IMPROVED SINGLE-CHANNEL NONSTATIONARY NOISE TRACKING BY AN OPTIMIZED MAP-BASED POSTPROCESSOR

Aleksej Chinaev, Reinhold Haeb-Umbach

University of Paderborn
Department of Communications Engineering
33098 Paderborn, Germany
{chinaev,haeb}@nt.upb.de

Jalal Taghia, Rainer Martin

Ruhr-Universität Bochum
Institute of Communication Acoustics
44780 Bochum, Germany
{jalal.taghia,rainer.martin}@rub.de

ABSTRACT

In this paper we present an improved version of the recently proposed Maximum A-Posteriori (MAP) based noise power spectral density estimator. An empirical bias compensation and bandwidth adjustment reduce bias and variance of the noise variance estimates. The main advantage of the MAP-based postprocessor is its low estimation variance. The estimator is employed in the second stage of a two-stage single-channel speech enhancement system, where eight different state-of-the-art noise tracking algorithms were tested in the first stage. While the postprocessor hardly affects the results in stationary noise scenarios, it becomes the more effective the more nonstationary the noise is. The proposed postprocessor was able to improve all systems in babble noise w.r.t. the perceptual evaluation of speech quality performance.

Index Terms— Noise power estimation, Maximum a posteriori estimation, Speech enhancement

1. INTRODUCTION

The noise power spectral density (PSD) estimation algorithm is a key component for many speech processing tasks, such as speech enhancement and automatic speech recognition. Many very sophisticated algorithms have been proposed in the past, and accurate noise tracking remains to be an important and challenging research topic to date [1–9]. Particularly difficult is the tracking of nonstationary noise from a single-channel noisy speech input, as most noise tracking algorithms assume that noise is ”more stationary” than speech and that time-frequency bins can be found, where only noise is present. However, when the distortion is highly nonstationary, the first assumption begins to break down. Further, it is not sufficient to update the noise PSD estimates in speech absence periods only.

Recently, algorithms have been proposed that try to overcome these limitations. For example, the transient noise reduction algorithm proposed in [10] that relies on non-local filtering, allows for the reduction of repetitive highly nonstationary noise. In [11] we have proposed a MAP-based (”MAP-B”) estimator which can update its noise estimate even if speech is dominant in the time-frequency bin under consideration. The estimator relied on approximating the posterior of the noise variance in the presence of an observation of the noise variance, which is ”distorted” by speech with known power, by a conjugate prior with the same mode as the true posterior. This mode could be efficiently computed and served as noise variance estimate. To have an estimate of the speech power available, the MAP estimator was used as a postprocessor to a first speech enhancement stage (SES).

A later performance analysis revealed that the MAP-B noise tracker was strikingly insensitive to zero mean estimation errors of the input speech power. However it was also observed that the estimator was not bias-free and performance degraded for large input signal-to-noise (SNR) ratios [12]. These shortcomings, however, can be overcome by an optimization, as is shown in this contribution.

Recently, an extensive performance evaluation of a total of eight state-of-the-art noise tracking algorithms under various adverse acoustic environments has been conducted [13]. In this evaluation not only the mean of a spectral distance but also the variance of the estimators has been assessed. The latter is related to undesirable fluctuations, known as musical tones. In this contribution we extend this evaluation and investigate whether the optimized MAP-B noise tracker is able to improve upon the result of the other eight noise trackers in the first SES. Indeed the results show that MAP-B reduces the variance of the noise estimates for all noise trackers. This leads to improved speech quality of the speech enhancement system, as measured by the perceptual evaluation of speech quality (PESQ) score, for nonstationary noise environments.

The paper is organized as follows. In the next section we briefly summarize the MAP-B algorithm and its use in a two-stage speech enhancement system. Sec. 3 addresses the optimization by an SNR-dependent bias removal and bandwidth adjustment. In Sec. 4 we present the experimental setup, followed by the results in Sec. 5. The paper is finished with the conclusions in Sec. 6.

2. MAP-BASED NOISE VARIANCE ESTIMATION IN A TWO-STAGE SPEECH ENHANCEMENT SYSTEM

In [11] we have presented a noise PSD estimation algorithm and its use in a single-channel speech enhancement system. Given the short-time Fourier transform (STFT) coefficients $Y_{kl} = X_{kl} + N_{kl}$ of the noisy speech, where k and l denote frequency bin and time frame index, respectively, and where X_{kl} and N_{kl} are the STFTs of speech and noise, the algorithm determines an approximate MAP estimate of the noise variance $\sigma^2_{N,kl} = E[|N_{kl}|^2]$, assuming that an estimate of the speech power $\sigma^2_{X,kl} = E[|X_{kl}|^2]$ is available. To this end the a-priori probability density function (PDF) $p_{\sigma^2_{N,kl}}$ of the time variant noise power for each frequency bin was modeled by a scaled inverse chi-square (SICS) distribution

$$p_{\sigma^2_{N,kl}}(\sigma^2; \nu_0, \lambda^2_{kl}) = \frac{(\nu_0 \cdot \lambda^2_{kl})^{\nu_0/2}}{\Gamma(\nu_0/2)} \cdot \left(\sigma^2\right)^{-\frac{\nu_0+2}{2}} \cdot e^{-\frac{\nu_0 \cdot \lambda^2_{kl}}{2\sigma^2}} \quad (1)$$

The work was in part supported by Deutsche Forschungsgemeinschaft under contract no. Ha3455/8-1.

with the degrees of freedom ν_0 and the scale λ^2_{kl}. (1) is a conjugate prior for the Gaussian observation PDF p_Y in the case of absence of speech. In order to maintain an efficient estimation procedure in the presence of speech, the posterior $p_{\sigma^2_N|Y}$ was approximated by a SICS distribution with the same mode as the exact posterior PDF.

To have an estimate of the speech power available, the MAP estimator was used as a postprocessor to a first SES. Fig. 1 shows this two-stage configuration, where the upper part depicts the first SES [14], while the lower part includes the proposed MAP-B postprocessor. It delivers an improved noise variance estimate $\hat{\hat{\sigma}}^2_{N,kl}$, which in the following leads to an improved clean speech STFT estimate $\hat{X}^{II}_{kl}$. The first stage forwards an estimate $\hat{\sigma}^2_{X,kl}$ of the speech PSD and a smoothed version $\hat{\zeta}_{kl}$ of the a-priori SNR $\hat{\xi}_{kl}$ to the second stage, where

$$\hat{\zeta}_{kl} = \alpha_\zeta \cdot \hat{\zeta}_{k,l-1} + (1 - \alpha_\zeta) \cdot \hat{\xi}_{kl} \tag{2}$$

with $\alpha_\zeta = 0.7$ being a typical value [1]. Note that the speech absence probability $\hat{q}_{kl}$ to be used in the gain calculation is not recalculated in the second stage, which saves some computations. Further it is important to note that the second stage operates on the same, undelayed, noisy speech signal Y_{kl} as the first stage. Thus the postprocessor does not incur any additional latency compared to single-stage speech enhancement system.

Any noise tracking algorithm may be used in the first stage. In [11] we employed the Improved Minimal Controlled Recursive Averaging algorithm (IMCRA) [4], while in [12] a simplified version of the Minimum Statistics (MS) approach was used [2].

The quality analysis carried out in [12] revealed that the MAP-B postprocessor delivers noise variance estimates with a positive bias (i.e., the variance is overestimated), which grows with increasing SNR. Further, it was observed that the improvements obtained by MAP-B diminished with increasing input SNR. On the other hand, the quality analysis also revealed the excellent immunity of the algorithm against (zero mean) estimation errors in the input speech power $\hat{\sigma}^2_{X,kl}$. The latter finding could hint to a potentially good performance in a nonstationary noise environment, where speech power estimation is particularly difficult. This insensitivity to speech power estimation errors could then be beneficial for the reduction of musical tones. However, first the mentioned shortcomings needed to be removed. The next section shows how this can be achieved.

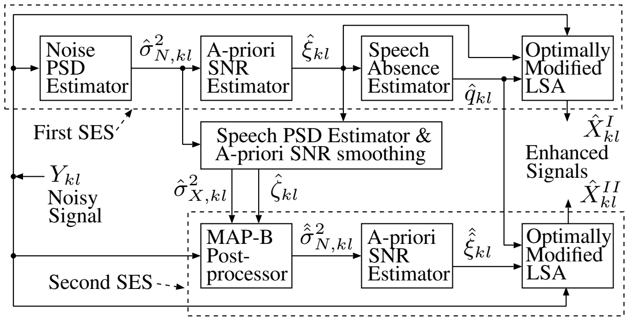

Fig. 1. Two-stage single-channel speech enhancement system

3. OPTIMIZED MAP-BASED NOISE POWER SPECTRAL DENSITY TRACKER

The quality analysis had shown that the MAP-B postprocessor of [11] delivers estimates with a positive bias, that grows with increasing SNR. It is therefore proposed to reduce the MAP-B estimate as a function of the SNR:

$$\hat{\hat{\sigma}}^2_{N,kl} = (1 - \beta_{kl}(\mathrm{SNR}_{kl})) \cdot \tilde{\sigma}^2_{N,kl} \tag{3}$$

employing the SNR-dependent bias compensation factor $\beta_{kl}(\mathrm{SNR}_{kl})$. Here, $\tilde{\sigma}^2_{N,kl}$ denotes the initial, biased MAP-B noise variance estimate. As an estimate of the SNR the smoothed a-priori SNR $\hat{\zeta}_{kl}$, see (2), can be employed. The experiments revealed that the following rule led to a simple, yet effective bias reduction

$$\beta_{kl}(\hat{\zeta}_{kl}) = \beta_{max} \cdot \left(\frac{\arctan(\hat{\zeta}_{kl})}{\pi} + \frac{1}{2} \right), \tag{4}$$

if $\beta_{kl}(\hat{\zeta}_{kl})$ replaces $\beta_{kl}(\mathrm{SNR}_{kl})$ in (3). Here, β_{max} is a bias compensation factor, which is set to $\beta_{max} = 0.01$, and $\hat{\zeta}_{kl}$ has to be given in dB.

The MAP-B postprocessor has a single tunable parameter, the degrees of freedom ν_0 of the SICS distribution (1), which in [11] was chosen to some constant value. The choice of ν_0 determines the weight of the a-priori information relative to the current observation. The larger ν_0 the narrower is the a-priori distribution and the more weight is given to the a-priori knowledge in comparison to the current observation. In other words, the parameter ν_0 controls the bandwidth of the MAP-B noise tracker.

The observation made in [11], that the performance degraded with increasing input SNR, seems to indicate that the bandwidth of the noise tracker should be reduced for large SNR. This can be achieved by using a time variant parameter ν_{kl}:

$$\nu_{kl}(\hat{\zeta}_{kl}) = \nu_0 + \frac{\Delta_\nu}{\pi} \cdot \arctan\left(\hat{\zeta}_{kl}\right), \tag{5}$$

with a constant degrees of freedom $\nu_0 = 40$ and an adjustment range $\Delta_\nu = 10$. This is reminiscent to many other algorithms, which halt the noise PSD estimation in the presence of large input SNR.

4. EXPERIMENTAL SETUP

In this section the robustness of the MAP-B postprocessor in adverse environments is examined by providing a variety of different nonstationary noises. We follow the evaluation setup introduced in [13] and consider eight noise PSD estimators, which are the subspace noise tracking (SNT) algorithm [6], two minimum mean-squared error (MMSE) based approaches, i.e. MMSE-Hendriks [9] and MMSE-Yu [7], four minima controlled recursive averaging (MCRA) based algorithms, i.e. the original MCRA algorithm [3], the IMCRA algorithm [4] as well as two other methods belonging to this category, such as EMCRA [5] and MCRA-MAP [8], and finally another state-of-the-art approach, i.e. the MS algorithm [2]. In the experiments we intend to show how much the MAP-B postprocessor can be helpful in improving the noise PSD tracking performance of the aforementioned algorithms, and subsequently how much effective it is in increasing the quality of the estimated speech derived by the first SES introduced in section 2. For a performance analysis of the MAP-B estimator without the optimizations of section 3 we refer to [11], where IMCRA was used in the first SES.

In our experiments the sampling frequency of all signals is 8 kHz. Clean speech signals are taken from the TIMIT database [15]. By concatenating sentences and removing the beginning and trailing silences, two clean speech signals are generated for our experiments; one for female speech and one for male speech. Each clean speech signal has a length of 2 minutes and includes speech of 8 different speakers spoken in English. At the beginning of each clean speech data 0.1 seconds silence is included. Clean speech is degraded by different noise types. Here we present the results for three noise types. We select babble noise (produced by a large crowd) from NOISEX-92 [16] as a representative of a highly nonstationary noise, and car noise (inside a car during acceleration and deceleration) from SOUND-IDEAS database [17] representing only mildly nonstationary noise. Moreover, we consider a modulated white Gaussian noise (WGN) which is called ”sinusoidal WGN” and which is obtained through modulating WGN by the following function

$$g(n) = 1 + \sin\left(\frac{2\pi n}{f_s} . f_{\text{mod}}\right), \tag{6}$$

where n is the sample index, f_s the sampling frequency and f_{mod} indicates the varying modulation frequency, which linearly increases in 30 seconds from 0 Hz to 0.25 Hz. In the experiments, for each type of noises, we varied the input overall SNR from -5 dB to 20 dB in steps of 5 dB.

The reference noise PSD as employed in [13] can be derived by a recursive temporal smoothing of noise periodograms with a smoothing parameter $\alpha = 0.9$. However, the recursive smoothing by the IIR-filter incurs a delay of $\frac{\alpha}{1-\alpha}$ samples, which can be advantageous for a noise tracker, which happens to have the same delay. Thus, for our experiments we employed a delayless filter realized by the Matlab function *filtfilt*(1-α, [1 -α], $|N_{kl}|^2$), which performs the smoothing over all frames by processing the original noise power $|N_{kl}|^2$ for each frequency bin in both the forward and reverse directions to obtain an undelayed reference noise PSD $\sigma^2_{N,kl}$.

Two performance measures LogErr$_{\text{mean}}$ and LogErr$_{\text{var}}$ are taken into account to examine the performance of noise PSD trackers [13]. LogErr$_{\text{mean}}$ is defined as the mean of the spectral distance between the reference noise PSD $\sigma^2_{N,kl}$ and the estimated noise PSD, either from the first stage ($\hat{\sigma}^2_{N,kl}$) or from the second ($\hat{\hat{\sigma}}^2_{N,kl}$). LogErr$_{\text{var}}$ computes the variance of the estimation error and it is more related to undesirable fluctuations in the estimated noise PSD. The first 3 seconds of the input signals are used for the initialization of the algorithms and are excluded from the computation of the performance measures. Moreover, the first 5 frequency bins as well as the 5 bins below the Nyquist frequency are excluded as well in order to reduce the influence of DC-removal and anti-aliasing filter.

All noise power estimators were implemented in a DFT-based spectral analysis system using overlapping square-root periodic Hann windows. The window length as well as the DFT length is 256 samples, and the amount of the overlap between frames is considered separately based on recommendations reported by the authors of the algorithms. The frame overlap factor for MS, MMSE-Yu, MMSE-Hendriks, SNT algorithms is set to 50%, and for MCRA, IMCRA, EMCRA, MCRA-MAP algorithms to 75%. Having different frame overlap factors results in producing different numbers of frames. Thus, to have the same number of frames for the evaluation of noise PSD estimators in terms of LogErr$_{\text{mean}}$ and LogErr$_{\text{var}}$ we sub-sample the reference and estimated noise PSD for those algorithms, which use more than 50% frame overlap.

The overall performance measures are defined as follows:

$$\Delta\text{LogErr}_{\text{mean}} = -(\text{LogErr}^{(II)}_{\text{mean}} - \text{LogErr}^{(I)}_{\text{mean}}), \tag{7}$$

$$\Delta\text{LogErr}_{\text{var}} = -(\text{LogErr}^{(II)}_{\text{var}} - \text{LogErr}^{(I)}_{\text{var}}), \tag{8}$$

$$\Delta\text{PESQ} = \text{PESQ}^{(II)} - \text{PESQ}^{(I)}, \tag{9}$$

where LogErr$^{(I)}_{\text{mean}}$ and LogErr$^{(I)}_{\text{var}}$ are computed from the estimated noise PSD $\hat{\sigma}^2_{N,kl}$ in the first stage. Similarly, LogErr$^{(II)}_{\text{mean}}$ and LogErr$^{(II)}_{\text{var}}$ are computed from the estimated noise PSD $\hat{\hat{\sigma}}^2_{N,kl}$ of the second stage. PESQ$^{(I)}$ and PESQ$^{(II)}$ are computed from the enhanced speech signal of the first and second stage, respectively.

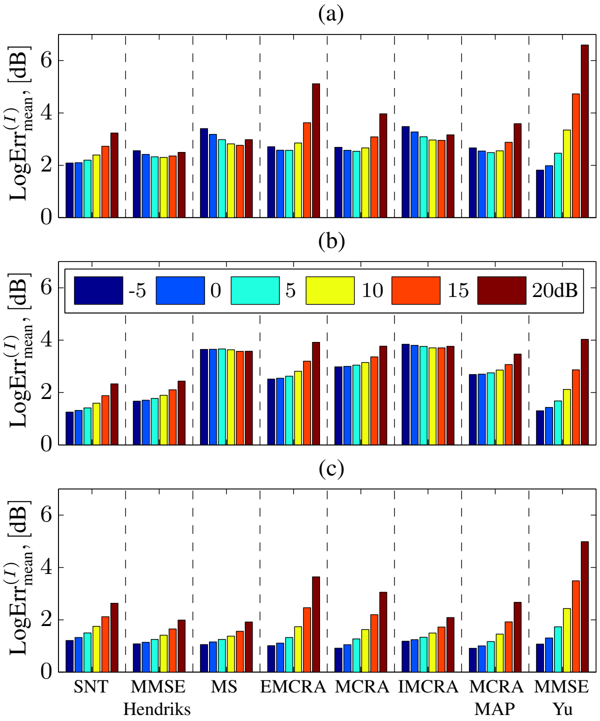

Fig. 2. Performance of the noise PSD estimators in the first stage before applying of MAP-B estimator for various noise types in terms of LogErr$^{(I)}_{\text{mean}}$: (a) babble, (b) sinusoidal WGN, (c) car noise.

ΔLogErr$_{\text{mean}}$ and ΔLogErr$_{\text{var}}$ show the amount of attenuation of the noise estimation error provided by the MAP-B postprocessor, and ΔPESQ expresses how much the MAP-B postprocessor is effective in improving the speech quality. For the performance measures in (7)-(9), the larger values show better performance.

5. EXPERIMENTAL RESULTS

The performance of the noise PSD estimators in the first stage before applying of MAP-B postprocessor is shown for various noise types in terms of LogErr$^{(I)}_{\text{mean}}$ in Fig. 2. One can see that the considered noise trackers perform quite differently. While the babble noise seems to be the most difficult noise type to track, the most easiest is the car noise. Furthermore the SNT and MMSE-Hendriks estimators seem to reach the best averaged performance.

In Fig. 3 we show the effect of the MAP-B postprocessor on the accuracy of the noise power estimation with respect to the performance measures ΔLogErr$_{\text{mean}}$ and ΔLogErr$_{\text{var}}$. Furthermore, the impact of the MAP-B postprocessor on the improvement of the speech quality as measured by ΔPESQ is presented in Fig. 4.

Looking at the results for ΔLogErr$_{\text{var}}$ in Fig. 3, it can be seen that for almost all tested environments the MAP-B postprocessor was able to reduce the estimator's variance, in particular for the more nonstationary noise types, i.e. babble noise and sinusoidal WGN.

In terms of reduction of the mean estimation error ΔLogErr$_{\text{mean}}$

the proposed approach performs quite well for babble noise, which is the most difficult to track. As a consequence of the better noise tracking, a consistent quality improvement of the enhanced speech signals was observed for all noise PSD estimators, see Fig. 4(a).

According to the results for $\mathrm{LogErr}^{(I)}_{\mathrm{mean}}$ from Fig.2(b) SNT and both MMSE-based approaches are able to track the sinusoidal WGN noise type better than MS and the MCRA-based approaches. Here,

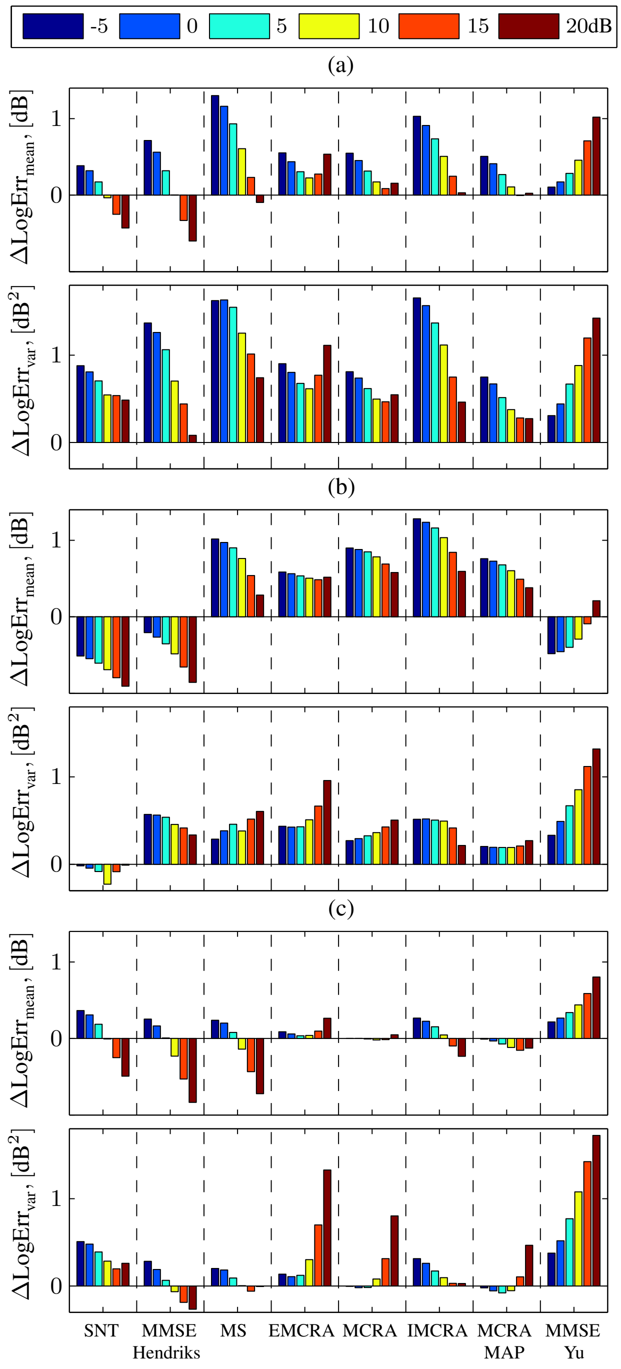

Fig. 3. Impact of the MAP-B postprocessor on the accuracy of noise power estimation for various noise types in terms of $\Delta\mathrm{LogErr}_{\mathrm{mean}}$ and $\Delta\mathrm{LogErr}_{\mathrm{var}}$: (a) babble, (b) sinusoidal WGN, (c) car noise.

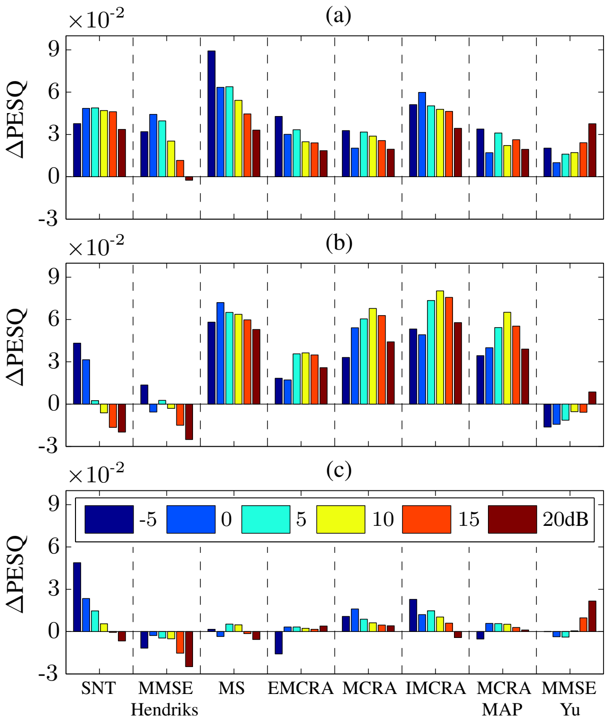

Fig. 4. The impact of the MAP-B postprocessor on the improvement of speech quality as measured by ΔPESQ for various noise types: (a) babble noise, (b) sinusoidal WGN, (c) car noise.

the MAP-B postprocessor could improve the noise tracking in terms of $\Delta\mathrm{LogErr}_{\mathrm{mean}}$ only for the last mentioned approaches. However, as one can see in Fig. 4(b), the quality of enhanced signals by using the SNT and both MMSE-based approaches was barely affected.

As can be seen from the results of Fig. 2(c), the car noise type can be tracked by all evaluated noise trackers quite well, and the MAP-B postprocessor can not improve on that except of noise PSD estimates of MMSE-Yu tracker, see Fig. 3(c). Notwithstanding, because of the attenuation of the variance of the noise PSD estimation, an average slight improvement of the PESQ measure was noticed, see Fig. 4(c).

6. CONCLUSIONS AND RELATION TO PRIOR WORK

The extensive performance analysis described in this contribution showed that a two-stage speech enhancement system that includes an optimized version of the MAP-B noise PSD estimator in the second stage is able to reduce the variance of all eight state-of-the-art noise estimation algorithms and consequently led to improved speech quality for nonstationary noise environments. For more stationary noise the first stage performs already well and the MAP-B estimator is only able to reduce the variance of the noise PSD estimate. In this case a second stage is not necessary. The second stage can be realized very efficiently adding no latency to the system.

The MAP-B estimator was proposed in [11], and the optimizations of the MAP-B noise estimator used here are based on an analysis described in [12]. These approaches lead to a reduced bias and improved performance in high SNR. The experimental framework under which the noise trackers were compared has been taken from [13] and was extended to include a speech quality measure.

7. REFERENCES

[1] I. Cohen and S. Gannot, *'Spectral Enhancement Methods' in Springer Handbook of Speech Processing*, J. Benesty, M.M. Sondhi and Y. Huang, Berlin, Germany: Springer-Verlag, Chapter 44, Part H, pp. 873–901, 2008.

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

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

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[5] N. Fan, J. Rosca, and R. Balan, "Speech noise estimation using enhanced minima controlled recursive averaging," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. IV–581–IV–584, June 2007.

[6] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 541–553, March 2008.

[7] R. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4421–4424, April 2009.

[8] J. M. Kum, Y. S. Park, and J. H. Chang, "Speech enhancement based on minima controlled recursive averaging incorporating conditional maximum a posteriori criterion," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4417–4420, April 2009.

[9] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4266–4269, March 2010.

[10] R. Talmon, I. Cohen, and S. Gannot, "Transient noise reduction using nonlocal diffusion filters," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1584–1599, August 2011.

[11] A. Chinaev, A. Krueger, Dang Hai Tran Vu, and R. Haeb-Umbach, "Improved noise power spectral density tracking by a MAP-based postprocessor," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4041–4044, March 2012.

[12] A. Chinaev and R. Haeb-Umbach, "Quality analysis and optimization of the MAP-based noise power spectral density tracker," *10. ITG Symposium on Speech Communication*, pp. 1–4, September 2012.

[13] J. Taghia, J. Taghia, N. Mohammadiha, S. Jinqiu, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4640–4643, May 2011.

[14] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, November 2001.

[15] "TIMIT, acoustic-phonetic continuous speech corpus," DARPA, NIST Speech Disc 1-1.1, October 1990.

[16] A. Varga and H. J. 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," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[17] B. Nimens, "Sound ideas: sound effects collection," Series 6000; http://www.sound-ideas.com/6000.html, March 2013.