NOISE POWER ESTIMATION BASED ON THE PROBABILITY OF SPEECH PRESENCE

Timo Gerkmann[*]

Sound and Image Processing Lab.
KTH Royal Institute of Technology
100 44 Stockholm, Sweden
gerkmann@kth.se

Richard C. Hendriks[†]

Signal and Information Processing Lab.
Delft University of Technology
2628 CD Delft, The Netherlands
R.C.Hendriks@tudelft.nl

ABSTRACT

In this paper, we analyze the minimum mean square error (MMSE) based spectral noise power estimator [1] and present an improvement. We will show that the MMSE based spectral noise power estimate is only updated when the *a posteriori* signal-to-noise ratio (SNR) is lower than one. This threshold on the *a posteriori* SNR can be interpreted as a voice activity detector (VAD).

We propose in this work to replace the hard decision of the VAD by a soft speech presence probability (SPP). We show that by doing so, the proposed estimator does not require a bias correction and safety-net as is required by the MMSE estimator presented in [1]. At the same time, the proposed estimator maintains the quick noise tracking capability which is characteristic for the MMSE noise tracker, results in less noise power overestimation and is computationally less expensive.

Index Terms— Noise power estimation, speech enhancement, noise reduction.

1. INTRODUCTION

Portable digital communication devices, such as hearing aids or mobile telephones, are often used in noisy environments. The noise signal that corrupts the target speech signal can be locally quite nonstationary. Nonstationary noise corruptions can be caused for example by passing cars when communicating while walking along the street, or babble noise while in the cafeteria or at a party. Speech enhancement algorithms aim at reducing the additive noise while keeping the target speech signal unaffected. One of the most important parameters of speech enhancement algorithms is the spectral noise power. The spectral noise power can be estimated whenever we know that speech is absent. However, in nonstationary noise scenarios the estimation of the noise power is particularly difficult as the it may change rapidly over time. Then, the estimated noise power has to be updated as often as possible, requiring a robust voice activity detector (VAD). However, deciding whether speech is present or absent is more difficult the more nonstationary the noise source is, as a sudden rise in the noise power may be misinterpreted as a speech onset.

Several approaches have been proposed for the estimation of the noise power. Among the most established estimators are those based on minimum statistics [2], [3]. For instance, in [2] the power of the noisy signal is estimated and observed over a time-span of about 1-3 seconds. The spectral noise power is then inferred from the minimum of the estimated power of the noisy signal, assuming that speech is absent at least for a short duration within the observed time-span. However, if the noise power rises within the observed time-span, the noise power will be underestimated. While in [2] mechanisms are proposed that allow for a tracking of rising noise powers within the observed time-span, rising noise powers as caused e.g. by passing cars, are usually tracked with a rather large delay. The local underestimation of the noise power is likely to result in annoying artifacts, so-called *musical noise*, when the noise power estimate is applied in a speech enhancement framework.

More recent spectral noise power estimators allow for a quicker tracking of the noise spectral power, e.g. the subspace-DFT approach [4], or minimum mean square error (MMSE) based approaches [5], [1]. While subspace based approaches are computationally rather demanding, the MMSE based algorithm [1] is computationally much less demanding and at the same time robust to increasing noise levels [6]. In the MMSE based estimator [1], first a limited maximum likelihood (ML) estimate of the *a priori* signal-to-noise ratio (SNR) is used to estimate the periodogram of the noise signal. However, this simple estimate results in a bias, which is then compensated based on a second estimate of the *a priori* SNR. In this work, we analyze the MMSE based estimator presented in [1] and present an improvement that makes the bias compensation unnecessary.

This work is organized as follows: after explaining the notations and assumptions in Section 2, we show in Section 3 that the MMSE based noise power estimator of [1] can be interpreted as a VAD based noise power estimator, where the noise power estimate is only updated if the *a posteriori* SNR is smaller than one. Then, in Section 4 we propose to replace the VAD of [1] by a soft speech presence probability (SPP), without the need of applying a bias compensation. In Section 5 we show that the proposed estimator results in a similar noise tracking performance as the estimator in [1], while being computationally and memory-wise more efficient.

2. SIGNAL MODEL

We assume the speech and noise signals to be additive in the short-time Fourier domain. The complex spectral noisy observation is thus given by $Y_k(l) = S_k(l) + N_k(l)$, where k is the frequency index, l is the segment index, S_k are the complex spectral speech coefficients and N_k are the complex spectral noise coefficients. For each k, the spectral speech and noise power are defined as $\sigma^2_{\mathrm{S},k}(l) = \mathrm{E}\big(|S_k(l)|^2\big)$ and $\sigma^2_{\mathrm{N},k}(l) = \mathrm{E}\big(|N_k(l)|^2\big)$, respectively. In the sequel, we omit the time and frequency index wherever possible. We define the *a posteriori* SNR as $\gamma = |Y|^2/\sigma^2_{\mathrm{N}}$ and the *a priori* SNR

[*]The research leading to these results has received funding from the European Community's Seventh Framework Programme under Grant Agreement PIAP-GA-2008-214699.

[†]The research is supported by the Dutch Technology Foundation STW.

as $\xi = \sigma_{\mathrm{S}}^2/\sigma_{\mathrm{N}}^2$. We assume that the speech and noise signals are uncorrelated and have zero mean so that $\mathrm{E}\big(|Y|^2\big) = \sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{N}}^2$. In addition, we assume that the real and imaginary part of the noise and speech spectral coefficients are independent and Gaussian distributed. Furthermore, estimated quantities are denoted by a hat symbol, e.g. $\widehat{\xi}_k$ is the estimate of ξ_k.

3. REVIEW OF MMSE BASED NOISE POWER ESTIMATION

In [5], [1] it is proposed to estimate the spectral noise power from an MMSE estimate of the noise periodogram. Given an estimate of the *a priori* SNR ξ and an estimate of the noise power σ_{N}^2, the estimate of the noise periodogram is obtained as

$$|\widehat{N}|^2 = \mathrm{E}\big(|N|^2 \mid Y\big) = \left(\frac{1}{1+\widehat{\xi}}\right)^2 |Y|^2 + \frac{\widehat{\xi}}{1+\widehat{\xi}}\,\widehat{\sigma}_{\mathrm{N}}^2. \quad (1)$$

Assuming that the spectral noise power does not change abruptly from one signal segment to the other, it is reasonable to employ the noise power estimate of the previous frame $\widehat{\sigma}_{\mathrm{N}}^2 = \widehat{\sigma}_{\mathrm{N}}^2(l-1)$ in (1). However, as the speech signal may change quickly over time, finding an appropriate estimate for the *a priori* SNR in (1) is rather difficult. In [1] a limited ML is employed as detailed in Section 3.1.

After estimating the noise periodogram via (1), the noise power spectral density is updated by a recursive smoothing, as

$$\widehat{\sigma}_{\mathrm{N}}^2(l) = \alpha\,\widehat{\sigma}_{\mathrm{N}}^2(l-1) + (1-\alpha)\,|\widehat{N}(l)|^2, \quad (2)$$

where, as in [1], we choose $\alpha = 0.8$.

Taking the expectation of (1) with respect to Y, we obtain

$$\mathrm{E}_Y\big(\mathrm{E}\big(|N|^2 \mid Y, \widehat{\sigma}_{\mathrm{N}}^2, \widehat{\sigma}_{\mathrm{S}}^2\big)\big) = \left(\frac{\widehat{\sigma}_{\mathrm{N}}^2}{\widehat{\sigma}_{\mathrm{S}}^2 + \widehat{\sigma}_{\mathrm{N}}^2}\right)^2 (\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{N}}^2) + \frac{\widehat{\sigma}_{\mathrm{S}}^2}{\widehat{\sigma}_{\mathrm{S}}^2 + \widehat{\sigma}_{\mathrm{N}}^2}\widehat{\sigma}_{\mathrm{N}}^2, \quad (3)$$

where we now explicity state that the estimator requires knowing $\widehat{\sigma}_{\mathrm{N}}^2$ and $\widehat{\sigma}_{\mathrm{S}}^2$. From (3) it follows that if $\widehat{\sigma}_{\mathrm{S}}^2 = \sigma_{\mathrm{S}}^2$ and $\widehat{\sigma}_{\mathrm{N}}^2 = \sigma_{\mathrm{N}}^2$ (1) is unbiased and we have $\mathrm{E}_Y\big(\mathrm{E}\big(|N|^2 \mid Y, \sigma_{\mathrm{N}}^2, \sigma_{\mathrm{S}}^2\big)\big) = \sigma_{\mathrm{N}}^2$. On the other hand, if $\widehat{\sigma}_{\mathrm{S}}^2 \neq \sigma_{\mathrm{S}}^2$ and/or $\widehat{\sigma}_{\mathrm{N}}^2 \neq \sigma_{\mathrm{N}}^2$ the estimator is biased, and we have $\mathrm{E}_Y\big(\mathrm{E}\big(|N|^2 \mid Y, \widehat{\sigma}_{\mathrm{N}}^2, \widehat{\sigma}_{\mathrm{S}}^2\big)\big) \neq \sigma_{\mathrm{N}}^2$. However, to compensate for the bias, again the true noise and speech spectral power are required.

3.1. Interpretation as a voice activity detector

In [1] it is proposed to employ a limited ML estimate of the *a priori* SNR in (1). In this section we show that by this the MMSE estimate of the noise periodogram is only updated when the *a posteriori* SNR is smaller than 1. This thresholding of the *a posteriori* SNR can be interpreted as a VAD; the spectral noise power is only updated when there is speech absence according to the *a posteriori* SNR.

In the way we wrote (1), it can be seen that the MMSE solution results in a weighted sum of the noisy observation and the previous estimate of the spectral noise power $\widehat{\sigma}_{\mathrm{N}}^2$. The weights are a function of the *a priori* SNR and gradually take values between zero and one, i.e., a *soft* decision between $|Y|^2$ and $\widehat{\sigma}_{\mathrm{N}}^2$. However, in [1] a limited ML estimate of the *a priori* SNR is employed, which is obtained, as

$$\widehat{\xi} = \max\Big(0, \widehat{\xi}^{\,\mathrm{ml}}\Big) = \max(0, \widehat{\gamma} - 1)\,, \quad (4)$$

where $\widehat{\gamma}(l) = |Y(l)|^2/\widehat{\sigma}_{\mathrm{N}}^2(l-1)$. Substituting (4) and $\widehat{\sigma}_{\mathrm{N}}^2 = \widehat{\sigma}_{\mathrm{N}}^2(l-1)$ into (1) we see that the MMSE estimator can be seen as a VAD based detector, as

$$|\widehat{N}(l)|^2 = \mathrm{E}\big(|N(l)|^2 \mid Y(l)\big) = \begin{cases} \widehat{\sigma}_{\mathrm{N}}^2(l-1) & ,\text{if } \widehat{\gamma}(l) \geq 1 \\ |Y(l)|^2 & ,\text{if } \widehat{\gamma}(l) < 1, \end{cases} \quad (5)$$

Using the *a priori* SNR estimator from (4) we thus have a *hard* instead of a soft decision between the noisy observation and the estimate of the spectral noise power $\widehat{\sigma}_{\mathrm{N}}^2(l-1)$.

In [1] the bias is derived for the case that the limited ML estimate of (4) is employed. However, the resulting bias again depends on the true and unknown *a priori* SNR. In [1] this *a priori* SNR is estimated using the decision-directed approach [7].

3.2. Safety-net

In addition to the bias compensation, in [1] a so-called *safety-net* is employed to prevent the spectral noise power tracker from stalling when the noise level would make an abrupt step from one segment to the next. In this safety-net, the last 0.8 seconds of the noisy speech periodogram, i.e. 50 signal segments $|Y(l)|^2$, are stored. The final estimate of the spectral noise power is obtained by comparing the current noise power estimate to the minimum of the last 0.8 seconds of $|Y(l)|^2$, as

$$\widehat{\sigma}_{\mathrm{N}}^2 \leftarrow \max\big(\widehat{\sigma}_{\mathrm{N}}^2, \min\big(|Y(l-49)|^2, ..., |Y(l)|^2\big)\big)\,. \quad (6)$$

4. PROPOSED APPROACH: SPP INSTEAD OF VAD

Instead of first using a limited ML estimate for the *a priori* SNR that results in the VAD behavior explained by (5), we argue in this paper that neither a bias compensation nor the safety-net of Section 3.2 is necessary if the hard decision of the VAD (5) is exchanged by a soft decision by means of the probability of speech presence.

Under speech presence uncertainty, an MMSE estimator for the noise periodogram is given by

$$\mathrm{E}\big(|N|^2 \mid Y\big) = P(\mathcal{H}_0 \mid Y)\,\mathrm{E}\big(|N|^2 \mid Y, \mathcal{H}_0\big) + P(\mathcal{H}_1 \mid Y)\,\mathrm{E}\big(|N|^2 \mid Y, \mathcal{H}_1\big), \quad (7)$$

where $\mathcal{H}_0$ indicates speech absence, while $\mathcal{H}_1$ indicates speech presence.

4.1. Estimation of the speech presence probability

As for the derivation of (1), we assume that the real and imaginary parts of the speech and noise spectral coefficients are Gaussian distributed. With Bayes' theorem, assuming uniform priors $P(\mathcal{H}_0) = P(\mathcal{H}_1)$, follows the probability of speech presence, e.g. [8]

$$P(\mathcal{H}_1 \mid Y) = \left(1 + (1+\xi_{\mathrm{opt}})\exp\left(-\frac{|Y|^2}{\widehat{\sigma}_{\mathrm{N}}^2}\,\frac{\xi_{\mathrm{opt}}}{1+\xi_{\mathrm{opt}}}\right)\right)^{-1}. \quad (8)$$

While in (1) $\widehat{\xi}$ is the local SNR, in (8) the *a priori* SNR ξ_{opt} reflects the SNR that is typical if speech were present [9]. In the radar or communication context, one would choose ξ_{opt} in order to guarantee a specified performance in terms of false alarms or missed detections [10]. Similarly, we find the fixed optimal *a priori* SNR $10\log_{10}(\xi_{\mathrm{opt}}) = 15\,\mathrm{dB}$ by minimizing the total probability of error when the true *a priori* SNR lies between $-\infty$ and $20\,\mathrm{dB}$, as detailed in [9].

4.2. Derivation of $\mathrm{E}\big(|N|^2|Y,\mathcal{H}_0\big)$ and $\mathrm{E}\big(|N|^2|Y,\mathcal{H}_1\big)$

From (8) it is possible to derive an expression for the *a posteriori* SNR $\gamma = |Y|^2/\widehat{\sigma}_{\mathrm{N}}^2$ in terms of ξ_{opt} and $P(\mathcal{H}_1 \mid Y)$, that is,

$$\gamma = \log\left(\frac{1+\xi_{\mathrm{opt}}}{P(\mathcal{H}_1 \mid Y)^{-1}-1}\right)\frac{1+\xi_{\mathrm{opt}}}{\xi_{\mathrm{opt}}}. \tag{9}$$

From this expression it follows that already for $P(\mathcal{H}_1 \mid Y) > 0.075$, the *a posteriori* SNR satisfies $\gamma > 1$ if $10\log_{10}(\xi_{\mathrm{opt}}) = 15\,\mathrm{dB}$. From this it can be concluded that under speech presence, i.e., when $P(\mathcal{H}_1 \mid Y)$ is sufficiently high, the ML estimate of the *a priori* SNR from (4) can be rewritten as $\widehat{\xi}^{\mathrm{ml}} = \widehat{\gamma} - 1$. The optimal estimator under speech presence can now be computed as

$$\mathrm{E}\Big(|N|^2 \mid Y,\widehat{\xi},\mathcal{H}_1\Big) = \mathrm{E}\Big(|N|^2 \mid Y,\widehat{\xi}^{\mathrm{ml}} = \widehat{\gamma}-1\Big) = \widehat{\sigma}_{\mathrm{N}}^2,$$

which follows from substitution of $\widehat{\xi}^{\mathrm{ml}} = \widehat{\gamma} - 1$ into (1). Under speech absence we have $Y = N$ and thus $\mathrm{E}\big(|N|^2 \mid Y,\mathcal{H}_0\big) = \mathrm{E}\big(|N|^2 \mid N\big) = |N|^2 = |Y|^2$. Then, similar to (1), we obtain

$$|\widehat{N}|^2 = \mathrm{E}\big(|N|^2 \mid Y\big) = P(\mathcal{H}_0 \mid Y)\,|Y|^2 + P(\mathcal{H}_1 \mid Y)\,\widehat{\sigma}_{\mathrm{N}}^2, \tag{10}$$

where $P(\mathcal{H}_0 \mid Y) = 1 - P(\mathcal{H}_1 \mid Y)$ and we employ the spectral noise power estimated of the previous frame $\widehat{\sigma}_{\mathrm{N}}^2$. The spectral speech power is then obtained by a recursive smoothing of $|\widehat{N}|^2$ as given in (2).

4.3. Avoiding stagnation

From (8) it can be seen that if the spectral noise power is underestimated, it may occur that $P(\mathcal{H}_1 \mid Y) = 1$ even though $|Y|^2$ is small with respect to the true, but unkown, noise power. Then, due to (10), the noise power may not be updated anymore, such that the noise power remains underestimated. To check and overcome that this happens, we recursively smooth $P(\mathcal{H}_1 \mid Y)$ over time by,

$$\overline{\mathcal{P}}(l) = 0.9\,\overline{\mathcal{P}}(l-1) + 0.1\,P(\mathcal{H}_1 \mid Y(l))\,, \tag{11}$$

and force the current estimate $P(\mathcal{H}_1 \mid Y)$ to be smaller than one, if $\overline{\mathcal{P}}(l)$ is larger than a threshold, as

$$P(\mathcal{H}_1 \mid Y(l)) \leftarrow \begin{cases} \min(0.99, P(\mathcal{H}_1 \mid Y(l))) & ,\overline{\mathcal{P}}(l) > 0.99 \\ P(\mathcal{H}_1 \mid Y(l)) & ,\text{else.} \end{cases} \tag{12}$$

This procedure fits well into the framework and is more memory efficient than the safety-net of Section 3.2 as we do not need to store 0.8 seconds of data. The proposed SPP based algorithm is summarized in Algorithm 1.

In Section 5 we show that the proposed approach results in slightly better results than the estimator proposed in [1], but does not require a bias correction and requires less memory storage.

Algorithm 1 The proposed algorithm for noise power estimation.

1: **for all** signal segments l **do**

2: Compute the *a posteriori* SPP
$P(\mathcal{H}_1 \mid Y) = \Big(1+(1+\xi_{\mathrm{opt}})\exp\Big(-\frac{|Y|^2}{\widehat{\sigma}_{\mathrm{N}}^2}\,\frac{\xi_{\mathrm{opt}}}{1+\xi_{\mathrm{opt}}}\Big)\Big)^{-1}$
where $\widehat{\sigma}_{\mathrm{N}}^2$ is the noise power estimate of the previous frame and $10\log_{10}(\xi_{\mathrm{opt}}) = 15\,\mathrm{dB}$.

3: Compute a smoothed *a posteriori* SPP, as
$\overline{\mathcal{P}}(l) = 0.9\,\overline{\mathcal{P}}(l-1) + 0.1\,P(\mathcal{H}_1 \mid Y_k(l))\,.$

4: Avoid stagnation, as
$P(\mathcal{H}_1 \mid Y(l)) \leftarrow \begin{cases} \min(0.99, P(\mathcal{H}_1 \mid Y(l))) & ,\overline{\mathcal{P}}(l) > 0.99 \\ P(\mathcal{H}_1 \mid Y(l)) & ,\text{else.} \end{cases}$

5: Update the noise periodogram estimate as
$|\widehat{N}|^2 = P(\mathcal{H}_0 \mid Y)\,|Y|^2 + P(\mathcal{H}_1 \mid Y)\,\widehat{\sigma}_{\mathrm{N}}^2.$

6: Obtain spectral noise power estimate by temporal smoothing
$\widehat{\sigma}_{\mathrm{N}}^2(l) = 0.8\,\widehat{\sigma}_{\mathrm{N}}^2(l-1) + 0.2\,|\widehat{N}(l)|^2.$

7: **end for**

5. EVALUATION

In this section, we compare the proposed spectral noise power estimators to the minimum statistics approach [2] and the MMSE approach with bias compensation proposed in [1]. For the evaluation we employ 320 sentences from the TIMIT database [11] and several synthetic and natural noise sources. In these evaluations we set the sampling rate at $f_{\mathrm{s}} = 16\,\mathrm{kHz}$. Further, we use a Hann-window of length $N = 512$ for spectral analysis, where successive segments overlap by 50%.

As proposed in [12] we compare the estimated noise power $\widehat{\sigma}_{\mathrm{N},k}^2$ to a reference $\sigma_{\mathrm{N},k}^2$ in terms of the log-error distortion measure. In contrast to [12], we separate the error measure into over and under estimation, i.e.

$$\mathrm{LogErr} = \mathrm{LogErrOver} + \mathrm{LogErrUnder}, \tag{13}$$

where LogErrOver measures the contributions of an overestimation of the true noise power, as

$$\mathrm{LogErrOver} = \frac{1}{NL}\sum_{l=0}^{L-1}\sum_{k=0}^{N-1}\left|\min\left(0, 10\log_{10}\left(\frac{\sigma_{\mathrm{N},k}^2(l)}{\widehat{\sigma}_{\mathrm{N},k}^2(l)}\right)\right)\right|,$$

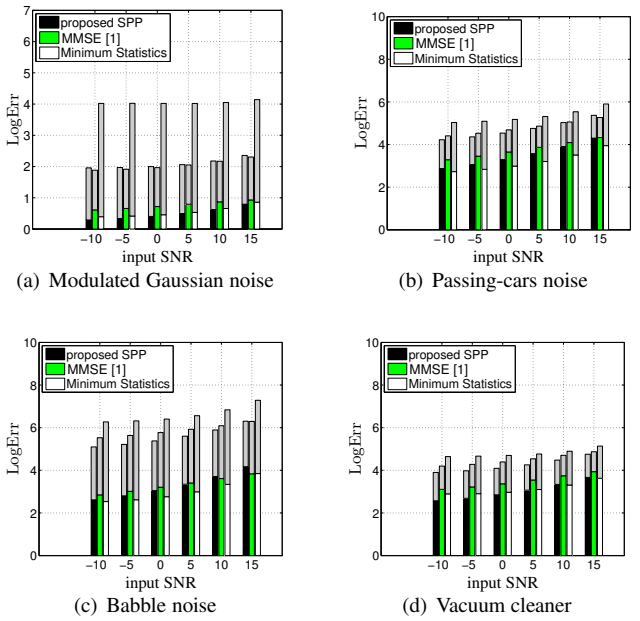
while LogErrUnder measures the contributions of an underestimation of the true noise power, as

$$\mathrm{LogErrUnder} = \frac{1}{NL}\sum_{l=0}^{L-1}\sum_{k=0}^{N-1}\max\left(0, 10\log_{10}\left(\frac{\sigma_{\mathrm{N},k}^2(l)}{\widehat{\sigma}_{\mathrm{N},k}^2(l)}\right)\right).$$

Note that an overestimation of the true noise power, as indicated by LogErrOver, is likely to result in an attenuation of the speech signal in a speech enhancement framework and thus in speech distortions. On the other hand, at time-frequency points where the noise power is underestimated, the noise signal is not reduced to the same extend as for the true noise power. Furthermore, if the noise power is underestimated locally, isolated time-frequency points are not attenuated, which may result in annoying artifacts perceived as so-called *musical noise*.

As noise types, we consider modulated white Gaussian noise, passing cars noise, nonstationary vacuum cleaner noise, and babble noise. The modulated noise is modulated with $f_{\mathrm{mod}} = 0.5\,\mathrm{Hz}$ using the function $f(m) = 1 + 0.5\,\sin(2\pi m f_{\mathrm{mod}}/f_{\mathrm{s}})$, where m is the sample index.

For the synthetic modulated white noise, the true noise power is known and is thus used for the evaluation. For the remaining nonstationary and thus non-ergodic noise sources the determination of the true spectral noise power is impossible, as only one realization



(a) Modulated Gaussian noise (b) Passing-cars noise

(c) Babble noise (d) Vacuum cleaner

Figure 1: Comparison in terms of the LogErr for 320 TIMIT sentences and various input SNRs. The lower part of the bars represents the amount of noise power overestimation LogErrOver, while the upper part represents the noise underestimation LogErrUnder. The total height of the bars corresponds to LogErr.

of the random variable is available in each time-frequency point. Therefore, we use the periodogram of the noise-only signal as an estimate of the true noise power, e.g. $\sigma^2_{\mathrm{N},k} = |N|^2$.

The results of our evaluation are given in Figure 1. It can be seen that for the modulated white Gaussian noise, the minimum statistics approach is not able to follow the rapid changed of the noise signal, resulting in a large amount of noise underestimation that is likely to result in musical noise in a speech enhancement framework. For the natural noise sources we considered, passing car noise, babble noise and vacuum cleaner noise, this effect is not as dramatic as for the synthetic modulated Gaussian noise, but still, the minimum statistics approach results in the largest noise underestimation, and is thus likely to result in the largest amount of musical noise. Comparing the proposed SPP based estimator to the MMSE based estimator [1] it can be seen that the overall performance in terms of the LogErr is rather similar. However, the MMSE [1] approach has the tendency to overestimate the noise power most. This may be because the bias compensation factor multiplied to the estimated noise signal in [1] is always larger or equal to one, even when the noise power estimate of the previous frame was overestimated.

While obtaining similar results in terms of the LogErr, the proposed SPP based estimator is more computationally and memory efficients, as we do not require to store 0.8 seconds of data for the safety-net of Section 3.2, nor do we need to compute the incomplete gamma function necessary for the bias compensation in [1].

The code for this noise power estimator is available at `www.ee.kth.se/~gerkmann/sppBasedNoisePow`.

6. CONCLUSIONS

In this work, we have refined the minimum mean square error (MMSE) based noise power estimator [1]. We have shown that when a limited maximum likelihood (ML) estimator is used for the estimation of the *a priori* signal-to-noise ratio (SNR), the resulting noise power estimate is only updated when the *a posteriori* SNR is below a certain threshold. We have argued that this thresholding can be interpreted as a voice activity detector (VAD). In addition, in order to function properly, the estimator in [1] requires a bias compensation and a so-called safety-net that requires storing the last 0.8 seconds of data.

In this paper we have shown that the bias compensation and the safety-net are unnecessary if the hard decision of the VAD is replaced by a soft speech presence probability (SPP) estimator. The proposed estimator is more memory efficient and results slightly better performance than the estimator from [1].

7. REFERENCES

[1] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *IEEE ICASSP*, pp. 4266–4269, Mar. 2010.

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

[3] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[4] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 16, no. 3, pp. 541–553, March 2008.

[5] R. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," in *IEEE ICASSP*, 2009, pp. 4421–4424.

[6] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," *IEEE ICASSP*, May 2011.

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

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

[9] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 16, no. 5, pp. 910–919, July 2008.

[10] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

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

[12] R. C. Hendriks, J. Jensen, and R. Heusdens, "DFT domain subspace based noise tracking for speech enhancement," in *Interspeech*, August 2007, pp. 830–833.