UTILIZING SPECTRO-TEMPORAL CORRELATIONS FOR AN IMPROVED SPEECH PRESENCE PROBABILITY BASED NOISE POWER ESTIMATION

Martin Krawczyk-Becker, Dörte Fischer, Timo Gerkmann

Speech Signal Processing Group, Department of Medical Physics and Acoustics
and Cluster of Excellence "Hearing4all"
University of Oldenburg, Germany
{martin.krawczyk, doerte.fischer, timo.gerkmann}@uni-oldenburg.de

ABSTRACT

For the enhancement of speech degraded by noise, accurate estimation of the noise power spectral density (PSD) is indispensable, especially if only a single microphone signal is available. Fast and accurate tracking of the noise PSD is particularly challenging in highly non-stationary noise types, since the distinction between speech and noise components becomes more difficult. Short-time discrete Fourier transform (STFT) based noise PSD estimation algorithms which employ estimates of the speech presence probability (SPP) with fixed priors have been shown to yield good tracking performance even in adverse noise conditions. In this paper, we compare two methods to incorporate spectro-temporal correlations to improve the tracking performance. The first method smoothes the noisy observation over time and frequency before computing the SPP, while the second is based on a Hidden Markov Model (HMM) of the speech presence and absence states. We show that the proposed modifications lead to improved noise PSD estimators which are less sensitive to spectral outliers of the noise and track changes in the noise PSD more quickly than the reference method. Further, when employed in a common speech enhancement setup, the proposed estimators achieve an increased noise reduction while keeping speech distortions at a comparable level.

***Index Terms*—** speech enhancement, noise reduction, noise power estimation

1. INTRODUCTION

For the reduction of additive noise, most single-channel speech enhancement approaches first transform the noisy speech signal into some spectro-temporal domain, most frequently the STFT domain. Then a multiplicative gain is applied to every time-frequency point, which allows for a frequency selective suppression of the noise while maintaining the speech component. To achieve this, accurate estimates of the spectral speech and noise PSD are indispensable. Over the years, several approaches for the estimation of the noise PSD have been proposed, of which the simplest limit the estimation to speech pauses based on a voice activity detector (VAD). Such approaches are however not capable of tracking changes in the noise PSD during speech activity. This limitation is to some degree overcome by the approach of Martin [1], where the noise PSD in each frequency band is inferred from spectral minima of the noisy observation in a search window of typically more than a second. Since even during speech activity the minima in frequency bands are assumed to correspond to the noise component, the estimation of the noise PSD is not limited to speech pauses anymore. Sudden rises in the noise PSD however remain a problem and are tracked only with a delay of the length of the search window, leading to suboptimal noise suppression. Spectral minima are also utilized in [2], where the noise PSD is estimated by recursively averaging the squared noisy spectrum. The averaging parameter is adapted according to an SPP estimate based on spectral minima. However, as compared to [1], the improvement in the tracking performance is limited.

Evolving from [3], a low complexity alternative for minimum mean square error (MMSE) based noise PSD estimation has been proposed in [4]. For this approach, a frequency dependent estimate of the SPP is required. The lower the probability of speech presence, the more the respective time-frequency point contributes to the noise PSD estimation. This approach has been shown to work reliably also in highly non-stationary noise types. In [4], the SPP estimation is performed based on fixed priors, as proposed in [5]. Although [5] allows to estimate the SPP based on a smoothed observation to reduce the influence of random outliers in the noise, in [4] the SPP is estimated independently for every time-frequency point to minimize the computational complexity of the algorithm. In this paper, we investigate if and how the incorporation of temporal and spectral correlations in the SPP estimation are beneficial for noise PSD estimation. For this, we consider two different approaches. First, we apply spectro-temporal smoothing to the noisy observation before estimating the SPP based on it according to [5]. Second, we utilize information of neighboring time-frequency points by means of a two dimensional HMM as proposed in [6], where one dimension covers the temporal correlations while the other covers correlations along frequency bands. In contrast to the first approach, where correlations between adjacent bins are utilized in terms of a pre-processing stage, the 2D HMM utilizes correlations directly in the SPP estimation procedure. Furthermore, in [6] it has been shown that this approach has the potential to outperform the spectro-temporal smoothing of e.g. [5] in terms of speech detection accuracy. The price to pay is an increased algorithmic delay and computational complexity.

Instead of the instantaneous SPP estimate in [4], here we employ the SPP estimators [5] and [6] for the estimation of the noise PSD. The noise PSD estimation performance is evaluated within a complete speech enhancement setup. Further, the capability to track non-stationary noises is assessed. We compare the results of the proposed noise PSD estimators that incorporate spectro-temporal correlations to the initial method [4] and to the minimum statistics approach [1].

This work was supported by grant GE2538/2-1 of the German Research Foundation (DFG).

2. SPP BASED NOISE PSD ESTIMATION

In this section, we briefly review the SPP-based noise PSD estimation proposed in [4]. We denote the complex spectral STFT coefficients of the noisy observation as $Y_\ell^k = S_\ell^k + N_\ell^k$, where S_ℓ^k and N_ℓ^k represent the mutually independent speech and noise coefficients. Further, k and ℓ are the frequency bin index and the segment index, respectively, which will be dropped for notational convenience wherever dispensable. Random variables are denoted by capital letters, while for their realizations the corresponding lower case letters are used. We assume S and N to follow zero-mean complex Gaussian distributions with variances $E\left\{|S|^2\right\} = \sigma_{\mathrm{S}}^2$ and $E\left\{|N|^2\right\} = \sigma_{\mathrm{N}}^2$, where the operator $E\left\{\cdot\right\}$ denotes statistical expectation. To distinguish estimated quantities from the true values we use a hat symbol, e.g. $\widehat{\sigma_{\mathrm{N}}^2}$ is an estimate of σ_{N}^2.

To estimate the noise PSD, two states are distinguished, namely speech presence $\mathcal{H}_1$, i.e. $Y = S + N$, and speech absence $\mathcal{H}_0$, i.e. $Y = N$. Under speech presence uncertainty an estimator of the noise periodogram can be formulated as [4]

$$\widehat{|N|^2} = P(\mathcal{H}_0 \mid y)\,|y|^2 + P(\mathcal{H}_1 \mid y)\,\widehat{\sigma_{\mathrm{N}}^2}, \tag{1}$$

where the posterior probability of speech presence and absence are related via $P(\mathcal{H}_0 \mid y) = 1 - P(\mathcal{H}_1 \mid y)$. In case $P(\mathcal{H}_0 \mid y) = 1$, the current bin is assumed to contain only noise and $E\left\{|N|^2 \middle| y\right\} = |y|^2 = |n|^2$. For $P(\mathcal{H}_1 \mid y) = 1$, however, we observe a mixture of speech and noise and thus assume that a previous estimate of the noise PSD is more reliable than the current observation, i.e. $\widehat{|N|^2} = \widehat{\sigma_{\mathrm{N}}^2}$. For all values in between, (1) yields a weighted mixture of $|y|^2$ and $\widehat{\sigma_{\mathrm{N}}^2}$.

To estimate the a posteriori SPP $P(\mathcal{H}_1 \mid y)$ we can employ Bayes' formula:

$$P(\mathcal{H}_1 \mid y) = \frac{P(\mathcal{H}_1)\ p(y \mid \mathcal{H}_1)}{P(\mathcal{H}_1)\ p(y \mid \mathcal{H}_1) + P(\mathcal{H}_0)\ p(y \mid \mathcal{H}_0)}, \tag{2}$$

with the prior probabilities of speech presence $P(\mathcal{H}_1)$ and absence $P(\mathcal{H}_0)$. Both priors are fixed and set to $P(\mathcal{H}_1) = P(\mathcal{H}_0) = 0.5$ independent of the observed signal, i.e. without knowledge of the signal, we assume that speech presence and absence are equally likely. Assuming zero-mean complex Gaussian distributed spectral coefficients, the likelihoods $p(y \mid \mathcal{H}_1)$ and $p(y \mid \mathcal{H}_0)$ in (2) are given by

$$p(y \mid \mathcal{H}_i) = \frac{1}{\widehat{\sigma_{\mathrm{N}}^2}\,(1+\xi_{\mathcal{H}_i})\,\pi}\exp\left(-\frac{|y|^2}{\widehat{\sigma_{\mathrm{N}}^2}\,(1+\xi_{\mathcal{H}_i})}\right),\ i \in \{0,1\}\,. \tag{3}$$

In speech absence, the a priori signal to noise ratio (SNR) is $\xi_{\mathcal{H}_0} = 0$. In speech presence, however, $\xi_{\mathcal{H}_1}$ denotes the *typical* a priori SNR if speech were present [5] and it is set to $10\log_{10}(\xi_{\mathcal{H}_1}) = 15\,\mathrm{dB}$ [4].

Inserting (3) into (2) and (2) into (1), the noise periodogram $\widehat{|N_\ell^k|^2}$ can be estimated, where on the right hand side the noise PSD estimate of the last frame $\widehat{\sigma_{\mathrm{N}}^2}\,(k,\ell-1)$ is employed. Finally, the current noise PSD estimate $\widehat{\sigma_{\mathrm{N}}^2}\,(k,\ell)$ is obtained by recursively smoothing $\widehat{|N_\ell^k|^2}$ over time with $\beta = 0.8$,

$$\widehat{\sigma_{\mathrm{N}}^2}\,(k,\ell) = \beta\widehat{\sigma_{\mathrm{N}}^2}\,(k,\ell-1) + (1-\beta)\,\widehat{|N_\ell^k|^2}. \tag{4}$$

3. INCORPORATING SPECTRO-TEMPORAL CORRELATIONS

The likelihood (3) and consequently the SPP (2) in [4] are computed independently for every time-frequency point, not considering temporal and spectral relations within the observed signal. In this section, we now outline two methods that overcome this limitation.

3.1. Spectro-temporal smoothing

The SPP estimator that is used for the noise PSD estimation in [4] is based on a fixed a priori SNR and a fixed a priori SPP as proposed in [5]. However, in [5] the estimator is formulated in a more flexible way and using the a posteriori SNR $\gamma = |y|^2/\widehat{\sigma_{\mathrm{N}}^2}$. In contrast to [5], where the SPP estimator is considered in isolation, assuming that an estimate of the noise PSD $\widehat{\sigma_{\mathrm{N}}^2}$ is available, here we first estimate the SPP to subsequently estimate $\widehat{\sigma_{\mathrm{N}}^2}$. Thus, γ is not readily available, and we instead use the noise PSD estimate of the previous segment $\widehat{\sigma_{\mathrm{N}}^2}\,(k,\ell-1)$ as an initial estimate. As a generalization of (3), the likelihood for the a posteriori SNR given $\mathcal{H}_i$ is modeled as a chi-squared distribution [5]:

$$p(\gamma|\mathcal{H}_i) = \left(\frac{r}{2\,(1+\xi_{\mathcal{H}_i})}\right)^{\frac{r}{2}} \frac{\gamma^{\frac{r}{2}-1}}{\Gamma\left(\frac{r}{2}\right)}\exp\left(-\frac{r\gamma}{2\,(1+\xi_{\mathcal{H}_i})}\right), \tag{5}$$

with the degree of freedom r, which is 2 for the case that the a posteriori SNR γ is not smoothed. To cover spectral and temporal correlations, it is possible to smooth γ along time and frequency. However, the a posteriori SNR after smoothing follows a different distribution than before smoothing. The authors therefore derived methods that relate the amount of smoothing to optimal values of $\xi_{\mathcal{H}_i}$ and r, which both depend on the amount of smoothing. Here we use the same causal setup that has been proposed in [5], i.e. no additional algorithmic delay is introduced. The resulting likelihoods (5) are used in (2) instead of $p(y \mid \mathcal{H}_i)$ (3) to obtain an SPP estimate, which is then used for the estimation of the noise PSD via (1) and (4). Please note that the employed smoothing setup can also be replaced by alternatively techniques, like e.g. temporal cepstrum smoothing [7].

3.2. HMM based SPP estimation

Recently, evolving from previous work of the same authors in the field of multichannel speech separation and noise reduction [8], in [6] a novel SPP estimation scheme has been proposed based on a two dimensional HMM that captures correlations along time and frequency. In each time-frequency point (k,l), the model is either in the speech presence state $Z_\ell^k = \mathcal{H}_1$ or in the speech absence state $Z_\ell^k = \mathcal{H}_0$. Depending on the current state, the model emits an a posteriori SNR that follows either $p(\gamma|\mathcal{H}_1)$ or $p(\gamma|\mathcal{H}_0)$ as defined in (5) with $r = 2$.

With the help of the HMM, we search for the SPP in each time-frequency point (k,l) given *all* observed a posteriori SNRs of a given utterance, i.e. $P\big(Z_\ell^k = \mathcal{H}_1 \mid \gamma_{1:L}^{1:K}\big)$, where $\gamma_{1:L}^{1:K}$ is a matrix containing all γ_ℓ^k. Here, L and K denote the total number of segments and frequency bins, respectively. By considering all observations instead of only the current one, spectral as well as temporal correlations are now incorporated into the SPP estimation process. On the other hand, this formulation imposes that the complete signal must be available at processing time.

Like in [4], for the HMM-based SPP estimation we choose $10\log_{10}(\xi_{\mathcal{H}_1}) = 15\,\mathrm{dB}$ and $P(\mathcal{H}_1) = P(\mathcal{H}_0) = 0.5$. For the estimation of $\gamma_{1:L}^{1:K}$, furthermore the transition probabilities

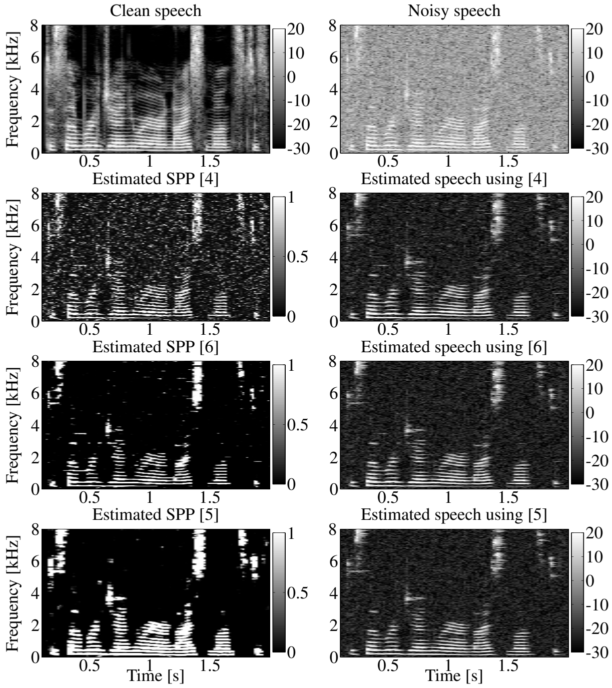

Fig. 1: Spectrograms of clean and noisy speech degraded by white Gaussian noise at 5 dB SNR in the top row. Then, from top to bottom: the SPP estimates of [4], [5], and [6] (left), together with the respective speech enhancement results (right).

along time $a_{i,j}^{H} = P\big(Z_{l+1}^{k} = \mathcal{H}_j \mid Z_{\ell}^{k} = \mathcal{H}_i\big)$ and along frequency $a_{i,j}^{V} = P\big(Z_{l}^{k+1} = \mathcal{H}_j \mid Z_{\ell}^{k} = \mathcal{H}_i\big)$ are needed, which denote the probability that the model switches to state $\mathcal{H}_j$ given that it previously was in state $\mathcal{H}_i$. Here we use the same transition probabilities $a_{i,j}$ along time and frequency and set $a_{1,1} = 0.95 = 1 - a_{1,0}$ and $a_{0,0} = 0.7 = 1 - a_{0,1}$, i.e. we assume that it is more likely for the HMM to stay in the current state than to switch it. Consequently, the resulting SPP estimates are expected to be smoother than the instantaneous estimate used in [4]. While a large value of $a_{0,0}$ reduces false detections of speech presence due to random outliers in the noise, a large value of $a_{1,1}$ leads to a preservation of the speech component. However, the tendency to stay in the same state may also lead to estimation errors at the boundaries between regions of speech presence and speech absence. To avoid missing relevant speech active bins we therefore chose $a_{1,1}$ larger than $a_{0,0}$, putting more emphasize on the preservation of the speech component rather than on the reduction of random outliers. Details on how the 2D HMM is decoded can be found in [6].

As the HMM is based on the a posteriori SNR γ, an initial noise PSD estimate is needed, which we obtain using [4]. The HMM-based SPP estimate is then employed in (1) to reestimate the noise PSD via (4). In principle, the new noise PSD estimate could then in turn be used to refine the SPP estimate and this procedure could be iterated for a number of times. Preliminary experiments however showed that no further improvements are achieved that would justify the increase in computational complexity.

4. EVALUATION

We now evaluate the influence of the alternative SPP estimates on the noise PSD estimation performance.

In Figure 1, we present the different SPP estimates for an example utterance which is degraded by white Gaussian noise at 5 dB SNR. Compared to the instantaneous SPP estimate used in [4], the HMM-based approach strongly reduces spectral outliers in noise only regions. This is due to the choice of the HMM parameters, especially of the transition probabilities, as discussed in Sec. 3.2. The spectral structure of the speech, e.g. the harmonics, is maintained, however, some weak speech components that are detected in the instantaneous estimate are missed, which leads to an undesired leakage of the speech PSD into the noise PSD estimate. Eventually, this results in a local overestimation of the noise PSD. The SPP estimates of [5] depict even less random fluctuations in noise only regions than the HMM-based approach. Since random outliers in the a posteriori SNR are reduced by the spectro-temporal smoothing, also the SPP estimate depending on it shows less outliers. At the same time, speech active regions are blurred, which may further lead to an increased protection of the speech.

This behavior is also reflected in the right column of Figure 1, where we present the noisy signal after speech enhancement using the respective SPP estimate on the left to estimate the noise PSD. The decision-directed approach with a smoothing parameter of $\alpha = 0.98$ [9] is employed for the estimation of the speech PSD and the noisy spectrum is weighted with the Wiener filter to obtain an estimate of the clean speech signal.

The incorporation of spectral and temporal correlations in the SPP estimation leads to an effective suppression of random outliers in the noise, i.e. musical tones, while the speech is well maintained. Due to the smooth SPP estimates, outliers of the noise are more likely to be included in the update of the noise PSD via (1) and (4) and are thus suppressed by the Wiener filter. Both of the proposed methods achieve a clear reduction of musical noise relative to [4], with a slightly more prominent suppression for the approach based on spectro-temporal smoothing.

We instrumentally evaluate the noise PSD estimators by applying the same enhancement scheme that has been used for the example in Figure 1 to a set of 128 utterances of the TIMIT database [10], which are degraded by babble noise [11] and modulated pink noise with a modulation frequency of $f_{\mathrm{mod}} = 0.5$ Hz at various SNRs. Similar to [12], we assess the performance in terms of the segmental speech SNR (spSSNR), the segmental noise reduction (NR), and the overall segmental SNR (SSNR) improvement:

$$\mathrm{NR} = \frac{10}{|\mathbb{L}|} \sum_{l \in \mathbb{L}} \log_{10} \frac{\sum_{m=1}^{M} n_t^2\,(lM+m)}{\sum_{m=1}^{M} \widetilde{n}_t^2\,(lM+m)} \tag{6}$$

$$\mathrm{spSSNR} = \frac{10}{|\mathbb{L}|} \sum_{l \in \mathbb{L}} \log_{10} \frac{\sum_{m=1}^{M} s_t^2\,(lM+m)}{\sum_{m=1}^{M} \left[s_t\,(lM+m) - \widetilde{s}_t\,(lM+m)\right]^2} \tag{7}$$

$$\mathrm{SSNR} = \frac{10}{|\mathbb{L}|} \sum_{l \in \mathbb{L}} \log_{10} \frac{\sum_{m=1}^{M} s_t^2\,(lM+m)}{\sum_{m=1}^{M} \left[s_t\,(lM+m) - \widehat{s}_t\,(lM+m)\right]^2}, \tag{8}$$

where NR measures the amount of noise reduction and spSSNR is a measure for speech distortions and is larger the less speech distortions are introduced. Finally SSNR gives a compromise between speech distortions and noise reduction. Here, we define the time domain speech s_t, noise n_t, and the estimated clean speech signal

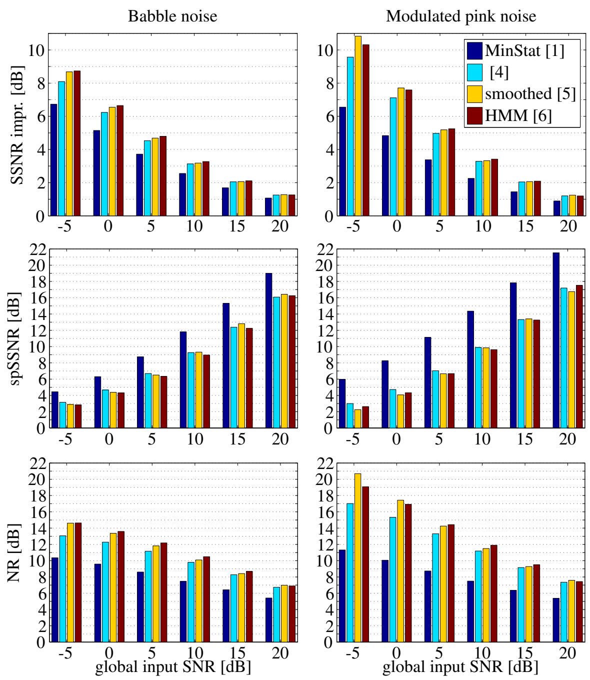

Fig. 2: Segmental SNR (SSNR) improvement relative to the noisy observation, speech SSNR (spSSNR), and segmental noise reduction (NR) (top to bottom) after speech enhancement using noise PSD estimates obtained via Minimum Statistics [1], [4], and the two proposed approaches incorporating signal correlations. The speech is degraded by babble (left) and modulated pink noise with a modulation frequency of $f_{\text{mod}} = 0.5$ Hz (right) over a range of SNRs

$\widehat{s}_t$. To obtain $\widetilde{s}_t$ and $\widetilde{n}_t$, the Wiener filter gain that has been used to estimate $\widehat{s}_t$ is applied to the speech and the noise spectrum separately. The signals are chopped into non-overlapping segments of M samples, where m is the time domain sample index within each segment. For the computation of the measures only the segments which contain relevant speech energy are included, which we denote as $\mathbb{L}$. The results are presented in Figure 2, where for the SSNR the improvement relative to the noisy input is depicted.

It can be seen that the SPP based noise PSD estimation outperforms the minimum statistics approach [1] in terms of the overall SSNR improvement for the presented non-stationary noise types. The incorporation of spectro-temporal correlations for the SPP estimation leads to a further increase in NR while only very little additional speech distortions are introduced, mainly at low SNRs, i.e. only a minor decrease in spSSNR is observed. This is very much in line with the results for the example utterance in Figure 1. Especially in lower SNRs the new estimators yield a favorable trade-off between noise reduction and speech distortion, which is also reflected in a small but consistent increase in the overall SSNR improvement for both approaches.

An important property of any noise PSD estimation algorithm is its capability to track non-stationary noises, which we illustrate in Figure 3 for the original SPP-based noise PSD estimator [4], the two proposed modifications, and the minimum statistics approach [1]. For this, the PSD of a white Gaussian noise signal which is modulated with $0.5\,\text{Hz}$ – as a well defined example of a non-stationary

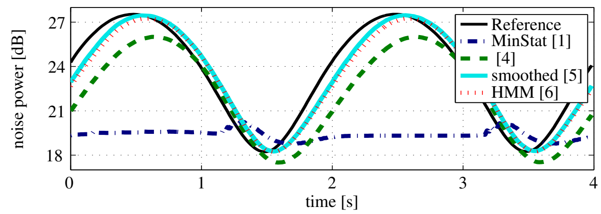

Fig. 3: Noise PSD tracking performance in white Gaussian noise modulated at a frequency of $f_{\text{mod}} = 0.5$ Hz

noise – is estimated with each of the four estimators. Then, the noise PSD estimates are averaged over frequency. To obtain (3), we further repeat the estimation for several realizations of the noise and average the results. As reported in [4], minimum statistics is not capable of tracking the PSD of such a strongly non-stationary signal. The estimation performance is strongly increased by all three SPP based approaches. Relative to the original proposal [4], the new SPP estimates lead to a faster and more accurate tracking of changes in the noise PSD, with a small advantage for [5]. The improved tracking performance can be explained by the reduced number of outliers in the SPP estimates, which have been observed in Figure 1. This allows for a more rapid update of the noise PSD estimates via (1) and (4).

Taking all the results into account, it can be stated that the incorporation of spectral and temporal correlations for SPP-based noise PSD estimation leads to an increased noise reduction, an effective suppression of musical noise, and an improved tracking of non-stationary noises. The price to pay are only marginally increased speech distortions at low SNRs and an increase in computational complexity. The HMM-based approach, as implemented here, only allows for batch processing, rendering it unsuitable for on-line applications. In contrast to that, the smooth SPP estimates of [5] are obtained without an additional algorithmic delay, at the same time achieving a similar performance as the HMM-based approach. Furthermore, the increase in computational complexity over the original method [4] is significantly lower. Nevertheless, the 2D HMM-based SPP estimator [6] is an extremely versatile tool, which allows for numerous setups, of which we have investigated only a few. Further improvements in the SPP and noise PSD estimation performance might be achieved by a different choice of the model parameters. While here the values for $\xi_{\mathcal{H}_1}$, $P(\mathcal{H}_1)$, and the transition probabilities have been set manually, these parameters – or a subset thereof – could for example be optimized based on a large training set of noise corrupted speech in an expectation maximization fashion.

5. CONCLUSIONS

In this contribution, we investigated how SPP-based noise PSD estimation can benefit from utilizing spectro-temporal signal correlations in the SPP estimation stage.While the original proposal [4] is based on an instantaneous SPP estimate, here we employed two alternative estimators of which one uses a 2D HMM [6] and the other operates on a smoothed observation [5] to utilize spectral and temporal information. We experimentally showed that both approaches, when employed in a common speech enhancement framework, lead to an increased noise reduction, less musical noise, and an improved tracking of non-stationary noises, at the expense of an only marginal increase of speech distortions relative to the initial method [4].

6. REFERENCES

[1] Rainer 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.

[2] Israel 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.

[3] Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4266–4269.

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

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

[6] Dang Hai Tran Vu and Reinhold Haeb-Umbach, "Using the turbo principle for exploiting temporal and spectral correlations in speech presence probability estimation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vancouver, Canada, May 2013, pp. 863–867.

[7] Timo Gerkmann, Martin Krawczyk, and Rainer Martin, "Speech presence probability estimation based on temporal cepstrum smoothing," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4254–4257.

[8] Dang Hai Tran Vu and Reinhold Haeb-Umbach, "Exploiting Temporal Correlations in Joint Multichannel Speech Separation and Noise Suppression using Hidden Markov Models," in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Aachen, Germany, Sep. 2012.

[9] Yariv Ephraim and David 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.

[10] John S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST)*, 1988.

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

[12] Thomas Lotter and Peter Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Applied Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Jan. 2005.