Improved *a posteriori* Speech Presence Probability Estimation Based on a Likelihood Ratio with Fixed Priors

Timo Gerkmann, Colin Breithaupt, and Rainer Martin, *Senior Member, IEEE*

***Abstract*—In this contribution we present an improved estimator for the speech presence probability at each time-frequency point in the short-time Fourier-transform domain. In contrast to existing approaches this estimator does not rely on an adaptively estimated and thus signal dependent *a priori* signal-to-noise ratio estimate. It therefore decouples the estimation of the speech presence probability from the estimation of the clean speech spectral coefficients in a speech enhancement task. Using both a fixed *a priori* signal-to-noise ratio and a fixed prior probability of speech presence, the proposed *a posteriori* speech presence probability estimator achieves probabilities close to zero for speech absence and probabilities close to one for speech presence. While state-of-the-art speech presence probability estimators use adaptive prior probabilities and signal-to-noise ratio estimates we argue that these quantities should reflect true *a priori* information that shall not depend on the observed signal. We present a detection theoretic framework for determining the fixed *a priori* signal-to-noise ratio. The proposed estimator is conceptually simple and yields a better trade-off between speech distortion and noise leakage than state-of-the-art estimators.**

***Index Terms*—Generalized likelihood ratio, softgain, speech analysis, speech enhancement, speech presence probability (SPP).**

I. Introduction

FOR many short-time Fourier transform (STFT) based speech processing systems an estimator for the speech-presence-probability (SPP) in each time-frequency point is of great interest. For instance in speech enhancement clean-speech estimators are often derived under the assumption that speech is actually present. Since this is neither true in speech pauses nor between the spectral bins of the harmonics of a voiced sound, the SPP should be taken into account [1], [2], [3], [4]. For clean-speech estimators, it is crucial that the SPP estimator does reliably recognize speech presence to avoid spectral distortion of low energy speech components. Most existing SPP estimators are designed in a way that they satisfy this demand, and yield high SPP estimates whenever speech is present. However, SPP estimators like [1], [2], [3], have the drawback that they usually do not yield small values for the SPP at time-frequency points where speech is absent, e.g. between the harmonics of voiced speech or even in speech pauses. The estimator in [4] overcomes this problem by making the *a priori* SPP signal dependent. As a consequence, the resulting *a posteriori* SPP is dominated by this *a priori* SPP. In this contribution we show why these problems arise and present a novel way of implementing the *a posteriori* SPP estimator that yields a better trade-off between speech-distortion and noise leakage than state-of-the-art SPP estimators. Furthermore, SPP estimators that rely on an observation of the noisy periodogram suffer from random fluctuations, since the periodogram, as an estimate of the power spectrum, has a high variance [5]. Therefore, we derive an SPP estimator from smoothed observations to reduce these random fluctuations.

This paper is structured as follows: in the next section we provide a deep insight into the mechanisms of *a posteriori* SPP estimation, and conclude that in an SPP estimator the well known decision-directed approach may not be appropriate for estimating the *a priori* signal-to-noise ratio (SNR). We also provide the theoretical basis for incorporating a smoothed observation into an SPP estimator. In Section III-A we show that a smoothed observation reduces the false-alarm rate and the missed-hit rate when the SPP estimator is interpreted as a detector. In Section III-B we propose to use an optimally derived fixed value for the *a priori* SNR that reflects the SNR that is expected when speech is present. In Section IV we combine SPPs gained by globally and locally smoothed observations, and summarize the overall algorithm and the determination of the parameters. In Section V we discuss the application of SPP to a speech enhancement task. In Section VI we show that the proposed method yields a better trade-off between speech distortion and noise leakage as compared to existing approaches.

II. Generalized Speech Presence Probability Estimator

In this section, we derive the SPP estimator in a generalized form, providing the theoretical background for incorporating smoothed observations. We show why the basic SPP estimators do not yield small SPP estimates at time-frequency points where speech is absent and discuss existing improvements.

We assume an additive mixture of speech, $S(k,l)$, and noise, $N(k,l)$, in the STFT domain. Here, k is the frequency index and l is the frame index. The observed signal is given by $Y(k,l) = S(k,l) + N(k,l)$ under the hypothesis, $\mathcal{H}_1$, that speech is present. Under the hypothesis, $\mathcal{H}_0$, that speech is absent it is given by $Y(k,l) = N(k,l)$. We will omit

The authors are with the Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum, 44780 Bochum, Germany (e-mail: timo.gerkmann@rub.de; colin.breithaupt@rub.de; rainer.martin@rub.de).

the indices (k, l) for notational convenience, whenever it is possible. For the short-time Fourier analysis we use Hann windows with 75% overlap and a window-length of 32 ms. The signals are sampled at $f_\mathrm{s} = 16\,\mathrm{kHz}$. We assume that the noise power, $\sigma_N^2 = \mathrm{E}\{|N|^2\}$, is known. In practice, the noise variance, σ_N^2, can be gained using the minimum statistics [6] or the improved minima controlled recursive averaging [7] noise power estimation approach. Introducing the *a posteriori* signal-to-noise ratio (SNR), $\gamma = \frac{|Y|^2}{\sigma_N^2}$, the probability of speech presence given the observation γ, may be written as [5]:

$$\mathcal{P} = P(\mathcal{H}_1|\gamma) = \frac{\Lambda}{1+\Lambda}\,. \tag{1}$$

The generalized likelihood ratio (GLR), Λ, is defined as the weighted ratio of the likelihood of speech presence and the likelihood of speech absence:

$$\Lambda \;=\; \frac{q \quad p(\gamma|\mathcal{H}_1)}{(1-q)\,p(\gamma|\mathcal{H}_0)}\,, \tag{2}$$

where $q = P(\mathcal{H}_1)$ is the *a priori* probability of speech presence and in principle does not depend on the observation. This term can be used to bias the GLR in favor of either speech presence ($q > 0.5$) or of speech absence ($q < 0.5$). All SPP estimators mentioned in Section I are implicitly or explicitly based on the GLR.

As $\mathcal{P}$ is a function of the random variable γ, $\mathcal{P}$ is also a random variable. When $\mathcal{P}$ is incorporated into a speech enhancement framework, random fluctuations in $\mathcal{P}$ may result in spectral peaks in the enhanced signal that may be perceived as *musical noise* [3]. In order to reduce random fluctuations in $\mathcal{P}$, we calculate the smoothed observation over a time-frequency region in the neighborhood of the time-frequency point under consideration:

$$\overline{\gamma}(k,l) = \frac{1}{N} \sum_{\substack{\kappa \,\in\, \mathbb{K} \\ \lambda \,\in\, \mathbb{L}}} \gamma(\kappa,\lambda)\,. \tag{3}$$

Here, $\mathbb{K}$ is the set of adjacent frequency bins, $\mathbb{L}$ is the set of successive time frames, and $N = |\mathbb{K}| \cdot |\mathbb{L}|$ is the number of spectral bins which are averaged. For speech absence $\gamma(\kappa,\lambda)$ is assumed to be short-time stationary in the time-frequency range $\mathbb{K} \times \mathbb{L}$. Assuming a Gaussian distribution for the STFT coefficients, the resulting values $\overline{\gamma}$ are approximately chi-squared distributed [8]:

$$p\left(\overline{\gamma} \;\middle|\; \mathcal{H}_0\right) = \left(\frac{\overline{r}}{2}\right)^{\frac{\overline{r}}{2}} \frac{\overline{\gamma}^{\frac{\overline{r}}{2}-1}}{\Gamma(\frac{\overline{r}}{2})} \exp\left(-\frac{\overline{\gamma}\,\overline{r}}{2}\right)\,. \tag{4}$$

Their degree of freedom $\overline{r} = 2N c_\mathrm{dof}$ is increased as compared to the unsmoothed case where only one spectral bin is considered, i.e. $\overline{r} = 2$. The correction factor c_dof for the degrees of freedom results from the fact that the different values $\gamma(\kappa,\lambda)$ in (3) are not independent, e.g. due to the overlapping analysis frames and the Hann window. It has to be determined empirically as detailed in Appendix A.

If speech is present in single bins within the $\mathbb{K} \times \mathbb{L}$ spectrogram region, the mean $\overline{\gamma}$ is likely to be larger than $\mathrm{E}\{\overline{\gamma}|\mathcal{H}_0\} = 1$. In order to model speech presence in $\mathbb{K}\times\mathbb{L}$ bins we assume the speech energy to be distributed homogeneously over the $\mathbb{K} \times \mathbb{L}$ spectrogram region. This homogeneously spread speech signal energy in $\mathbb{K} \times \mathbb{L}$ is thus reflected in the *a priori* SNR

$$\overline{\xi}(k,l) = \frac{1}{N} \sum_{\substack{\kappa \,\in\, \mathbb{K} \\ \lambda \,\in\, \mathbb{L}}} \xi(\kappa,\lambda)\,, \tag{5}$$

with $\xi(\kappa,\lambda) = \frac{\mathrm{E}\{|S(\kappa,\lambda)|^2\}}{\sigma_N^2(\kappa,\lambda)}$. Note that in case $N = 1$, $\overline{\xi} = \xi$ and $\overline{\gamma} = \gamma$. These assumptions allow us to compute the likelihood of $\overline{\gamma}$ given speech presence, as

$$p\left(\overline{\gamma} \;\middle|\; \mathcal{H}_1\right) = \left(\frac{\overline{r}}{2\,(1+\overline{\xi})}\right)^{\frac{\overline{r}}{2}} \frac{\overline{\gamma}^{\frac{\overline{r}}{2}-1}}{\Gamma(\frac{\overline{r}}{2})} \exp\left(-\frac{\overline{\gamma}\,\overline{r}}{2\,(1+\overline{\xi})}\right)\,. \tag{6}$$

The GLR thus gives

$$\Lambda(\overline{\gamma}) = \frac{q}{1-q} \cdot \left(\frac{1}{1+\overline{\xi}}\right)^{\frac{\overline{r}}{2}} \exp\left(\frac{\overline{\xi}}{1+\overline{\xi}}\,\frac{\overline{r}}{2}\,\overline{\gamma}\right), \tag{7}$$

which is then used in (1) to compute the *a posteriori* SPP $P(\mathcal{H}_1|\overline{\gamma}) = \mathcal{P}$.

The GLR (2) is the ratio of the likelihoods (6) and (4) weighted by their priors. In order to illustrate the effect of $\mathcal{P}$, Figure 1 shows the numerator and denominator of the GLR (2) and the resulting SPPs for an *a priori* SNR of $\xi = 8\,\mathrm{dB}$ and $\xi = -40\,\mathrm{dB}$ as a function of the *a posteriori* SNR. For this investigation no smoothing of the observations is performed, i.e. $N = 1$ in (3). Note that while all computations are done in the linear domain, for the illustrations the *a posteriori* SNR is converted from linear scale to decibels, as $\gamma[\mathrm{dB}] = 10\log_{10}(\gamma)$. The intersection of the weighted likelihoods $q\,p(\gamma|\mathcal{H}_1)$ and $(1-q)\,p(\gamma|\mathcal{H}_0)$ occurs at

$$\gamma_\mathrm{intersect} = \frac{1+\overline{\xi}}{\overline{\xi}} \log\left(\frac{1-q}{q}[1+\overline{\xi}]\right) \tag{8}$$

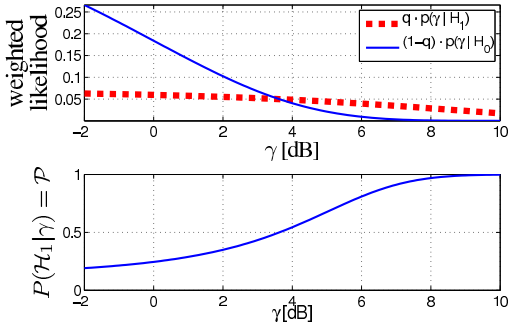
and demarks the point where the GLR is $\Lambda = 1$ and where the resulting *a posteriori* SPP is $\mathcal{P} = 0.5$ (cf. Figure 1(a)).

Besides the observation, $\overline{\gamma}$, and its degrees of freedom, $\overline{r}$, the estimate of $\mathcal{P}$ depends on the *a priori* SPP, $q = P(\mathcal{H}_1)$, and the *a priori* SNR $\overline{\xi}$ (equations (1) and (7)). Since its introduction in [2], an estimate, $\widehat{\xi}$, of the *a priori* SNR is usually obtained using the *decision-directed* approach [2], [3], [4], [9], as:

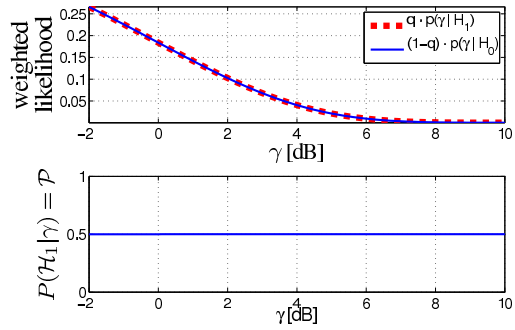
$$\widehat{\xi}(k,l) = \alpha \frac{|\widehat{S}(k,l-1)\,|^2}{\sigma_N^2(k,l-1)} + (1-\alpha)(\gamma(k,l)-1)\,. \tag{9}$$

The smoothing factor α is an important tuning factor, responsible for the trade-off between speech distortion and random fluctuations [10]. The estimate $\widehat{\xi}$ is usually bound to be larger than a minimum value ξ_min. Note that as the decision-directed approach depends on an estimate of the clean speech, $\widehat{S}$, the estimation of the SPP and the estimation of clean speech are coupled if the decision-directed approach is used for SPP estimation. While the decision-directed approach is a very powerful approach to estimate the *a priori* SNR for filter gains, there is an intrinsic disadvantage in using the decision-directed approach for the SPP estimator: at time-frequency points where speech is absent, the *a priori* SNR as gained

with the decision-directed approach is very small and thus the two likelihoods (4) and (6) that are compared in the GLR (2) are approximately *the same* (cf. Figure 1(b)). In this case the *a posteriori* SPP estimate, $\widehat{\mathcal{P}}$, does not make use of any information in the observation, but depends only on the *a priori* SPP $P(\mathcal{H}_1) = q$.



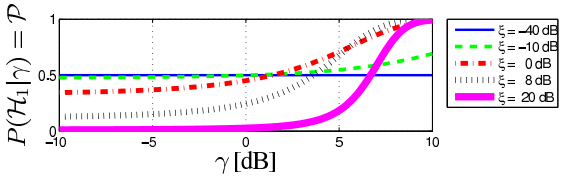
(a) Case $\xi = 8\,\text{dB}$. For $\gamma > \gamma_{\text{intersect}} = 3.6\,\text{dB}$ the weighted ratio (2) of the likelihoods (6) and (4) is larger than one and the SPP, $\mathcal{P}$, is larger than 0.5.



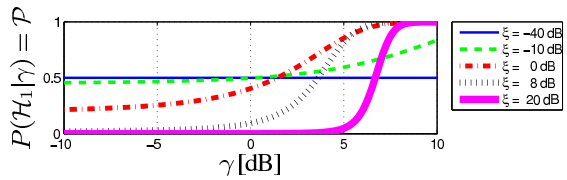
(b) Case $\xi = -40\,\text{dB}$. The likelihoods (4) and (6) are effectively identical and the *a posteriori* SPP, $\mathcal{P}$, equals the prior $q = 0.5$ for all γ.

Fig. 1. Numerator and denominator of the GLR (2) and the resulting *a posteriori* SPP, $\mathcal{P}$ for different *a priori* SNRs, ξ. No smoothing is applied to the observation γ, i.e. $N = 1$ in (3) and $\overline{r} = 2$ in (7). For large ξ (a), the model works well in detecting speech presence or absence. The prior $q = P(\mathcal{H}_1)$ provides for an overall bias. When ξ is small (b), the *a posteriori* SPP yields the prior q (here: $q = 0.5$), independent of the observation.

To overcome this problem Malah, Cox, and Accardi [3] suggested to perform two iterations on the SPP estimator: starting with a fixed $q = 0.5$, the resulting SPP estimate of the first iteration of (1) is used as a frequency dependent *a priori* SPP estimate, $\widehat{q}(k,l)$, in the second iteration. In Figure 2 we compare the *a posteriori* SPP, $\mathcal{P}$, obtained with the conventional method to the iterative method proposed in [3]. The second iteration of (1) causes a steeper transition of $\mathcal{P}$ from its minimum to its maximum, as may be seen by comparing figures 2(a) and 2(b). Note that in case that $\widehat{\xi}$ is very small (e.g. $\xi = -40\,\text{dB}$ in Figure 2(b)), the resulting SPP still equals the prior q, and the second iteration has no effect. This situation is somewhat improved, if we limit $\widehat{\xi}$ to be larger than $\xi_{\text{min}} = -10\,\text{dB}$ as proposed in [3]. The lower limit on $\widehat{\xi}$ enables the SPP estimate to differ from q even in speech pauses, because the two likelihoods (6) and (4) cannot become identical. The second iteration emphasizes this difference, but the resulting SPP estimate is still far from zero for low SNR



(a) The conventional GLR approach (1), (2) considering a single spectral bin ($\overline{r} = 2$) with $q = 0.5$.



(b) The GLR approach considering a single spectral bin ($\overline{r} = 2$) with two iterations and an initial $q = 0.5$ [3].

Fig. 2. Speech presence probability, $\mathcal{P}$, $P(\mathcal{H}_1|\overline{\gamma})$ for different configurations. No smoothing is applied to the observation γ, i.e. $N = 1$ in (3) and $\overline{r} = 2$ in (7). The iterative approach [3] results in a steeper transition as compared to the conventional approach.

conditions, when $\widehat{\xi} = \xi_{\text{min}} = -10\,\text{dB}$ (cf. Figure 2(b)).

Cohen and Berdugo [4] developed the idea of adapting the *a priori* SPP, $\widehat{q}(k,l)$, further. Their approach exploits the correlation of speech presence in neighboring frequency bins of consecutive frames. This is done by taking local and global averages on the *a priori* SNR, $\widehat{\xi}$, as gained via the decision-directed approach. The averages are then mapped on values between 0 and 1 and reinterpreted as *a priori* SPPs. Since the resulting *a priori* SPP estimate, $\widehat{q}$, is mostly either very close to one, or very close to zero, it dominates the *a posteriori* SPP, $\mathcal{P}$. The likelihood-ratio in (2) has then only a minor effect.

III. Proposed Improvement

In the previous section we have given the theoretical basis for the *a posteriori* SPP estimator based on a smoothed observation. In this section we show that smoothing the observation via (3) has two major benefits when compared to the case where only a single bin is considered. First, the estimator yields less random fluctuations because the variance of the observation is reduced. Second, the transition curves for the SPP estimator are steeper, yielding smaller values for the SPP at time-frequency points where speech is absent and larger values where speech is present. Further, we propose using an optimally derived fixed *a priori* SNR, $\overline{\xi}_{\text{fix}}$, and a fixed *a priori* SPP, q, for the estimaton of the *a posteriori* SPP, $\mathcal{P}$.

A. Smoothed observation

Without loss of generality, we discuss the smoothing defined in (3) in the context of a causal system. Instead of considering only one spectral bin of the observation $\gamma(k,l)$ for the GLR, as done in [1], [2], [3], [4], we consider the spectrally neighboring bins $\mathbb{K} = \{k - \Delta k, ..., k, ..., k + \Delta k\}$, and preceeding frames $\mathbb{L} = \{l - \Delta l, ..., l\}$ giving $N = (\Delta l + 1) \cdot (2\Delta k + 1)$ in (3). Figure 3 illustrates those N bins used for smoothing. With (6) and [11, (3.381.4)] it can be shown that the random variable $\overline{\gamma}$

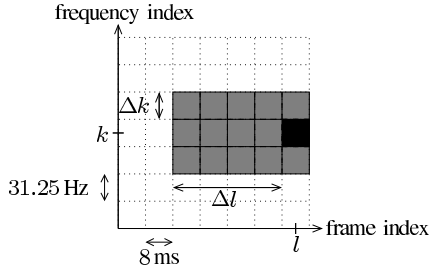


Fig. 3. Illustration of the computation of the smoothed observation $\overline{\gamma}(k,l)$ via (3) in the time-frequency domain. The current bin (k,l) is marked black. The gray area illustrates the neighboring bins used for the smoothing, giving $N = [\Delta l + 1] \cdot [2\Delta k + 1]$ bins.

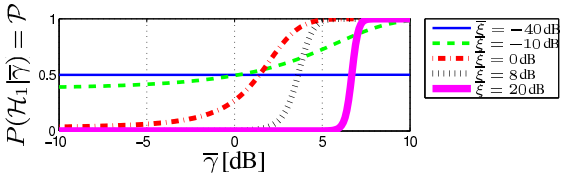


Fig. 4. The speech presence probability, $\mathcal{P}$, for the GLR approach (1), (7) with the proposed smoothing (3) considering 15 spectral bins ($\overline{r} = 10.2$) with $q = 0.5$. Smoothing the observation results in an even steeper transition of $\mathcal{P}$ than the iterative approach in [3] (cf. Figure 2(b)).

has the same mean value $\mathrm{E}\{\overline{\gamma}\} = 1 + \overline{\xi}$ as γ, but its variance is reduced by the factor $2/\overline{r}$.

In Figure 4 the resulting speech presence probability is plotted versus the observation $\overline{\gamma}$ for $\Delta k = 1$, $\Delta l = 4$, $N = 15$, $c_{\mathrm{dof}} = 0.34$, and $\overline{r} = 2Nc_{\mathrm{dof}} = 10.2$. c_{dof} is determined as detailed in Appendix A. Similar to using multiple iterations (Figure 2(b)), the separation between speech and noise bins is more pronounced with the proposed method (Figure 4) as compared to the basic approach in Figure 2(a). The advantage of a steeper transition is that low values of $\overline{\gamma}$ yield a low SPP, when $\overline{\xi}$ is larger than a lower bound.

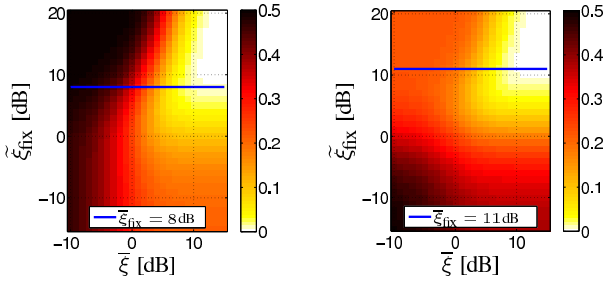
For further theoretical analyses we employ the *false-alarm rate* and the *missed-hit rate*, as used in classical detection and estimation theory, e.g. [12, Ch. 2]. Interpreting the SPP estimator as a detector, we define the false-alarm rate as the probability that a noise-only bin yields an SPP higher than 0.5. Accordingly, the missed-hit rate is the probability that a bin that contains speech yields an SPP lower than 0.5. We show that the smoothed observation $\overline{\gamma}$ reduces both the false-alarm rate and the missed-hit rate. Using [11, (3.381.3)], the false-alarm rate can be written as

$$P_{F,\overline{r}} = \int_{\gamma_{\mathrm{intersect}}}^{\infty} p(\overline{\gamma}|\mathcal{H}_0)\mathrm{d}\overline{\gamma} = \frac{\Gamma(\frac{\overline{r}}{2}, \frac{\overline{r}}{2}\gamma_{\mathrm{intersect}})}{\Gamma(\frac{\overline{r}}{2})}, \qquad (10)$$

where $\gamma_{\mathrm{intersect}}$ is determined according to (8). For $\overline{r} = 2$ this results in $P_{F,\overline{r}} = \exp(-\gamma_{\mathrm{intersect}}) = (\frac{1-q}{q}[1+\overline{\xi}])^{-\frac{1+\overline{\xi}}{\overline{\xi}}}$. For $\overline{\xi} = 8\,\mathrm{dB}$ and $q = 0.5$ the false-alarm rate reduces from $P_{F,\overline{r}} = 10\%$ if one bin is considered ($\overline{r} = 2$) to $P_{F,\overline{r}} = 1\%$ for $N = 15$ bins ($\overline{r} = 2 \cdot N\, c_{\mathrm{dof}} = 10.2$, $c_{\mathrm{dof}} = 0.34$). The missed-hit rate can be written as:

$$P_{M,\overline{r}} = \int_{0}^{\gamma_{\mathrm{intersect}}} p(\overline{\gamma}|\mathcal{H}_1)\mathrm{d}\overline{\gamma} = 1 - \frac{\Gamma(\frac{\overline{r}}{2}, \frac{\overline{r}}{2}\frac{\gamma_{\mathrm{intersect}}}{1+\overline{\xi}})}{\Gamma(\frac{\overline{r}}{2})}. \qquad (11)$$

For $\overline{r} = 2$ this results in $P_{M,\overline{r}} = 1 - \exp(-\frac{\gamma_{\mathrm{intersect}}}{1+\overline{\xi}}) = 1 - (\frac{1-q}{q}[1+\overline{\xi}])^{-\frac{1}{\overline{\xi}}}$. For $\overline{\xi} = 8\,\mathrm{dB}$ and $q = 0.5$ the missed-hit rate reduces from $P_{M,\overline{r}} = 27\%$ considering a single bin ($\overline{r} = 2$) to $P_{M,\overline{r}} = 2\%$ for $N = 15$ bins ($\overline{r} = 10.2$).



(a) $c_F = c_M$ (b) $c_F = 4\,c_M$

Fig. 5. The risk $\mathcal{R}(\overline{\xi}, \widetilde{\xi}_{\mathrm{fix}})$ according to (12) as a function of the unknown *a priori* SNR $\overline{\xi}$ and the assumed $\widetilde{\xi}_{\mathrm{fix}}$. A risk of zero corresponds to perfect detection. The larger the risk, the larger the probability of incorrectly assigning a bin to be speech or having missed a true speech bin. An integration of $\mathcal{R}(\overline{\xi}, \widetilde{\xi}_{\mathrm{fix}})$ along the horizontal line in the linear domain achieves the minimum overall risk. Here a smoothing with $\Delta k = 1$ and $\Delta l = 4$ is assumed. The *a priori* SPP is $q = 0.5$.

B. *Fixed* a priori *SNR and* a priori *SPP*

In Section III-A we have shown that smoothing the observation leads to a steeper transition of $\mathcal{P}$, enabling values closer to zero for the SPP at time-frequency points where speech is absent and values closer to one where speech is present. In time-frequency points where speech is absent the decision-directed approach (9) yields an estimate for the *a priori* SNR, $\widehat{\xi}$, that is very small. For small values of the *a priori* SNR the SPP based on smoothed observations does still not make much use of the observation, because the likelihoods (4) and (6) are effectively identical. In this paper however, we argue that for SPP estimation the *a priori* SNR should reflect the SNR that a typical speech sound would have *if speech were present* in the considered bin. We therefore propose to use a fixed prior q and a constant $\overline{\xi}_{\mathrm{fix}}$ instead of the decision-directed *a priori* SNR estimate $\widehat{\xi}$. This $\overline{\xi}_{\mathrm{fix}}$ should be carefully chosen. If it is too high, the missed-hit rate increases, i.e. weak speech components are not recognized. If it is too low, the false-alarm rate increases, i.e. random fluctuations occur in $\mathcal{P}$.

The optimal choice for $\overline{\xi}_{\mathrm{fix}}$ is found by minimizing the average cost for a detection, which is denoted as the risk $\mathcal{R}$. With an assumed $\widetilde{\xi}_{\mathrm{fix}}$, $\gamma_{\mathrm{intersect}} = \frac{1+\widetilde{\xi}_{\mathrm{fix}}}{\widetilde{\xi}_{\mathrm{fix}}} \log\left(\frac{1-q}{q}[1+\widetilde{\xi}_{\mathrm{fix}}]\right)$, (10), and (11), the risk combines the false-alarm rate, $P_{F,\overline{r}}$, and the missed-hit rate, $P_{M,\overline{r}}$, as

$$\mathcal{R}(\overline{\xi}, \widetilde{\xi}_{\mathrm{fix}}) = c_F\,[1-q]\,P_{F,\overline{r}}(\widetilde{\xi}_{\mathrm{fix}}) + c_M\,q\,P_{M,\overline{r}}(\widetilde{\xi}_{\mathrm{fix}}, \overline{\xi}), \qquad (12)$$

where c_F, c_M are the respective costs. The probabilities for correct detection are not considered in (12), since their cost is assumed to be zero. Note that the missed-hit rate depends on the assumed *a priori* SNR, $\widetilde{\xi}_{\mathrm{fix}}$, and the unknown *a priori* SNR, $\overline{\xi}$. The false-alarm rate, however, is independent of the signal power and depends only on the assumed $\widetilde{\xi}_{\mathrm{fix}}$. In Figure 5, the risk is illustrated for different costs c_F and c_M.

We find an optimal $\overline{\xi}_{\mathrm{fix}}$ by minimizing the risk for all $\overline{\xi}$ between -10 dB and 15 dB. The corresponding integral is

solved numerically in the linear domain, i.e. from $\overline{\xi}$ = 0.1 to 32:

$$\overline{\xi}_{\text{fix}} = \arg\min_{\widetilde{\xi}_{\text{fix}}} \int_{0.1}^{32} \mathcal{R}(\overline{\xi}, \widetilde{\xi}_{\text{fix}}) \mathrm{d}\overline{\xi} \,. \tag{13}$$

When the resulting SPP estimate is applied to a speech enhancement framework, the costs for false-alarms, c_F, and missed-hits, c_M, control the trade-off between noise-leakage and speech distortion. These costs as well as the range of the integral in (13) can be adjusted, such that the performance of the SPP estimator is optimal for the application of interest. A choice of $c_M = c_F = 1$ and zero cost for perfect detection minimizes the total probability of error [12]. When $N = 15$ bins with $c_{\text{dof}} = 0.34$ are considered the optimization (13) yields $\overline{\xi}_{\text{fix}} = 8\,\text{dB}$. Note that using a value of $\overline{\xi}_{\text{fix}} = 8\,\text{dB}$ in the GLR does not mean that the *a posteriori* SNR has to be higher than $\overline{\gamma} \approx 8\,\text{dB}$ to "detect" speech presence ($\mathcal{P} > 0.5$), but only higher than $\gamma_{\text{intersect}} = 3.6\,\text{dB}$ (cf. figures 1(a), 2, and 4).

As in [1] we assume that the probabilities of speech presence and speech absence in each bin are *a priori* equal, and thus set $q = 0.5$.

IV. Implementation

In the previous section we have shown that smoothing the observation reduces the false-alarm rate and the missed-hit rate when the SPP is interpreted as a detector. The smoothing proposed in Section III-A is done by averaging over N neighboring bins. The number of bins, N, depends on the spectral neighbors, Δk, and temporal neighbors, Δl, as $N = [\Delta l + 1] \cdot [2\Delta k + 1]$ (cf. Figure 3). The parameters Δl and Δk should be chosen large enough to ensure a low false-alarm rate, but small enough to preserve the fine structure of speech. In [4] and [13] the combination of two initial SPPs $\widehat{\mathcal{P}}_{\text{local}}$ and $\widehat{\mathcal{P}}_{\text{global}}$ has been successfully applied. These initial SPPs are based on different averaging windows. $\widehat{\mathcal{P}}_{\text{global}}$ is based on a relatively large averaging window. Thus its variance is greatly reduced, but the fine structure of the speech signal is lost (see example in Figure 8(a)). On the other hand $\widehat{\mathcal{P}}_{\text{local}}$ is based on a much smaller averaging window. It has a high variance but is able to resolve the fine structure of the speech signal (see example in Figure 8(b)). The two initial SPPs are then combined such that the final SPP estimator yields values close to one only if the global *and* the local SPP have values close to one. This is achieved via a multiplicative combination [4], [13], as:

$$\widehat{\mathcal{P}} = \widehat{\mathcal{P}}_{\text{local}} \cdot \widehat{\mathcal{P}}_{\text{global}} \,. \tag{14}$$

In Figure 12(e) it can be seen that the combined SPP, $\widehat{\mathcal{P}} = \widehat{\mathcal{P}}_{\text{local}} \cdot \widehat{\mathcal{P}}_{\text{global}}$, based on the local and global averages presented in Figure 8, has a low variance but resolves the fine structure of speech.

For our purposes, the following averaging parameters were found to yield a good trade-off between tempo-spectral resolution, missed-hit rate, and false-alarm rate of the proposed SPP estimator. For the temporal smoothing, we propose to average over $\bar{T} = 64\,\text{ms}$ of speech. With

$$\Delta l = (\bar{T} - T_{\text{seg}})/T_{\text{shift}} \,, \tag{15}$$

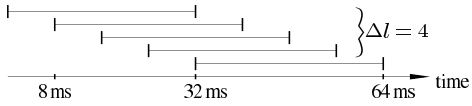


Fig. 6. Illustration of overlapping time segments. With an analysis window of 32 ms and a frame-shift of 8 ms the overall time averaging window of 64 ms results in $\Delta l + 1 = 5$.

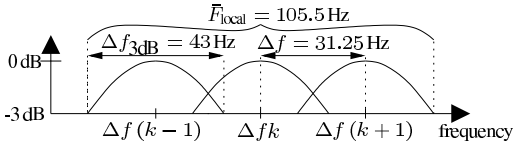


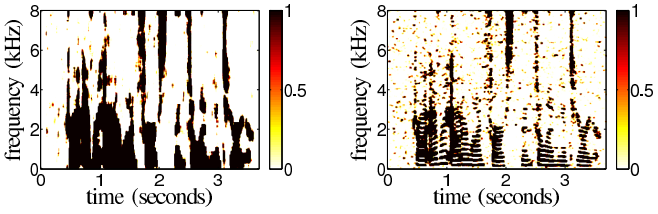
Fig. 7. Illustration of frequency averaging. With a distance between frequency bands of 31.25 Hz and a 3 dB mainlobe bandwidth of 43 Hz the overall frequency averaging window of 105.5 Hz results in $2\,\Delta k + 1 = 3$.

the analysis frame-length of $T_{\text{seg}} = 32\,\text{ms}$ and a frame-shift of $T_{\text{shift}} = (1 - 0.75)\,T_{\text{seg}} = 8\,\text{ms}$ (75% overlap), this results in $\Delta l = 4$ (cf. Figure 6). For the smoothing along frequency we have

$$\Delta k_\chi = \frac{1}{2}(\bar{F}_\chi - \Delta f_{3\,\text{dB}})/\Delta f \,, \tag{16}$$

with a frequency bin distance of $\Delta f = 1/T_{\text{seg}} = 31.25\,\text{Hz}$, and a 3 dB mainlobe bandwidth of the Hann window of approximately $\Delta f_{3\,\text{dB}} \approx 43\,\text{Hz}$. Here χ stands for either the local or the global average. For the local average we want to apply only little smoothing to preserve the fine structure of speech. We propose to average over a frequency window of $\bar{F}_{\text{local}} = 105.5\,\text{Hz}$ which results in $\Delta k_{\text{local}} = 1$ (cf. Figure 7). For the global average we want to have a relatively large frequency window to reduce fluctuations in the observation. We choose a frequency window of 543 Hz that results in $\Delta k_{\text{global}} = 8$.

Using these values for the averaging in the time-frequency plane and assuming costs of $c_F = c_M = 1$ for the computation of the optimal $\overline{\xi}_{\text{fix}}$, we get the parameters given in Table I. The subscript χ stands for either the local or the global average. The parameters are determined as summarized in Figure 9. Note that the resulting parameters $\overline{r}_\chi$ and $\overline{\xi}_{\text{fix},\chi}$ are insensitive to different choices of window overlaps, if the number of time frames Δl is chosen accordingly, as the difference in correlation is considered in c_{dof} (cf. Table I). The algorithm for estimating the SPP is summarized in Figure 10.



(a) $\widehat{\mathcal{P}}_{\text{global}}$ (b) $\widehat{\mathcal{P}}_{\text{local}}$

Fig. 8. The local and global SPPs of the proposed method for the noisy speech in Figure 12(b). The resulting SPP, $\widehat{\mathcal{P}} = \widehat{\mathcal{P}}_{\text{local}} \cdot \widehat{\mathcal{P}}_{\text{global}}$, can be seen in Figure 12(e).

window-overlap	χ	Δk_χ	Δl	N_χ	$c_{\text{dof},\chi}$	$\overline{r}_\chi$	$\overline{\xi}_{\text{fix},\chi}$
75%	local	1	4	15	0.34	10.2	8 dB
	global	8	4	85	0.29	49.3	3 dB
50%	local	1	2	9	0.59	10.7	8 dB
	global	8	2	51	0.50	51.3	3 dB

TABLE I
THE PARAMETER SETTINGS FOR THE PROPOSED SPP ESTIMATOR WITH $\bar{T} = 64$ MS, $\bar{F}_{\text{LOCAL}} = 105.5$ HZ AND $\bar{F}_{\text{GLOBAL}} = 543$ HZ. χ STANDS FOR EITHER THE LOCAL OR THE GLOBAL AVERAGE. THE PARAMETERS ARE DETERMINED FOR A HANN WINDOW WITH 75% OVERLAP AND 50% OVERLAP, RESPECTIVELY.

- choose averaging window, e.g. $\bar{T} = 64\,\text{ms}$, $\bar{F}_{\text{global}} = 543\,\text{Hz}$, and $\bar{F}_{\text{local}} = 105.5\,\text{Hz}$
- compute Δl and Δk_χ via (15), (16)
- compute the number of bins, $N_\chi = (\Delta l + 1) \cdot (2\Delta k_\chi + 1)$
- determine the correction factor, $c_{\text{dof},\chi}$ (Appendix A)
- compute the degrees of freedom, $\overline{r}_\chi = 2N_\chi c_{\text{dof},\chi}$
- compute the optimal *a priori* SNR, $\overline{\xi}_{\text{fix},\chi}$ (13)

Fig. 9. Determination of the parameters in Table I. χ stands for either the local or the global average.

In each signal frame do:
- compute smoothed observation $\overline{\gamma}_\chi$ (3)
- compute $\widehat{\mathcal{P}}_\chi$ via (1) and (7) using $\overline{\xi}_{\text{fix},\chi}$ and $q = 0.5$
- compute the overall SPP: $\widehat{\mathcal{P}} = \widehat{\mathcal{P}}_{\text{global}} \cdot \widehat{\mathcal{P}}_{\text{local}}$

Fig. 10. Proposed SPP estimation algorithm. χ stands for either the local or the global average.

V. APPLICATION TO A SPEECH ENHANCEMENT TASK

A Minimum Mean Square Error (MMSE) estimate $\widehat{S} = \mathrm{E}\{S|Y\}$ of the clean speech spectral coefficients, S, can be obtained in each time-frequency point via the multiplicative gain functions $G_{\mathcal{H}_1}$ and $G_{\mathcal{H}_0}$ for the cases of speech presence and speech absence, respectively:

$$\begin{aligned} \widehat{S} &= \mathcal{P} \cdot \mathrm{E}\{S|Y,\mathcal{H}_1\} + (1-\mathcal{P})\mathrm{E}\{S|Y,\mathcal{H}_0\} \\ &= [\mathcal{P} \cdot G_{\mathcal{H}_1} + (1-\mathcal{P}) \cdot G_{\mathcal{H}_0}] \cdot Y \\ &= \widetilde{G} \cdot Y\,. \end{aligned} \tag{17}$$

For speech absence the clean speech estimator $\mathrm{E}\{S|Y,\mathcal{H}_0\}$ is zero [1], [3], and the resulting gain function is given by

$$\widetilde{G} = \mathcal{P} \cdot G_{\mathcal{H}_1}\,. \tag{18}$$

We use the log spectral amplitude (LSA) estimator as proposed in [14]. The LSA estimator is especially popular because of its robustness against estimation errors in $\widehat{\sigma}_N^2$ which results in less musical noise [4]. If speech presence uncertainty is incorporated into the LSA, this results in [15]

$$\begin{aligned} \widehat{|S|}_{\text{LSA}} &= e^{\mathcal{P}\cdot\mathrm{E}\{\log(|S|)|Y,\mathcal{H}_1\}+(1-\mathcal{P})\cdot\mathrm{E}\{\log(|S|)|Y,\mathcal{H}_0\}} \\ &= G_{\mathcal{H}_1}^{\mathcal{P}} \cdot G_{\mathcal{H}_0}^{1-\mathcal{P}} \cdot |Y| \\ &= \widetilde{G} \cdot |Y|\,. \end{aligned} \tag{19}$$

Again, in speech absence the MMSE optimal estimator yields a gain function $G_{\mathcal{H}_0}$ that is zero. However, then $G_{\mathcal{H}_0}^{1-\mathcal{P}}$ is only one for $\mathcal{P} \equiv 1$ and zero otherwise. Therefore, in practice $G_{\mathcal{H}_0}$ has to be set to a lower limit, e.g. $20\log_{10} G_{\mathcal{H}_0} = -25\,\text{dB}$ as proposed in [4]. The estimator in (19) results in larger improvements in the segmental SNR than the multiplicative estimator (17) but also in more speech distortions [5]. This dilemma justifies using the multiplicative modification (17) also for the LSA estimator as proposed in [3].

Often, a lower bound, G_{min}, is used on the overall gain function resulting in the clean speech estimate

$$\widehat{S} \;=\; \max\{\widetilde{G}, G_{\text{min}}\} \cdot Y\,. \tag{20}$$

A higher value of G_{min} helps masking musical noise [3] and limiting speech distortion at the price of a reduced noise reduction. The optimization of speech enhancement filters therefore aims at finding estimators $\widehat{\mathcal{P}}$ and $G_{\mathcal{H}_1}$ that exhibit as few statistical outliers as possible while introducing minimal speech distortion, so that lower choices for G_{min} are possible.

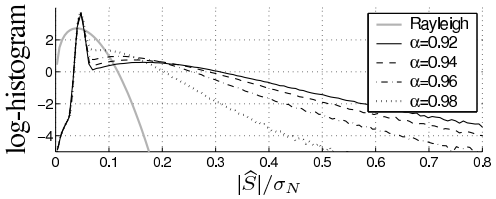
VI. EVALUATION

In this section, we apply the SPP $\widehat{\mathcal{P}}$ to a speech enhancement filter. Here, $\widehat{\mathcal{P}}$ is applied as a multiplicative *soft-gain* as in (18). First, we assess the tendency of a filter to produce musical noise when the estimators under investigation are used. Then, we compare the estimated SPP as obtained by the proposed estimator with the results obtained by the SPP estimators according to [3] and [4]. We measure the speech distortion and the noise leakage introduced by different SPP estimators, as well as the segmental SNR improvement for different noise types.

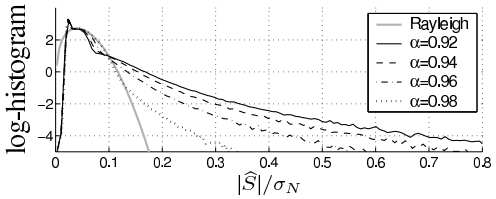
The filter used in this experiment is the LSA estimator [14], as this filter is known to produce little musical noise in stationary noise with a high value of α. The *a priori* SNR ξ, that is a parameter of the LSA-estimator, is estimated by the decision-directed estimation approach (9). For all our experiments we set $10\log_{10}\xi_{\text{min}} = -25\,\text{dB}$ and $20\log_{10} G_{\text{min}} = -25\,\text{dB}$.

We first assess the amount of statistical outliers that may be perceived as musical noise in signals processed by the different approaches. In [16] listening test showed that reducing the amount of spectral outliers yields a higher signal quality. To assess the amount of spectral outliers, we filter a noise-only signal consisting of stationary white Gaussian noise, i.e. $Y = N$. The outlier statistics reflects spectral bins with very low speech energy like in the higher frequencies during voiced speech. As the noise is stationary and white, the spectral bins greater than the DC frequency bin and smaller than the Nyquist frequency bin have the same statistics in all frames and are combined for a statistical analysis of the processed noise.

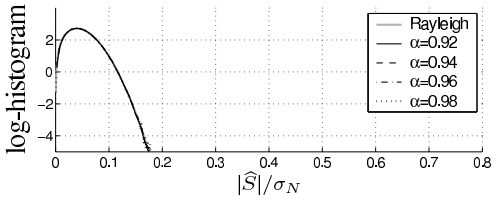
In Figure 11 log-histograms of the normalized filter output $|\widehat{S}|/\sigma_N$ are given. Note that in the experiment of Figure 11, $\widehat{S}$ stands for processed noise, as gained by (18) and (20). Fluctuations in the processed noise appear as an increased rate of spectral outliers, i.e. values much greater than the mean. Informal listening experiments show a strong correlation between the heaviness of the tails in the log-histogram and the perceived amount of musical noise. In Figure 11, different



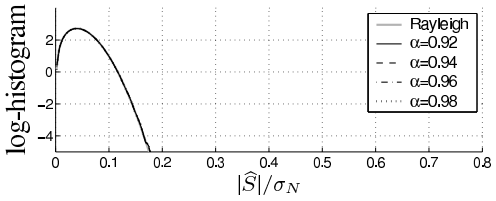
(a) LSA, no soft-gain



(b) LSA with soft-gain [3]



(c) LSA with soft-gain [4]



(d) LSA with the proposed soft-gain

Fig. 11. Log-histogram of the normalized filtered spectrum $|\widehat{S}|/\sigma_N$ for Gaussian noise. For comparison, the Rayleigh distribution of the magnitude of Gaussian noise is also given. The soft-gain [3] does not prevent musical noise. As in the case of no soft-gain, the lower the smoothing constant α of the decision-directed approach (9), the more spectral outliers can be observed. Large outliers are perceived as musical noise. The soft-gain [4] and the proposed soft-gain method result in hardly any spectral outliers for stationary Gaussian noise.

values of α are considered for (9). For lower values of α the amount of spectral outliers in processed noise increases [10]. This can be observed as heavy tailed histograms, which indicates an increased amount of musical noise.

The auditory relevance of the outliers in the histograms is also related to the shape of the histogram around its mean value. E.g., the mean values are different in Figures 11(a) and 11(b). For the LSA without soft-gain, as shown in Figure 11(a), musical noise is better masked by the higher mean value of $|\widehat{S}|/\sigma_N$. Note that this higher mean value is equivalent to a lower overall noise suppression. In the histogram 11(b) the mean value is smaller than in Figure 11(a). In relation to the lowered mean value, the tails are more pronounced. Therefore, in combination with the multiplicative soft-gain [3], the amount of spectral outliers is increased while a higher noise suppression is achieved.

For [4] and the proposed approach we find that audible outliers only occur once or twice per second for stationary Gaussian noise. Otherwise, the processed signal sounds like an attenuated version of the input signal. Accordingly, the histograms are almost identical to the Rayleigh distribution of the scaled input signal, $\widehat{S}(k,l) = G_{\min} Y(k,l)$.

Note that a principal difference between the approach [3] and the two other methods is obvious in Figure 11. As the soft-gain method [3] never fully suppresses the filter output in noise-only regions (cf. Section II), the histogram in Figure 11(b) still exhibits the characteristic shape of Gaussian noise filtered by the LSA-estimator. As the other two methods clearly indicate speech absence, the resulting filter gain is constantly $G_{\min}$, rendering the residual noise more Gaussian.

For the following experiments, we set the smoothing factor in (9) to $\alpha = 0.96$, which is a good trade-off between speech distortion that comes with higher values of α [3], and musical noise.

In the next step of the evaluation, the soft-gain factors, $\widehat{\mathcal{P}}$, are analyzed for a signal that contains speech. Figures 12(c) to 12(e) depict the soft-gain factors for the evaluated estimators. The corresponding spectrograms of the clean and the noisy input signal are given in Figures 12(a) and 12(b), respectively. In Figure 12(c) it becomes clear that the conventional SPP estimator does not result in a SPP of zero in noise-only spectral regions. The method [4] and the proposed method clearly distinguish between speech and noise. Nevertheless, spectral speech structures of low energy are recovered more often by the proposed method (cf. Figures 12(d) and 12(e)).

In Section III-A we have shown that smoothing the observation reduces the false-alarm rate (10). In Figure 12(e) the lower false-alarm rate results in much less estimates $\widehat{\mathcal{P}} > 0.5$ in noise-only spectral regions, as compared to Figure 12(c). In Figure 12(c), false-alarms cause dark speckles in the gray spectral regions where no speech is present.

Inspired by [17] we evaluate the SPP estimators, $\widehat{\mathcal{P}}$, in terms of speech distortion (SD) and noise leakage (NL), which can be seen as measures for missed-hit rate and false-alarm rate, respectively. As in [18] we create an ideal binary speech presence mask, $\mathcal{P}_{\text{id}}(k,l)$, from the clean speech signal $S(k,l)$ that contains ones at all STFT-bins, (k,l), where the energy is no less than 50 dB below the maximum bin energy in the particular speech signal. We then compute two error-signals, $E_{\text{SD}}(k,l)$ and $E_{\text{NL}}(k,l)$ as:

$$E_{\text{SD}}(k,l) = \max\left\{\mathcal{P}_{\text{id}}(k,l) - \widehat{\mathcal{P}}(k,l), 0\right\} S(k,l) \quad (21)$$

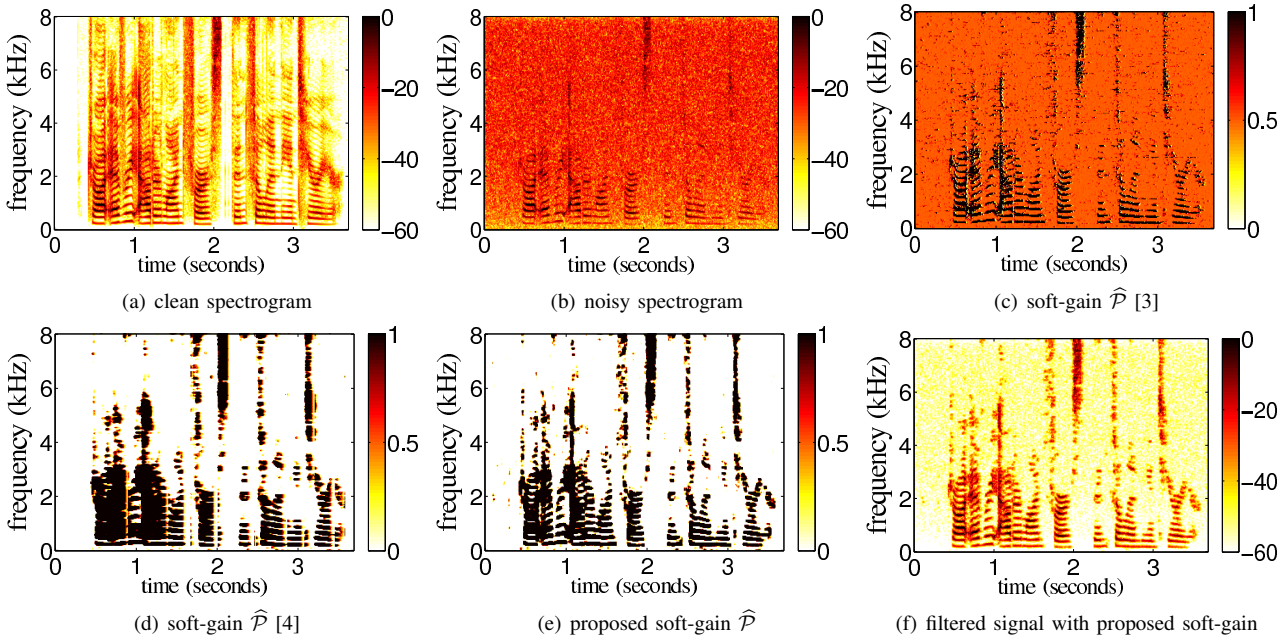
$$E_{\text{NL}}(k,l) = \max\left\{\widehat{\mathcal{P}}(k,l) - \mathcal{P}_{\text{id}}(k,l), 0\right\} N(k,l) \quad (22)$$

E_{SD} contains those speech bins that are marked as speech by the ideal mask, $\mathcal{P}_{\text{id}}$, but are attenuated by the SPP estimator $\widehat{\mathcal{P}}$. E_{NL} contains those noise bins that are marked as noise by the ideal mask, $\mathcal{P}_{\text{id}}$, but are not fully suppressed by the SPP estimator $\widehat{\mathcal{P}}$. These error signals are then related to the ideal speech signal, S_{id}, and the corresponding ideal noise signal, N_{id}, which are gained as:

$$S_{\text{id}}(k,l) = \mathcal{P}_{\text{id}}(k,l)\, S(k,l) \quad (23)$$

$$N_{\text{id}}(k,l) = (1 - \mathcal{P}_{\text{id}}(k,l)) N(k,l)\ . \quad (24)$$

After taking the inverse Fourier transform and reconstructing the time signal by overlapping and adding the signal segments



(a) clean spectrogram (b) noisy spectrogram (c) soft-gain $\widehat{\mathcal{P}}$ [3]

(d) soft-gain $\widehat{\mathcal{P}}$ [4] (e) proposed soft-gain $\widehat{\mathcal{P}}$ (f) filtered signal with proposed soft-gain

Fig. 12. Soft-gain for a speech signal in pink noise at 5 dB segmental SNR. Graph (a): clean spectrogram of the prompt "Draw every outer line first, then fill in the interior". Graph (b): spectrogram of the signal with pink noise at $\text{SNR}_{\text{seg}}\{y(t)\} = 5\,\text{dB}$. Graph (c): the soft-gain with two iterations according to [3]. Spectral regions without speech presence are not clearly recognized as such. Graph (d): Soft-gain according to Cohen and Berdugo [4]; speech and non-speech are distinguished properly, but spectrally narrow regions of speech are not detected. Graph (e): Proposed method; more low energy speech is preserved. Speech absence in between pitch harmonics is also indicated more clearly. Graph (f): Spectrogram of signal filtered with proposed soft-gain. Note that we pre-emphasized the signals in the spectrograms (a),(b),(f) for a better visualization of high-frequency components.

we get the time domain signals $e_{\text{SD}}(t)$, $e_{\text{NL}}(t)$, $s_{\text{id}}(t)$ and $n_{\text{id}}(t)$. The final measures for speech distortion and noise leakage are then gained as:

$$\text{SD} = \frac{\sum_t e_{\text{SD}}^2(t)}{\sum_t s_{\text{id}}^2(t)} \tag{25}$$

$$\text{NL} = \frac{\sum_t e_{\text{NL}}^2(t)}{\sum_t n_{\text{id}}^2(t)}\,. \tag{26}$$

The measure for speech distortion, SD, thus indicates the percentage of the speech energy that the corresponding SPP estimator neglects while the measure for noise leakage, NL, indicates how much energy from the noise-only bins is not attenuated (in percent). Thus SD equals 100% if all speech coefficients indicated by the ideal mask, $\mathcal{P}_{\text{id}}$, are attenuated by $\widehat{\mathcal{P}}$ and SD = 0% if $\widehat{\mathcal{P}}(k,l) = 1$ wherever $\mathcal{P}_{\text{id}}(k,l) = 1$. The NL equals 0% if $\widehat{\mathcal{P}}(k,l) = 0$ for all noise-only bins.

Furthermore we compute the segmental SNR improvement, when the SPP estimators are employed in a speech enhancement framework via (18) and (20). With $\widehat{s}(t)$ the enhanced signal in the time domain, obtained from $\widehat{S}(k,l)$, the segmental SNR, $\text{SNR}_{\text{seg}}\{\widehat{s}(t)\}$, compares the clean speech signal, $s(t)$, and its estimate, $\widehat{s}(t)$, on a segmental basis:

$$\text{SNR}_{\text{seg}}\{\widehat{s}(t)\} = \frac{10}{|\mathbb{L}_{\text{sp}}|}\sum_{\tilde{l}\in\mathbb{L}_{\text{sp}}}\log_{10}\frac{\sum_{t=0}^{\tilde{L}-1} s^2(\tilde{l}\tilde{L}+t)}{\sum_{t=0}^{\tilde{L}-1}(s(\tilde{l}\tilde{L}+t)-\widehat{s}(\tilde{l}\tilde{L}+t))^2}, \tag{27}$$

where $\mathbb{L}_{\text{sp}}$ is the set of frame indices $\tilde{l}$ that belong to non-overlapping speech active frames of an utterance $s(t)$. The length of each frame is set to $\tilde{L}/f_s = 10\,\text{ms}$. Note that $\text{SNR}_{\text{seg}}\{\widehat{s}(t)\}$ considers both noise suppression and speech distortion. Also note that in general, for input SNRs below 0 dB, the segmental output SNR is not necessarily reasonable, since for input SNRs below 0 dB the segmental SNR would indicate an improvement even if $\widehat{\mathcal{P}}$ is zero for all time-frequency points. Therefore it has to be read together with SD.

Table II gives the results for stationary (white, pink, car) and nonstationary (babble, traffic, subway) noise types. Table II states the SD, the NL, and the improvement of the segmental SNR, $\Delta\text{SNR}_{\text{seg}} = \text{SNR}_{\text{seg}}\{\widehat{s}(t)\} - \text{SNR}_{\text{seg}}\{y(t)\}$, with $y(t)$ the unprocessed noisy time signal. The results are given as the average of 10 phonetically balanced sentences from the TIMIT database [19] (5 male, 5 female), for input SNRs between -10 dB and 15 dB. It can be clearly seen that the noise leakage, NL, for method [3] is the highest. This is due to the fact that speech absence is not clearly indicated, but results in SPP estimates close to 0.5 (cf. Figure 12(c)). The proposed method and the method [4] clearly reduce the noise leakage as compared to [3]. The price is a higher speech distortion, SD. However, the proposed method yields a better trade off than [4], as the proposed method yields lower values for the SD as well as the NL. This means that more speech components are preserved *and* more noise is attenuated. This holds for stationary and nonstationary noise and all considered input SNRs. In terms of the SNR improvement, $\Delta\text{SNR}_{\text{seg}}$, the proposed method indicates only slightly better performance as compared to both competing estimators [3] and [4]. For stationary car noise at 0 dB input SNR these improvements are 1.9 dB and 1.3 dB as compared to [3] and [4], respectively.

According to (13) the proposed method has been optimized for input SNRs between -10 dB to 15 dB. Nevertheless, we report that the principal performance stays the same below and above these values. For extremely low input SNRs below -20 dB and the above optimization the proposed method and method [4] indicate mostly speech absence while [3] yields *a posteriori* SPP estimates close to the *a priori* SPP. As the parameters used in the proposed method are determined in an optimal way, they can be easily adjusted if the estimator is meant to perform in different application scenarios for instance by changing the SNR range of the integral in (13) or by changing the costs for false-alarms and missed-hits in (12).

VII. Conclusion

We have given the theoretical basis for an *a posteriori* speech-presence-probability (SPP) estimator based on the smoothed *a posteriori* signal-to-noise ratio (SNR). Smoothing the *a posteriori* SNR has the major benefit of reducing the variance of the SPP estimate. By interpreting the estimator as a detector, we have shown that this increases the estimation performance in terms of a lower false-alarm rate and a lower missed-hit rate. The perceptual benefits are less musical noise and less speech distortion when the SPP estimator is incorporated into a speech enhancement framework.

The *a posteriori* SPP estimator is based on the ratio of the likelihoods of speech presence and speech absence, weighted by their prior probabilities. In state-of-the-art *a posteriori* SPP estimators the likelihood-ratio is usually based on an adaptively estimated *a priori* SNR estimate that takes very small values at time-frequency points where speech is absent (e.g. between the harmonics of voiced speech). We have shown that then the resulting *a posteriori* SPP estimators yield only the prior probabilities. Existing approaches attempt to mend this undesired behavior by using signal dependent speech presence priors, i.e. by adaptively tracking the prior probabilities.

In this paper we present an approach to overcome the necessity for adaptively tracking the *a priori* SPP and the *a priori* SNR. We argue that for speech presence probability estimation the *a priori* SNR should reflect the SNR that is expected when speech is present. We therefore use an optimal fixed *a priori* SNR that minimizes the total probability of error. Our modifications provide low *a posteriori* SPP estimates at time-frequency points where speech is absent, without the necessity for adaptively tracking the *a priori* SPP. Since we do not use an adaptively estimated *a priori* SNR the proposed procedure enables a decoupling of the estimation of the speech-presence-probability and the estimation of the clean-speech coefficients. Furthermore, it is shown that the proposed method results in a better trade-off between speech distortion and noise leakage than state-of-the-art SPP estimators.

Appendix

A. Determination of c_{dof}

With (4) and [11, (3.381.4)], in case of speech absence the spectro-temporal smoothing (3) yields a variance of $\overline{\gamma}(k,l)$ given by

$$\operatorname{var}\{\overline{\gamma}\} = \frac{2}{\overline{r}}\,, \tag{28}$$

with $\overline{r}$ the degrees of freedom of $\overline{\gamma}$. In case of uncorrelated spectral bins $\gamma(\kappa,\lambda)$ in (3), each bin contributes 2 degrees of freedom to the sum. However, as the frame-wise spectral analysis uses overlapping frames of limited length, neighboring spectral bins are correlated. This reduces the degree of freedom per bin which we express by the correction factor $c_{\text{dof}} \leq 1$ giving

$$\overline{r} = 2c_{\text{dof}}N\,. \tag{29}$$

Substituting (29) in (28) and solving for c_{dof} we get

$$c_{\text{dof}} = \frac{1}{N\operatorname{var}\{\overline{\gamma}\}}\,. \tag{30}$$

Thus, the correction factor c_{dof} can be empirically determined with (30) by measuring $\operatorname{var}\{\overline{\gamma}\}$ for stationary noise that is segmented with a given frame-overlap and a given type of analysis window and then transformed into the spectral domain.

B. Acknowledgement

The work of Colin Breithaupt was funded by the German Research Foundation **DFG**.

References

[1] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Proc.*, vol. 28, no. 2, pp. 137–145, 1980.

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

[3] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *Proceedings, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 2, pp. 789–792, 1999.

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

[5] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. John Wiley & Sons, 2006.

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

[7] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

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

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

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

[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals Series and Products*, 6th ed., A. Jeffrey and D. Zwillinger, Ed. Academic Press, 2000.

[12] H. Van Trees, *Detection, Estimation, and Modulation Theory. Part I*. John Wiley & Sons, 1968.

[13] K. V. Sørensen and S. V. Andersen, "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 18, pp. 2954–2964, 2005.

	-10 dB			-5 dB			0 dB			5 dB			10 dB			15 dB		
	SD [%]	NL [%]	ΔSNR$_{seg}$ [dB]	SD [%]	NL [%]	ΔSNR$_{seg}$ [dB]	SD [%]	NL [%]	ΔSNR$_{seg}$ [dB]	SD [%]	NL [%]	ΔSNR$_{seg}$ [dB]	SD [%]	NL [%]	ΔSNR$_{seg}$ [dB]	SD [%]	NL [%]	ΔSNR$_{seg}$ [dB]
	SPP according to [3]																	
white	2.8	30.1	11.8	1.3	30.4	9.6	0.6	30.8	7.7	0.2	31.1	5.9	0.1	31.3	4.2	0.0	31.7	2.5
pink	3.1	32.5	11.0	1.4	32.8	8.8	0.6	33.0	7.0	0.2	33.2	5.5	0.1	33.3	4.1	0.0	33.8	2.8
car	0.2	46.5	14.0	0.1	45.7	13.5	0.1	45.2	12.4	0.0	45.2	11.0	0.0	46.3	9.3	0.0	46.4	7.6
babble	2.5	66.2	5.1	1.1	66.2	4.2	0.5	66.0	3.4	0.2	65.6	2.7	0.1	65.2	1.8	0.0	64.7	0.9
traffic	2.7	58.9	9.3	1.2	59.4	7.6	0.5	59.9	6.1	0.2	60.2	4.6	0.1	59.9	3.2	0.0	59.8	1.9
subway	2.7	59.5	8.1	1.2	59.2	6.9	0.5	59.0	5.6	0.2	58.7	4.3	0.1	58.4	3.1	0.0	58.0	1.7
	SPP according to [4]																	
white	14.4	2.8	12.1	6.1	3.2	9.7	2.3	4.8	7.7	0.8	7.5	5.8	0.2	11.6	4.1	0.1	17.2	2.5
pink	15.9	3.9	11.4	6.6	5.1	9.0	2.3	7.0	7.0	0.7	9.4	5.4	0.2	12.6	4.0	0.1	17.1	2.8
car	0.6	26.6	14.7	0.3	25.3	14.1	0.2	24.7	13.0	0.1	25.6	11.4	0.0	28.4	9.6	0.0	30.0	7.8
babble	7.4	56.9	5.3	3.0	57.7	4.3	1.1	59.1	3.4	0.3	60.5	2.6	0.1	62.1	1.8	0.0	63.5	0.9
traffic	10.5	46.8	9.8	4.3	48.6	7.8	1.7	51.9	6.1	0.6	56.0	4.6	0.2	59.9	3.2	0.1	63.9	1.9
subway	9.4	49.2	8.7	4.0	50.6	7.1	1.4	52.6	5.6	0.5	54.9	4.3	0.1	57.6	3.0	0.1	60.3	1.7
	proposed SPP estimator																	
white	11.1	1.9	12.3	4.6	2.1	9.9	1.6	2.7	8.0	0.5	3.8	6.1	0.2	5.8	4.3	0.0	9.1	2.7
pink	12.2	2.9	11.7	4.9	3.4	9.3	1.7	4.1	7.3	0.5	5.3	5.7	0.1	6.8	4.3	0.0	9.4	3.0
car	0.6	13.2	16.6	0.3	12.7	15.6	0.2	12.8	14.3	0.1	12.9	12.7	0.0	13.7	11.0	0.0	14.2	9.1
babble	7.1	56.2	5.4	2.9	56.5	4.4	1.0	56.8	3.5	0.3	56.8	2.7	0.1	57.2	1.9	0.0	57.8	1.0
traffic	9.0	45.3	9.8	3.6	46.4	7.9	1.3	48.2	6.2	0.5	49.7	4.7	0.2	51.1	3.3	0.1	53.5	2.0
subway	7.8	49.3	8.7	3.3	49.2	7.2	1.1	49.6	5.8	0.4	50.2	4.4	0.1	51.3	3.1	0.1	52.8	1.8

TABLE II
THE RESULTS IN TERMS OF SPEECH DISTORTION, SD, NOISE LEAKAGE, NL, AND SEGMENTAL SNR IMPROVEMENT, ΔSNR$_{SEG}$. THE PROPOSED METHOD YIELDS A GOOD TRADE-OFF BETWEEN SD AND NL AND SLIGHT IMPROVEMENTS IN TERMS OF ΔSNR$_{SEG}$.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Proc.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[15] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook*, R. Dorf, Ed. CRC Press, 2006.

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

[17] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, Sept. 2004.

[18] 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 Transactions on Audio, Speech, and Language Proc.*, vol. 15, no. 06, pp. 1741–1752, Aug. 2007.

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

Timo Gerkmann studied Electrical Engineering at the University of Bremen, Bremen, Germany, and the Ruhr-Universität Bochum, Bochum, Germany. He received the Dipl.-Ing. degree from the Ruhr-Universität Bochum in 2004. He is currently pursuing the Dr.-Ing. degree at the Institute of Communication Acoustics, Ruhr-Universität Bochum.

From January 2005 to July 2005 he visited Siemens Corporate Research in Princeton, NJ, where he worked on artificial bandwidth extension. His main research interests are digital speech and audio processing, including single- and multichannel speech enhancement.

Colin Breithaupt was born in 1974 in Böblingen, Germany. He received the Dipl.-Ing. degree in electrical engineering at RWTH Aachen University, Aachen, Germany, in 2002.

He then worked as a researcher at the Technische Universität Braunschweig, Braunschweig, Germany. In 2003 he joined the Institute of Communication Acoustics at the Ruhr-Universität Bochum, Bochum, Germany, to continue his research there. The subject of his research is statistical signal processing for noise reduction in the field of audio applications and automatic speech recognition systems. At the ATR Spoken Language Communication Research Laboratories in Kyoto, Japan, he gathered experience in the field of noise reduction for robust automatic speech recognition. At the Speech and Hearing group (SpandH) of the department of computer science (DCS) at the University of Sheffield, Sheffield, U.K., he investigated the auditory perception of noises by humans.

Rainer Martin (S'86-M'90-SM'01) received the Dipl.-Ing. and Dr.-Ing. degrees from RWTH Aachen University, Aachen, Germany, in 1988 and 1996, respectively, and the MSEE degree from Georgia Institute of Technology, Atlanta, in 1989.

From 1996 to 2002, he has been a Senior Research Engineer with the Institute of Communication Systems and Data Processing, RWTH Aachen University. From April 1998 to March 1999 he was on leave to the AT&T Speech and Image Processing Services Research Lab, Florham Park, N.J. From April 2002 until October 2003 he was a professor of Digital Signal Processing at the Technische Universität Braunschweig, Braunschweig, Germany. Since October 2003, he has been a Professor of Information Technology and Communication Acoustics at Ruhr-Universität Bochum, Bochum, Germany. His research interests are signal processing for voice communication systems, hearing instruments, and human-machine interfaces. He is coauthor with P. Vary of *Digital Speech Transmission - Enhancement, Coding and Error Concealment* (Wiley, 2006) and coeditor with U. Heute and C. Antweiler of *Advances in Digital Speech Transmission* (Wiley, 2008).