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 shall not depend on the observed signal. We present a detection theoretic framework for estimating the *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.

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 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 \(f_s = 16\text{ kHz}\). We assume that the noise power, \(\sigma^2_N = \mathbb{E}\{|N|^2\}\), is known. In practice, the noise variance, \(\sigma^2_N\), can be gained using the minimum statistics [6] or the improved minima controlled recursive averaging [7] noise power estimation approach. Introducing the \textit{a posteriori} signal-to-noise ratio (SNR), \(\gamma = \frac{|Y|^2}{\sigma^2_N}\), the probability of speech presence given the observation \(\gamma\), may be written as [5]:

\[
\mathcal{P} = P(\mathcal{H}_1 | \gamma) = \frac{\Lambda}{1 + \Lambda}.
\]  

(1)

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

\[
\Lambda = \frac{q}{(1 - q) p(\gamma | \mathcal{H}_0)}.
\]  

(2)

where \(q = P(\mathcal{H}_1)\) is the \textit{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 \(\gamma\), \(\mathcal{P}\) 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 \textit{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:

\[
\bar{\gamma}(k, l) = \frac{1}{N} \sum_{\kappa \in \mathbb{K}} \sum_{\lambda \in \mathbb{L}} \gamma(\kappa, \lambda).
\]  

(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 \(\bar{\gamma}\) are approximately chi-squared distributed [8]:

\[
p(\gamma | \mathcal{H}_0) = \left(\frac{\pi}{2}\right)^{\frac{\gamma}{2}} \frac{\Gamma\left(\frac{\gamma}{2}\right)}{\Gamma\left(\frac{\gamma}{2}\right)} \exp\left(-\frac{\gamma}{2}\right).
\]  

(4)

Their degree of freedom \(\tau = 2Nc_{\text{dof}}\) is increased as compared to the unsmoothed case where only one spectral bin is considered, i.e. \(\tau = 2\). The correction factor \(c_{\text{dof}}\) 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 \(\bar{\gamma}\) is likely to be larger than \(\mathbb{E}\{\gamma | \mathcal{H}_0\} = 1\). 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 \textit{a priori} SNR

\[
\xi(k, l) = \frac{1}{N} \sum_{\kappa \in \mathbb{K}} \sum_{\lambda \in \mathbb{L}} \xi(\kappa, \lambda),
\]  

(5)

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

\[
p(\gamma | \mathcal{H}_1) = \left(\frac{\pi}{2(1 + \xi)}\right)^{\frac{\gamma}{2} - 1} \frac{\Gamma\left(\frac{\gamma}{2}\right)}{\Gamma\left(\frac{\gamma}{2}\right)} \exp\left(-\frac{\gamma}{2(1 + \xi)}\right).
\]  

(6)

The GLR thus gives

\[
\Lambda(\bar{\gamma}) = \frac{q}{1 - q} \left(\frac{1 + \bar{\xi}}{1 + \frac{\bar{\xi}}{2}}\right)^{\frac{\bar{\tau}}{2}} \exp\left(-\frac{\bar{\tau}}{2(1 + \bar{\xi})}\right),
\]  

(7)

which is then used in (1) to compute the \textit{a posteriori} SPP

\[P(\mathcal{H}_1 | \bar{\gamma}) = \mathcal{P}.
\]

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 \textit{a priori} SNR of \(\xi = 8\text{ dB}\) and \(\xi = -40\text{ dB}\) as a function of the \textit{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 \textit{a posteriori} SNR is converted from linear scale to decibels, as \(\gamma[\text{dB}] = 10\log_{10}(\gamma)\). The intersection of the weighted likelihoods \(q p(\gamma | \mathcal{H}_1)\) and \((1 - q) p(\gamma | \mathcal{H}_0)\) occurs at

\[
\bar{\gamma}_{\text{intersect}} = \frac{1 + \frac{\bar{\tau}}{2}}{\bar{\tau}} \log\left(\frac{1 - q}{q} \frac{1 + \bar{\xi}}{1 + \frac{\bar{\xi}}{2}}\right),
\]  

(8)

and marks the point where the GLR is 1 and where the resulting \textit{a posteriori} SPP is \(\mathcal{P} = 0.5\) (cf. Figure 1(a)).

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

\[
\hat{\xi}(k, l) = \alpha \frac{|\hat{S}(k, l - 1)|^2}{\sigma^2_N(k, l - 1)} + (1 - \alpha)(\hat{\xi}(k, l - 1)).
\]  

(9)

The smoothing factor \(\alpha\) is an important tuning factor, responsible for the trade-off between speech distortion and random fluctuations [10]. The estimate \(\hat{\xi}\) is usually bound to be larger than a minimum value \(\xi_{\text{min}}\). Note that as the decision-directed approach depends on an estimate of the clean speech, \(\hat{S}\), 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 \textit{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 \textit{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, $\hat{P}$, does not make use of any
information in the observation, but depends only on the a
priori SPP $P(H_1) = q$.

$$P(H_1) = P$$

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

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

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, $q(k, l)$, in the second iteration. In
Figure 2 we compare the a posteriori SPP, $P$, obtained with
the conventional method to the iterative method proposed in
[3]. The second iteration of (1) causes a steeper transition of
$P$ from its minimum to its maximum, as may be seen by
comparing figures 2(a) and 2(b). Note that in case that $\xi$
is very small (e.g. $\xi = -40\,$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 $\xi$
to be larger than $\xi_{\text{min}} = -10\,$dB as proposed in [3]. The lower limit
on $\xi$ 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
conditions, when $\hat{\xi} = \xi_{\text{min}} = -10\,$dB (cf. Figure 2(b)).

Cohen and Berdugo [4] developed the idea of adapting
the a priori SPP, $q(k, l)$, 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, $\xi$, 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, $q$, is mostly either very close
to one, or very close to zero, it dominates the a posteriori SPP, $P$. 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, $\xi_{\text{fix}}$, and a fixed a priori SPP, $q$, for the estimation of the a posteriori SPP, $P$.

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 $\mathbb{I} = \{k - \Delta k, ..., k, ..., k + \Delta k\}$, 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 $\gamma$
has the same mean value $E\{ \gamma \} = 1 + \xi$ as $\gamma$, but its variance is reduced by the factor $2/\tau$.

In Figure 4 the resulting speech presence probability is plotted versus the observation $\gamma$ for $\Delta k = 1$, $\Delta l = 4$, $N = 15$, $c_{\text{df}} = 0.34$, and $\tau = 2Nc_{\text{df}} = 10.2$. $c_{\text{df}}$ 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 $\gamma$ yield a low SPP, when $\xi$ 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 $\gamma$ reduces both the false-alarm rate and the missed-hit rate. Using [11, (3.381.3)], the false-alarm rate can be written as

$$P_{F,\tau} = \int_0^{\gamma_{\text{intersect}}} p(\gamma | H_0) d\gamma = \frac{\Gamma(\frac{\tau}{2}, \frac{\xi}{\tau} \gamma_{\text{intersect}})}{\Gamma(\frac{\tau}{2})},$$  \hspace{0.5cm} (10)

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

$$P_{M,\tau} = \int_0^{\gamma_{\text{intersect}}} p(\gamma | H_1) d\gamma = 1 - \frac{\Gamma(\frac{\tau}{2}, \frac{\xi}{\tau} 2^{-\frac{1}{q}})}{\Gamma(\frac{\tau}{2})}.$$  \hspace{0.5cm} (11)

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

### 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 $P$, 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, $\xi$, 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 $\xi_{\text{fix}}$ instead of the decision-directed a priori SNR estimate $\xi$. This $\xi_{\text{fix}}$ 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 $P$.

The optimal choice for $\xi_{\text{fix}}$ is found by minimizing the average cost for a detection, which is denoted as the risk $R$. With an assumed $\xi_{\text{fix}}$, $\gamma_{\text{intersect}} = \frac{1+\xi_{\text{fix}}}{\xi_{\text{fix}}} \log \left( \frac{1-2^q(1+\xi_{\text{fix}})}{q} \right)$, (10), and (11), the risk combines the false-alarm rate, $P_{F,\tau}$, and the missed-hit rate, $P_{M,\tau}$, as:

$$R(\xi, \xi_{\text{fix}}) = c_F [1 - q] P_{F,\tau}(\xi_{\text{fix}}) + c_M q P_{M,\tau}(\xi_{\text{fix}}),$$  \hspace{0.5cm} (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, $\xi_{\text{fix}}$, and the unknown a priori SNR, $\xi$. The false-alarm rate, however, is independent of the signal power and depends only on the assumed $\xi_{\text{fix}}$. In Figure 5, the risk is illustrated for different costs $c_F$ and $c_M$.

We find an optimal $\xi_{\text{fix}}$ by minimizing the risk for all $\xi$ between -10 dB and 15 dB. The corresponding integral is
solved numerically in the linear domain, i.e. from $\tilde{\xi} = 0.1$ to $32$:

$$\bar{\xi}_{\text{fix}} = \arg \min_{\xi_{\text{fix}}} \int_{0.1}^{32} R(\xi, \bar{\xi}_{\text{fix}}) / d\xi.$$  \hspace{1cm} (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 $\bar{\xi}_{\text{fix}} = 8$ dB in the GLR does not mean that the a posteriori SNR has to be higher than $\gamma \approx 8$ dB to “detect” speech presence ($P > 0.5$), but only higher than $\gamma_{\text{intersect}} = 3.6$ 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, $\Delta k$, and temporal neighbors, $\Delta t$, as $N = [\Delta t + 1] \cdot [2\Delta k + 1]$ (cf. Figure 3). The parameters $\Delta t$ and $\Delta k$ 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 $\hat{P}_{\text{local}}$ and $\hat{P}_{\text{global}}$ has been successfully applied. These initial SPPs are based on different averaging windows. $\hat{P}_{\text{global}}$ 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 $\hat{P}_{\text{local}}$ 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:

$$\hat{P} = \hat{P}_{\text{local}} \cdot \hat{P}_{\text{global}}.$$  \hspace{1cm} (14)

In Figure 12(e) it can be seen that the combined SPP, $\hat{P} = \hat{P}_{\text{local}} \cdot \hat{P}_{\text{global}}$, 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 $T = 64$ ms of speech. With

$$\Delta t = (\bar{T} - T_{\text{seg}}) / T_{\text{shift}},$$  \hspace{1cm} (15)

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

$$\Delta k_{\chi} = \frac{1}{2} \left( \hat{F}_{\chi} - \Delta f_{3,\text{db}} \right) / \Delta f,$$  \hspace{1cm} (16)

with a frequency bin distance of $\Delta f = 1 / T_{\text{seg}} = 31.25$ Hz, and a 3 dB mainlobe bandwidth of the Hann window of approximately $\Delta f_{3,\text{db}} \approx 43$ Hz. Here $\chi$ 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 $F_{\text{local}} = 105.5$ 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 $\tilde{\xi}_{\text{fix}}$, we get the parameters given in Table I. The subscript $\chi$ stands for either the local or the global average. The parameters are determined as summarized in Figure 9. Note that the resulting parameters $\tau_{\chi}$ and $\tilde{\xi}_{\text{fix},\chi}$ are insensitive to different choices of window overlaps, if the number of time frames $\Delta t$ is chosen accordingly, as the difference in correlation is considered in $c_{\text{dof}}$ (cf. Table I). The algorithm for estimating the SPP is summarized in Figure 10.

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 $\Delta t = 1 + 5$.

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 $2 \Delta k + 1 = 3$.

Fig. 8. The local and global SPPs of the proposed method for the noisy speech in Figure 12(b). The resulting SPP, $\hat{P} = \hat{P}_{\text{local}} \cdot \hat{P}_{\text{global}}$, can be seen in Figure 12(e).
Again, in speech absence the MMSE optimal estimator yields a gain function \( G_{\eta_0} \) that is zero. However, then \( G_{\eta_0}^{-1} \) is only one for \( P = 1 \) and zero otherwise. Therefore, in practice \( G_{\eta_0} \) has to be set to a lower limit, e.g., \( 20 \log_{10} G_{\eta_0} = -25 \text{ dB} \) 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, \( G_{\text{min}} \), is used on the overall gain function resulting in the clean speech estimate

\[
\tilde{S} = \max\{\tilde{G}, G_{\text{min}}\} \cdot Y. \tag{20}
\]

A higher value of \( G_{\text{min}} \) 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 \( \tilde{P} \) and \( G_{\eta_0} \) that exhibit as few statistical outliers as possible while introducing minimal speech distortion, so that lower choices for \( G_{\text{min}} \) are possible.

VI. EVALUATION

In this section, we apply the SPP \( \tilde{P} \) to a speech enhancement filter. Here, \( \tilde{P} \) 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].

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., \( Y = \tilde{N} \). 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 \( |\tilde{S}|/\sigma_N \) are given. Note that in the experiment of Figure 11, \( \tilde{S} \) 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

<table>
<thead>
<tr>
<th>window</th>
<th>overlap</th>
<th>( \Delta k_X )</th>
<th>( \Delta f )</th>
<th>( N_X )</th>
<th>( c_{\text{dir}}(\chi) )</th>
<th>( T_X )</th>
<th>( \xi_{\text{SPP}}(\chi) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>75%</td>
<td>local</td>
<td>1</td>
<td>4</td>
<td>15</td>
<td>0.34</td>
<td>10.2</td>
<td>8 dB</td>
</tr>
<tr>
<td></td>
<td>global</td>
<td>8</td>
<td>4</td>
<td>85</td>
<td>0.29</td>
<td>49.3</td>
<td>3 dB</td>
</tr>
<tr>
<td>50%</td>
<td>local</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>0.50</td>
<td>10.7</td>
<td>8 dB</td>
</tr>
<tr>
<td></td>
<td>global</td>
<td>8</td>
<td>2</td>
<td>51</td>
<td>0.50</td>
<td>51.3</td>
<td>3 dB</td>
</tr>
</tbody>
</table>

TABLE I

The parameter settings for the proposed SPP estimator with \( T = 64 \text{ ms} \), \( f_{\text{local}} = 105.5 \text{ Hz} \) and \( f_{\text{global}} = 543 \text{ Hz} \). \( \chi \) stands for either the local or the global average. The parameters are determined for a Hann window with 75% overlap and 50% overlap, respectively.
values of $\alpha$ are considered for (9). For lower values of $\alpha$ 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 $|\hat{S}|/\sigma N$. 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, $S(k, l) = G_{\text{min}}Y(k, l)$.

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 $G_{\text{min}}$, 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 $\alpha$ [3], and musical noise.

In the next step of the evaluation, the soft-gain factors, $\hat{P}$, 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 $\hat{P} > 0.5$ in noise-only spectral regions, as compared to Figure 12(c). In Figure 12(c), false-alarmers cause dark speckles in the gray spectral regions where no speech is present.

Inspired by [17] we evaluate the SPP estimators, $\hat{P}$, 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, $P_{\text{id}}(k, l)$, 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:

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

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

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

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

$$N_{\text{id}}(k, l) = (1 - P_{\text{id}}(k, l)) N(k, l) \quad (24)$$

After taking the inverse Fourier transform and reconstructing the time signal by overlapping and adding the signal segments.
we get the time domain signals $e_{SD}(t)$, $e_{NL}(t)$, $s_{id}(t)$ and $n_{id}(t)$. The final measures for speech distortion and noise leakage are then gained as:

$$\begin{align*}
SD &= \sum_t \frac{e_{SD}^2(t)}{\sum_t s_{id}^2(t)} \\
NL &= \sum_t \frac{e_{NL}^2(t)}{\sum_t n_{id}^2(t)}.
\end{align*}$$

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, $\mathcal{P}_{id}$, are attenuated by $\mathcal{P}$ and SD = 0% if $\mathcal{P}(k, l) = 1$ wherever $\mathcal{P}_{id}(k, l) = 1$. The NL equals 0% if $\mathcal{P}(k, l) = 0$ 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 $\tilde{s}(t)$ the enhanced signal in the time domain, obtained from $\tilde{S}(k, l)$, the segmental SNR, $\text{SNR}_{\text{seg}} \{\tilde{s}(t)\}$, compares the clean speech signal, $s(t)$, and its estimate, $\tilde{s}(t)$, on a segmental basis:

$$\text{SNR}_{\text{seg}} \{\tilde{s}(t)\}$$

$$\begin{align*}
&= \frac{10}{|L_{\text{seg}}|} \sum_{t \in L_{\text{seg}}} \log_{10} \left( \frac{\sum_{k=0}^{L-1} s^2(\hat{L} + t)}{\sum_{k=0}^{L-1} (s(\hat{L} + t) - \tilde{s}(\hat{L} + t))^2} \right),
\end{align*}$$

where $L_{\text{seg}}$ is the set of frame indices $\hat{L}$ that belong to non-overlapping speech active frames of an utterance $s(t)$. The length of each frame is set to $\hat{L}/f_s = 10\text{ ms}$. Note that $\text{SNR}_{\text{seg}} \{\tilde{s}(t)\}$ 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 $\mathcal{P}$ 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, $\Delta \text{SNR}_{\text{seg}} = \text{SNR}_{\text{seg}} \{\tilde{s}(t)\} - \text{SNR}_{\text{seg}} \{y(t)\}$, 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 $\sigma(\kappa, \lambda)$ given by

$$\text{var}\{\tau\} = \frac{2}{\tau}, \quad (28)$$

with $\tau$ the degrees of freedom of $\tau$. 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 $c_{dof} \leq 1$ giving

$$\tau = 2c_{dof}N \cdot \quad (29)$$

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

$$c_{dof} = \frac{1}{N \text{var}\{\tau\}}. \quad (30)$$

Thus, the correction factor $c_{dof}$ can be empirically determined with (30) by measuring $\text{var}\{\tau\}$ 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

The results in terms of speech distortion, SD, noise leakage, NL, and segmental SNR improvement, $\Delta SNR_{seg}$. The proposed method yields a good trade-off between SD and NL and slight improvements in terms of $\Delta SNR_{seg}$.

### Table II

<table>
<thead>
<tr>
<th></th>
<th>-10 dB</th>
<th>-5 dB</th>
<th>0 dB</th>
<th>5 dB</th>
<th>10 dB</th>
<th>15 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SD [%]</td>
<td>NL [%]</td>
<td>$\Delta SNR_{seg}$ [dB]</td>
<td>SD [%]</td>
<td>NL [%]</td>
<td>$\Delta SNR_{seg}$ [dB]</td>
</tr>
<tr>
<td>white</td>
<td>2.8</td>
<td>33.1</td>
<td>11.8</td>
<td>1.3</td>
<td>33.4</td>
<td>9.6</td>
</tr>
<tr>
<td>pink</td>
<td>3.1</td>
<td>32.5</td>
<td>11.0</td>
<td>1.4</td>
<td>32.8</td>
<td>8.8</td>
</tr>
<tr>
<td>car</td>
<td>0.2</td>
<td>46.5</td>
<td>14.0</td>
<td>0.1</td>
<td>45.7</td>
<td>13.5</td>
</tr>
<tr>
<td>bubble</td>
<td>2.5</td>
<td>66.2</td>
<td>5.1</td>
<td>1.1</td>
<td>66.2</td>
<td>4.2</td>
</tr>
<tr>
<td>traffic</td>
<td>2.7</td>
<td>58.9</td>
<td>9.3</td>
<td>1.2</td>
<td>59.4</td>
<td>7.6</td>
</tr>
<tr>
<td>subway</td>
<td>2.7</td>
<td>59.5</td>
<td>8.1</td>
<td>1.2</td>
<td>59.2</td>
<td>6.9</td>
</tr>
</tbody>
</table>

### References


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).

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, N.J, where he worked on artificial bandwidth extension. His main research interests are digital speech and audio processing, including single- and multichannel speech enhancement.