s}$ achieves significantly more attenuation than $\widehat{H}_w$. In particular, the noisy weights in the low $\widehat{H}_w$ range are attenuated to a flat and positive threshold. This mechanism will eventually suppress musical noise to a comfortable noise floor. The high and supposedly reliable values of $\widehat{H}_w$ are however preserved asymptotically. This amounts to a very sensible behavior of the robust MMSE estimator $\widehat{H}_{\Phi_s}$ and, hence, its nonlinear transition between low and high values of $\widehat{H}_w$ as well.

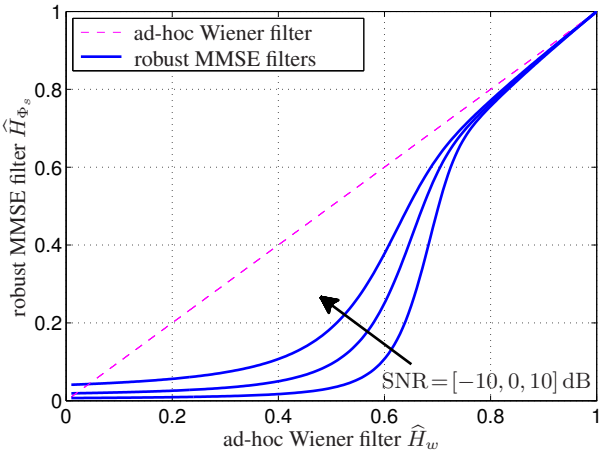

Fig. 3. Robust MMSE filter weights, given ad-hoc Wiener weights.

The actual characteristic $\widehat{H}_{\Phi_s}$ of the proposed filter is eventually controlled by the noise level Φ_n of the likelihood function $p(\widehat{\Phi}_s|\Phi_s)$ and by the prior speech model $p(\Phi_s)$, which is mainly quantified by the hyperparameter $\mu_{\rm dB}$. For fixed relationship of $\mu_{\rm dB}$ and Φ_n, i.e., for fixed global $\mathrm{SNR} = 10\log_{10}(\sigma_s^2/\sigma_n^2)$, where $\sigma_s^2 = \mathcal{E}\{s^2(k)\}$, $\sigma_n^2 = \mathcal{E}\{n^2(k)\}$, and for the speech material at hand, we also have a fixed relationship of $\widehat{H}_{\Phi_s}$ and $\widehat{H}_w$. This property can be confirmed easily from the structure of the likelihood function (14) and extends the same well-known property of the linear Wiener filter. For a given "speech in noise condition", i.e., global SNR, the robust MMSE estimator $\widehat{H}_{\Phi_s}$ is therefore conveniently computed via the ad-hoc weight $\widehat{H}_w$ and a look-up table to comprise the characteristics of Fig. 3.

5. APPLICATION TO SPEECH ENHANCEMENT

We shall eventually study the potential of the proposed filter in terms of the actual speech enhancement that can be achieved. Experiments are however restricted to the demonstration of the effects related to short-time speech reconstruction based on reliable statistical knowledge of the noise power. Hence, we consider a simulation of clean speech in white noise, where the speech is a new utterance from the TIMIT database. Other noise types or the related noise power estimation task is beyond the scope of the paper.

We then assess a time-varying (segmental) SNR at block-time κ,

$$\mathrm{SNR}_{\widehat{s}}(\kappa) \;=\; 10\,\log_{10}\Big(\frac{\sum_{\Omega_\mu}\Phi_s(\Omega_\mu,\kappa)}{\sum_{\Omega_\mu}\Phi_{s-\widehat{s}}(\Omega_\mu,\kappa)}\Big)\,, \qquad (18)$$

in which the required PSDs are computed in sliding windows of the same interval Q as used for spectral estimation (6), i.e.,

$$\Phi_{s-\widehat{s}}(\Omega_\mu,\kappa) \;=\; \frac{1}{Q}\sum_{i=\kappa-Q+1}^{\kappa}|S(\Omega_\mu,i)-\widehat{S}(\Omega_\mu,i)|^2\,, \qquad (19)$$

and $\mathrm{SNR}_y(\kappa)$ of the input signal just employs Y in place of $\widehat{S}$.

Fig. 4 illustrates, besides the SNR_y lower bound, the enhanced $\mathrm{SNR}_{\widehat{s}}$ by a "spectral-subtraction" (ad-hoc) Wiener filter $\widehat{S}_s = \widehat{H}_w Y$, the "robust" MMSE filter $\widehat{S}_{s,\Phi_s} = \widehat{H}_{\Phi_s} Y$, and an "informed" Wiener filter $\widehat{S} = H_w Y$ meant as an upper bound. The spectral-subtraction filter approaches the upper bound during speech presence, but is unfortunately missing optimal noise attenuation and SNR enhancement during speech pause. The robust MMSE filter essentially raises the SNR during speech pause, even approaching the informed Wiener filter, while preserving already high SNR in speech presence.

Tab. 1 then generalizes the previous illustration by presenting single-number results of common instrumental measures for several "speech in noise conditions", i.e., different input SNR. Here it turns out that the robust MMSE filter half-way closes the gap between pure spectral subtraction and the idealized informed Wiener filter, with some advantage of robust MMSE in low-noise conditions.

noisy signal	spec.-sub. WF	robust MMSE	informed WF
-10 dB // 1.50	1.5 dB // 1.80	3.5 dB // 1.90	4.5 dB // 2.60
0 dB // 2.02	8.1 dB // 2.50	8.8 dB // 2.57	9.4 dB // 2.82
10 dB // 2.72	15.0 dB // 3.10	15.3 dB // 3.12	15.6 dB // 3.12

Table 1. Global enhancement in terms of time-averaged segmental SNR [dB] // perceptual evaluation of speech quality (PESQ) [33].

6. CONCLUSIONS

The robust MMSE filter is meant to compensate for the uncertainty related to short-time spectral estimation in noise. The proposed filter then inherently amounts to a rigorous form of signal over-attenuation in mid-SNR ranges and noise flooring in low-SNR ranges, while preserving the already reliable high-SNR domain. In this way, the underlying Bayesian model overcomes more heuristic forms of noise over-estimation and spectral flooring known in spectral subtraction. The filtering can thus achieve improved musical noise suppression.

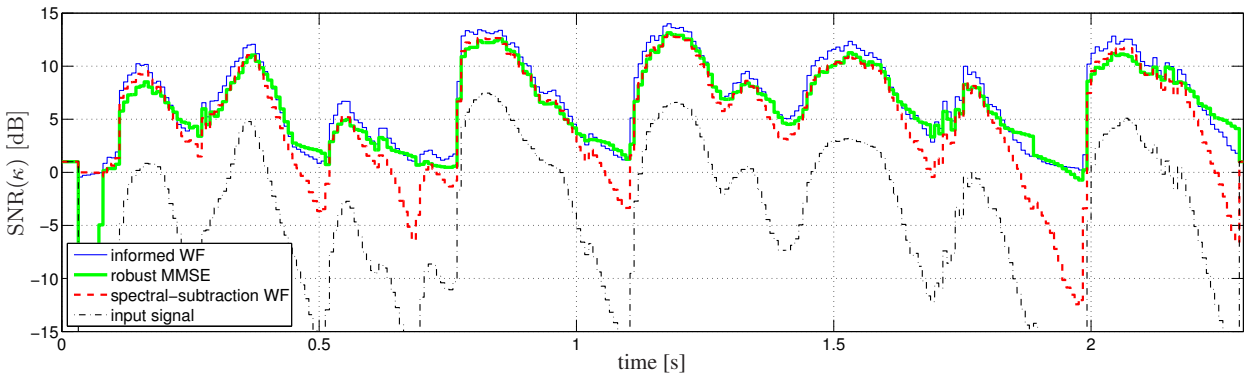

Fig. 4. SNR of noisy speech and various enhanced signals. Global input $\mathrm{SNR} = 0$ dB, sliding window length $Q = 5$, DFT analysis frame-size $M = 512$ samples, Hanning windowed, overlap-add frame-shift $R = 128$ samples, sampling frequency $f_s = 16$ kHz.

7. REFERENCES

[1] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, Wiley, New York NY, 1949.

[2] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, Prentice-Hall, Upper Saddle River, New Jersey, 1996.

[3] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ, 4th edition, 2002.

[4] L. L. Scharf, *Statistical Signal Processing*, Addison-Wesley Publishing Company, 1991.

[5] P. Vary and R. Martin, *Digital Speech Transmission - Enhancement, Coding and Error Concealment*, John Wiley & Sons, Ltd., Chichester, England, 2006.

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

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 33, pp. 443–445, April 1985.

[8] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Audio Proc.*, vol. 10, no. 5, pp. 245–256, Jul 2002.

[9] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Tr. Speech and Audio Process.*, vol. 7, no. 2, pp. 126–137, March 1999.

[10] P. Wolfe and S. Godsill, "Efficient alternatives to the Ephraim and Malah suppressoin rule for audio enhancement," *EURASIP Jrnl. Adv. Signal Process*, pp. 1043–1051, 2003.

[11] P. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 857–869, Sept. 2005.

[12] M. Parchami, W. Zhu, B. Champagne, and E. Plourde, "Bayesian STSA estimation using masking properties and generalized Gamma priors for speech enhancement," *EURASIP Jrnl. Adv. Signal Process.*, 2003.

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

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

[15] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, April 1980.

[16] ETSI 3GPP TS 126.094, "Universal mobile telecommunications system (UMTS); mandatory speech-codec speech processing functions; AMR speech codec; voice activity detector (VAD)," 2001.

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

[18] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Proc. Lett.*, vol. 9, no. 1, pp. 12–15, January 2002.

[19] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 28, no. 2, pp. 137–145, April 2012.

[20] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. 27, no. 2, pp. 113–120, April 1979.

[21] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Proc.*, vol. 2, no. 2, pp. 345–349, April 1994.

[22] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Proc. Lett.*, vol. 14, no. 12, pp. 1036–1039, Dec. 2007.

[23] T. Esch and P. Vary, "Modified Kalman filter exploiting interframe correlation of speech and noise magnitudes," in *Proc. Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, Sept. 2008.

[24] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[25] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 9, pp. 1355–1365, Sept. 2014.

[26] R. Gunning and H. Rossi, *Analytic Functions of Several Complex Variables*, Prentice-Hall, Englewood Cliffs, 1965.

[27] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 8, pp. 799–807, Nov. 2001.

[28] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," in *IEEE Intl. Conf. Acoustics, Speech and Signal Process.*, April 2009, pp. 4409–4412.

[29] T. Wolf and M. Buck, "Spatial maximum a posteriori postfiltering for arbitrary beamforming," in *IEEE Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, Trento, May 2008, pp. 53–56.

[30] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori SNR estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[31] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "An iterative speech model-based a priori SNR estimator," in *Proc. of Interspeech*, Dresden, Sept. 2015, pp. 1740–45.

[32] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium," 1993.

[33] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.