On Optimal Smoothing in Minimum Statistics Based Noise Tracking

Aleksej Chinaev, Reinhold Haeb-Umbach

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

Abstract

Noise tracking is an important component of speech enhancement algorithms. Of the many noise trackers proposed, Minimum Statistics (MS) is a particularly popular one due to its simple parameterization and at the same time excellent performance. In this paper we propose to further reduce the number of MS parameters by giving an alternative derivation of an optimal smoothing constant. At the same time the noise tracking performance is improved as is demonstrated by experiments employing speech degraded by various noise types and at different SNR values.

Index Terms: speech enhancement, noise tracking, optimal smoothing

1. Introduction

In the short time Fourier transform (STFT) domain speech is known to have a sparse representation, i.e., the vast majority of energy is concentrated in a few time-frequency (TF) slots. It is this property that is exploited in many speech signal extraction algorithms, and it is also this property that can be exploited for estimating the power spectral density (PSD) of the noise. Even in time intervals, where a person is speaking, TF slots exisit that have no or negligible speech energy, such that the STFT values of these slots are governed by the background noise. If the noise is sufficiently stationary, these TF slots suffice to track the noise PSD which can then be used in speech enhancement algorithms.

However, a major issue is how to reliably identify those TF slots in which the noise spectrum can be observed. One way to track the noise PSD without the necessity of an error-prone speech presence probability estimation is the famous Minimum Statistics (MS) approach [1].

The MS estimator is developed on the observation that the PSD of a noisy speech signal often drops to the noise power level [2]. To bypass the intermediate time spans with speech activity the MS algorithm carries out a minimum search on the smoothed spectrum of the noisy speech, using a causal sliding window of a certain length. From the found minima an unbiased noise PSD estimate is provided after a bias compensation. Besides an efficient realisation of the minimum search and a sophisticated calculation of the bias compensation factor, a third key component of the MS approach is an analytical derivation of the optimal smoothing parameter to be used for smoothing the noisy speech spectrum before the minimum search is applied.

A recursive averaging of the PSD of the noisy speech is also done in many other state-of-the-art noise trackers [1–8]. Most of them consider the noise PSD as an unknown but fixed parameter, rather than a random variable.

In this contribution we treat the noise PSD as a random variable and employ Bayesian estimation. The update equations for the maximum a posteriori or minimum mean squared error estimates result in a recursive averaging, thus giving a nice statistical derivation and interpretation of the smoothing operation mentioned above. This derivation also delivers the optimal time-variant smoothing parameter, however only for the case of a constant noise PSD. It needs to be adjusted for the practical case of a time-variant noise PSD, for which we also propose a solution.

In [1] a different derivation of the smoothing parameter is given. The noise PSD is treated as a deterministic parameter, which is estimated in the Maximum Likelihood sense. The optimal smoothing parameter is derived analytically from the statistics of the noisy observations. However, it exhibits some issues, such as a so-called dead lock for values of the a posteriori SNR close to one and poor performance of the noise tracking in high levels of nonstationary noise. For these issues heuristic solutions are given in [1]. Here we show that the new derivation based on Bayesian estimation mitigate these issues.

The paper is organized as follows. In the next section we introduce the statistical modeling and present the smoothing of the observed noise PSD as a consequence of Bayesian estimation of a constant, however, random noise PSD. In Section 3 we consider the practical case of time-variant noise statistics and adjust the smoothing parameter derived in the Section 2 correspondingly. We then compare its behavior with the solution given in [1]. Experimental results on noise tracking and speech enhancement are presented in Section 4, before we conclude the paper in Section 5.

2. Statistical modelling

2.1. ML estimation of noise PSD

The STFT coefficients of the clean speech signal $S_{\ell,k}$ and of the noise signal $N_{\ell,k}$ at frame number ℓ and frequency bin k are modelled as uncorrelated complex-valued zero-mean normally distributed random processes with statistically independent real and imaginary parts. In this case an additive superposition of the two signals results in an exponentially distributed power spectral density of the noisy speech $|Y_{\ell,k}|^2 = |S_{\ell,k}|^2 + |N_{\ell,k}|^2$. Its probability density function (PDF) is given by:

$$p_{|Y_{\ell,k}|^2}(x;\lambda_{Y,\ell,k}) = \frac{1}{\lambda_{Y,\ell,k}} \cdot \exp\left(-\frac{x}{\lambda_{Y,\ell,k}}\right), \qquad (1)$$

where $\lambda_{Y,\ell,k} = E\{|Y_{\ell,k}|^2\} = \lambda_{S,\ell,k} + \lambda_{N,\ell,k}$, and where $\lambda_{S,\ell,k} = E\{|S_{\ell,k}|^2\}$ and $\lambda_{N,\ell,k} = E\{|N_{\ell,k}|^2\}$ are the speech and noise variances. The goal of a noise tracker is to estimate $\lambda_{N,\ell,k}$ from the observed noisy spectrogram $|Y_{\ell,k}|^2$. Since the PSD estimator treats each frequency component identically and independently of the others, we will drop the frequency bin index k in the following.

The noise PSD estimation is especially challenging in the presence of speech and if the noise is nonstationary. On the other hand, in the absence of speech, i.e., if $\lambda_{Y,\ell} = \lambda_{N,\ell}$ and

if the noise is stationary $\lambda_{N,\ell} = \lambda_N$ for a certain number L of consecutive observations, the unbiased Maximum Likelihood (ML) estimate of the noise PSD is simply given by the sample mean $\hat{\lambda}_N^{\text{ML}} = (1/L) \cdot \sum_{\ell=1}^{L} |Y_\ell|^2$ with $\text{var}\left(\hat{\lambda}_N^{\text{MS}}\right) = 2\lambda_N^2/L$.

Although in the ML estimation the true noise PSD λ_N is treated as an unknown but fixed parameter, the estimate $\hat{\lambda}_N^{\text{ML}}$ is a random variable, since it is a function of the random observations $|Y_\ell|^2$. It has a scaled chi-squared (χ_s^2) PDF:

$$p_{\hat{\lambda}_N^{\text{ML}}}(x;\nu,\tau^2) = \frac{(\tau^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} \exp\left(-x \cdot \tau^2\nu/2\right) \quad (2)$$

for $x > 0$ and zero else. Here, $\nu > 0$, $\tau^2 = 1/\lambda_N > 0$ and $\Gamma(\cdot)$ denote the degrees of freedom, the inverse of the noise variance λ_N and the gamma function, respectively. It is worthwhile noting, that due to the signal processing steps in the STFT computation the degrees of freedom parameter should be usually reduced according to $\nu \approx L/a$ for the DC and the Nyquist frequency and otherwise $\nu \approx 2L/a$, where a is a function of L and of the STFT parameters such as the type of analysis window and the frame length [2].

Rather than using the $\hat{\lambda}_N^{\text{ML}}$ estimate a common estimation technique in noise PSD tracking is a first-order recursive averaging over the past noisy PSDs, assuming speech absence:

$$\tilde{\lambda}_{N,\ell} = \alpha_\ell \cdot \tilde{\lambda}_{N,\ell-1} + (1-\alpha_\ell) \cdot |Y_\ell|^2 \quad (3)$$

using a smoothing parameter α_ℓ. While the noise PSD estimators in [2–5] apply a constant smoothing parameter $\alpha_\ell = \alpha$, other sophisticated algorithms as in [1, 6–8] use a time-variant smoothing parameter. Similar to the sample mean $\hat{\lambda}_N^{\text{ML}}$, the estimate $\tilde{\lambda}_{N,\ell}$ in (3) can be also approximately modelled to be a χ_s^2 distributed random variable with appropriate degrees of freedom ν_ℓ dependend on α_ℓ.

In the MS approach the smoothed noisy speech according to eq. (3) is the input to the minimum search algorithm.

2.2. Bayesian estimation for stationary noise

We new treat the sought-after noise PSD λ_N as a random variable which we seek to estimate by means of Bayesian estimation. For the exponentially distributed observations, see eq. (1), the scaled inverse chi-squared (χ_{si}^2) distribution is a conjugate a priori distribution in speech absence. For $x > 0$ it is given by:

$$p_{\lambda_N}(x;\nu,\tau^2) = \frac{(\tau^2\nu/2)^{\nu/2}}{\Gamma(\nu/2)} x^{-\nu/2-1} \exp\left(-\frac{\tau^2\nu/2}{x}\right), \quad (4)$$

where $\nu > 0$ and $\tau^2 > 0$ denote the degrees of freedom parameter and the scale parameter, respectively.

For this conjugate prior the update equations for the hyper parameters degrees of freedom and scale are given by:

$$\nu_\ell = \nu_{\ell-1} + 2, \quad (5) \qquad \tau_\ell^2 = \frac{2}{\nu_\ell}|Y_\ell|^2 + \frac{\nu_{\ell-1}}{\nu_\ell}\tau_{\ell-1}^2, \quad (6)$$

where $\{\nu_{\ell-1}; \tau_{\ell-1}^2\}$ and $\{\nu_\ell; \tau_\ell^2\}$ are the parameters of the a priori and a posteriori distributions, respectively. To get a desired recursive equation we need to condense the posterior PDF to a point estimate. The most popular are the mode and the mean of the posterior resulting in the maximum a-posteriori (MAP) and minimum mean squared error (MMSE) estimates:

$$\tilde{\lambda}_{N,\ell}^{\text{MAP}} = \frac{\tau_\ell^2\nu_\ell}{\nu_\ell + 2} \quad (7) \qquad \tilde{\lambda}_{N,\ell}^{\text{MMSE}} = \frac{\tau_\ell^2\nu_\ell}{\nu_\ell - 2}. \quad (8)$$

We summarize the two equations in a single one via

$$\tilde{\lambda}_{N,\ell} = \frac{\nu_\ell}{\nu_\ell + \Delta\nu} \cdot \tau_\ell^2, \quad (9)$$

which for $\Delta\nu = 2$ gives the MAP estimate and for $\Delta\nu = -2$ the MMSE estimate of the noise PSD.

Using (5) and (9) in (6) results in the recursive equation:

$$\tilde{\lambda}_{N,\ell} = \frac{\nu_{\ell-1} + \Delta\nu}{\nu_{\ell-1} + \Delta\nu + 2} \cdot \tilde{\lambda}_{N,\ell-1} + \frac{2}{\nu_{\ell-1} + \Delta\nu + 2} \cdot |Y_\ell|^2. \quad (10)$$

By comparison of (3) and (10) the smoothing parameter α_ℓ of Bayesian recursion turns out to be a function of $\nu_{\ell-1}$ and $\Delta\nu$:

$$\alpha_\ell = \frac{\nu_{\ell-1} + \Delta\nu}{\nu_{\ell-1} + \Delta\nu + 2}. \quad (11)$$

According to (11) $\alpha_l \in (0;1)$ is guaranteed for all $\nu_{\ell-1} > 0$, if $\Delta\nu > 0$. For $\Delta\nu < 0$, it should be ensured that $\nu_{\ell-1} \geq \Delta\nu$. In the Bayesian estimation a small value of α_ℓ means, that the estimator does not rely on its a-priori knowledge and that the current observation $|Y_\ell|^2$ is more emphasized. Otherwise, if α_ℓ is close to 1, the estimator is confident in its a-priori knowledge and the observation $|Y_\ell|^2$ is deeply distrusted.

Although the two equations (3) and (10) apply the same recursive smoothing and $\tilde{\lambda}_{N,\ell}$ is in both cases a χ_s^2 distributed random variables, the modeling of $\lambda_{N,\ell}$ is quite different. While in (3) $\lambda_{N,\ell}$ is treated as an unknown fixed parameter, is it a χ_{si}^2 distributed random variable in (10). The identity of (3) and (10) is a strong analytical argument for the choice of the χ_{si}^2 distribution as a distribution of $\lambda_{N,\ell}$ and its practical relevance.

It should be mentioned that the update equation for the degrees of freedom parameter in (5) is an appropriate choice for stationary noise only, and it should be modified for nonstationary noise tracking. For instance in [9] we used a constant degrees of freedom $\nu_\ell = \nu_0$, that is similar to recursive smoothing with a constant smoothing parameter like in [2–5]. In this contribution we also drop the update equation $\nu_{\ell-1} = \nu_{\ell-2} + 2$ and derive an alternative in the next section.

3. Smoothing for nonstationary noise

As mentioned above in the MS approach, $\lambda_{N,\ell}$ is treated to be a deterministic parameter. An optimal smoothing parameter is analytically derived in [1] for observations containing speech pauses by minimizing the mean squared error:

$$\alpha_\ell^{\text{MS-opt}} = \underset{\alpha}{\text{argmin}} \; E\left[\left(\tilde{\lambda}_{N,\ell} - \lambda_{N,\ell}\right)^2 \middle| \tilde{\lambda}_{N,\ell-1}\right]. \quad (12)$$

This results in a smoothing constant that is a function of the smoothed a posteriori SNR $\hat{\gamma}_\ell$:

$$\alpha_\ell^{\text{MS-opt}} = \frac{1}{1 + (\hat{\gamma}_\ell - 1)^2}, \quad (13) \qquad \hat{\gamma}_\ell = \frac{\tilde{\lambda}_{N,\ell-1}}{\lambda_{N,\ell}}, \quad (14)$$

where in a practical implementation $\lambda_{N,\ell} = \hat{\lambda}_{N,\ell-1}^{\text{MS}}$ is used, employing the noise PSD estimate $\hat{\lambda}_{N,\ell-1}^{\text{MS}}$ obtained from the conventional MS approach [1]. Note, that $\alpha_\ell^{\text{MS-opt}}$ may take a very small values, so that $\tilde{\lambda}_{N,\ell}$ is able to follow the noisy speech spectrum $|Y_\ell|^2$ even in speech presence, where $\hat{\gamma}_\ell \gg 1$.

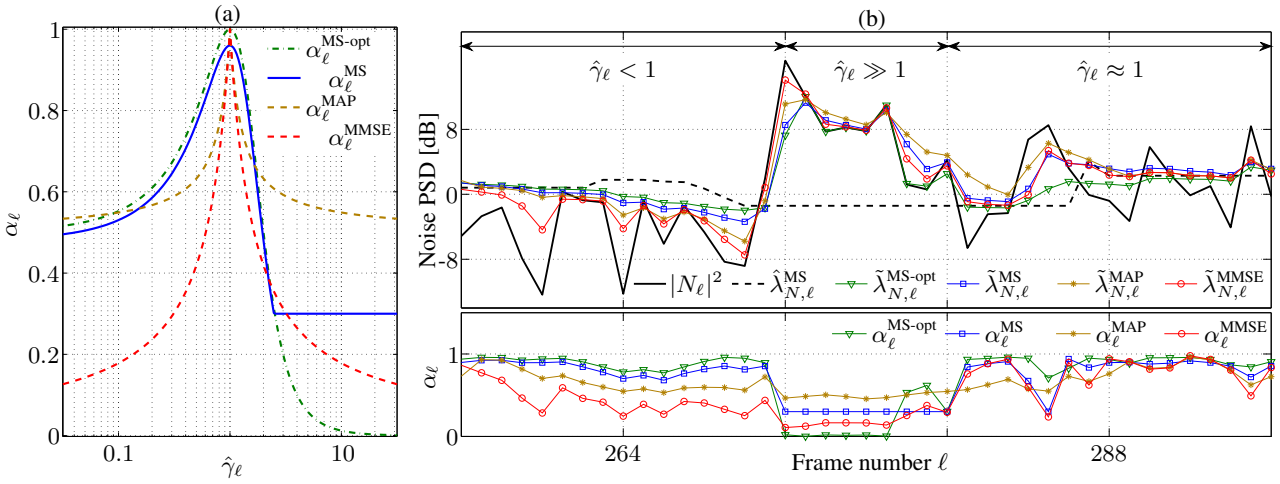

Figure 1: *(a) Smoothing parameter α_ℓ as a function of the smoothed a posteriori SNR $\hat{\gamma}_\ell$ over logarithmic scale; (b, top figure) sample noise trajectory for frequency bin 781, 25 Hz (black solid line) and trajectories of smoothed PSD of microphone signal according to eq.* (3) *for different smoothing parameters (colored lines). The dahsed line is the bias compensated output of the MS algorithm, which is used by all smoothing parameter estimators to compute $\hat{\gamma}_\ell$; (b, bottom figure) corresponding trajectories of the smoothing parameters.*

In Fig.1(a) the green dash-dotted line depicts the optimal $\alpha_\ell^{\text{MS-opt}}$ as a function of $\hat{\gamma}_\ell$ from eq. (13). However, there are two issues with this smoothing parameter [1]. The first is the so called dead lock problem for $\hat{\gamma}_\ell \approx 1$: if $\hat{\gamma}_\ell \approx 1$, then $\alpha_\ell^{\text{MS-opt}} \approx 1$ and the noise PSD estimate according to eq. (3) will hardly change and can get stuck at its current value. A second issue is the limited tracking performance in high levels of nonstationary noise: for large values of $\hat{\gamma}_\ell \gg 1$, $\alpha_\ell^{\text{MS-opt}}$ will become very small, $\tilde{\lambda}_{N,\ell}$ will almost instantly follow the observations and thus a too large estimator variance is induced.

A third, equally important issue is a weak tracking behavior of $\alpha_\ell^{\text{MS-opt}}$ for $0 < \hat{\gamma}_\ell < 1$ with the theoretical lower bound $\alpha_\ell^{\text{MS-opt}} = 0.5$ for $\hat{\gamma}_\ell = 0$. This results in a limited ability of the recursive averaging to follow the noisy speech spectrum for small power level $\tilde{\lambda}_{N,\ell} < \hat{\lambda}_{N,\ell}^{\text{MS}}$. Bins with such a low power level will most likely contain noise only. Tracking of these values is crucial, since the recursive estimates $\tilde{\lambda}_{N,\ell}$ are used for the minimum search in the MS approach, as mentioned above.

Therefore in a practical implementation Martin proposed to use the suboptimal smoothing parameter α_ℓ^{MS} according to:

$$\alpha_\ell^{\text{MS}} = \max\left(\alpha_{\text{max}} \cdot \alpha_\ell^{\text{MS-opt}},\ \alpha_{\text{min}}(\text{SNR})\right) \tag{15}$$

with an upper bound $\alpha_{\text{max}} = 0.96$ and a SNR dependent lower bound $\alpha_{\text{min}}(\text{SNR})$ calculated for positive values of the overall signal-to-noise ratios SNR > 0 measured in dB as per:

$$\alpha_{\text{min}}(\text{SNR}) = \min\left(0.3\,,\ \text{SNR}^{-\frac{R}{0.064 \cdot f_S}}\right), \tag{16}$$

where R and f_S are the frame shift (in number of samples) and the sampling frequency, respectively [1]. (16) allows α_{min} to be smaller than 0.3 for large SNR values. In the Fig.1(a) α_ℓ^{MS} is depicted by the solid blue curve for $\alpha_{\text{min}} = 0.3$. Additionally α_ℓ^{MS} is adjusted according to an error monitoring, that does not need any parameter and therefore is not taken into account in the Fig.1(a). It is a heuristic adjustment of the theoretical result to solve only the first two issues discussed above.

To overcome all mentioned issues with $\alpha_\ell^{\text{MS-opt}}$ we suggest to control the degrees of freedom parameter $\nu_{\ell-1}$ of (11) and thus the smoothing parameter α_ℓ by a measure of the instantaneous degree of nonstationarity (DN) d_ℓ according to

$$\nu_{\ell-1} = \begin{cases} 1/d_\ell & \text{for } \Delta\nu \geq 0, \\ |\Delta\nu| + 1/d_\ell & \text{for } \Delta\nu < 0. \end{cases} \tag{17}$$

where DN is defined as follows:

$$d_\ell = |\ln(\hat{\gamma}_\ell)|, \tag{18}$$

which is similar to the measure of degree of nonstationary defined in [10]. In comparision to [10] our definition of DN is more intuitive and needs no additional parameter. Inserting (18) and (17) in (11) leads to the proposed smoothing parameter:

$$\alpha_\ell^{\text{DN}} = \begin{cases} \left(1 + \frac{2 \cdot |\ln \hat{\gamma}_\ell|}{1 + \Delta\nu \cdot |\ln \hat{\gamma}_\ell|}\right)^{-1} & \text{for } \Delta\nu \geq 0, \\ \left(1 + 2 \cdot |\ln \hat{\gamma}_\ell|\right)^{-1} & \text{for } \Delta\nu < 0. \end{cases} \tag{19}$$

As a function of $\hat{\gamma}_\ell$, α_ℓ^{DN} has a single parameter $\Delta\nu$, which determines, which point estimate is used in (9) either the MAP or the MMSE estimate. The curves:

$$\alpha_\ell^{\text{MAP}} = \frac{1 + 2|\ln \hat{\gamma}_\ell|}{1 + 4|\ln \hat{\gamma}_\ell|}, \tag{20}$$

$$\alpha_\ell^{\text{MMSE}} = \frac{1}{1 + 2|\ln \hat{\gamma}_\ell|} \tag{21}$$

are depicted in the Fig.1(a) as dashed lines. It is obvious that $\alpha_\ell^{\text{DN}} = f(\hat{\gamma}_\ell)$ avoids the dead lock problem, since its first derivative $\partial\alpha_\ell^{\text{DN}}/\partial\hat{\gamma}_\ell \neq 0$ for $\hat{\gamma}_\ell = 1$. Its minimum value is calculated to $\alpha_{\text{min}}^{\text{DN}} = \Delta\nu/(\Delta\nu+2)$ for $\Delta\nu \geq 0$ and $\alpha_{\text{min}}^{\text{DN}} = 0$ else. Moreover the proposed α_ℓ^{DN} is symmetric with respect to the axis $\hat{\gamma}_\ell = 1$ and is able to follow the noisy spectrogram rapidly not only for large $\hat{\gamma}_\ell > 1$ but also for small values $\hat{\gamma}_\ell < 1$.

To justify the choice of the proposed function we first consider stationary noise in speech absence. In this case the smoothed PSD of the microphone signal is approximately equal to the noise PSD estimate delivered by the MS approach, $\tilde{\lambda}_{N,\ell} \approx \hat{\lambda}_{N,\ell}^{\text{MS}}$, resulting in $\hat{\gamma}_\ell \approx 1$ according to eq. (14). Consequently $d_\ell \approx 0$, therefore $\nu_{\ell-1} \to \infty$ and the smoothing parameter α_ℓ^{DN} assumes a value close to 1, which is desired in

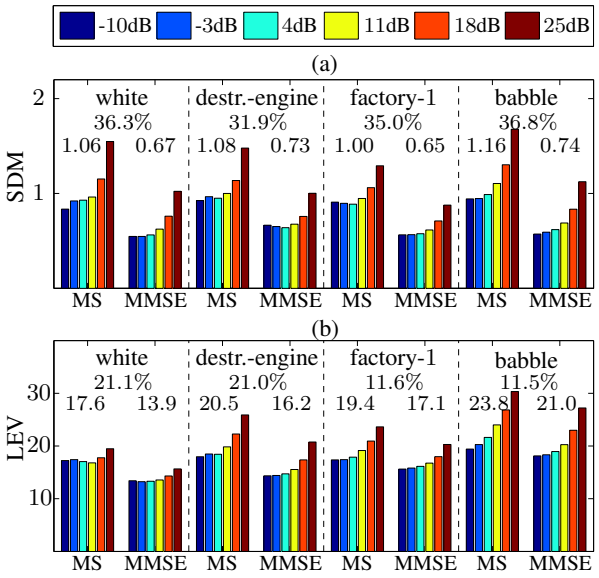

Figure 2: *Noise tracking performance of the MS approach with α_l^{MS}-based and with α_l^{MMSE}-based smoothing in terms of (a) spectral distance measure (SDM) and (b) logarithmic error variance (LEV) for 4 noise types of the NOISEX-92 database.*

stationary noise. However this is valid for quite a small range of $\hat{\gamma}_\ell$ only. For $\hat{\gamma}_\ell \neq 1$, α_ℓ^{DN} quickly decreases, and the smoothed PSD is able to follow changes in the noise spectrum, both for speech presence and for decaying noise power levels. Further α_ℓ^{DN} decreases for $\hat{\gamma}_\ell > 1$ not as rapidly as $\alpha_\ell^{\text{MS-opt}}$. So we do not expect a too large estimator variance and consequently do not need any lower bound even for $\alpha_\ell^{\text{DN}} = \alpha_\ell^{\text{MMSE}}$.

In Fig.1(b) trajectories of the noise PSD and the smoothed PSD of the microphone signal according to eq. (3) are shown for the different options for the smoothing parameter α_ℓ, as well as the trajectories of the smoothing paramters themselves. For all smoothed PSDs, the same $\hat{\lambda}_{N,\ell}^{\text{MS}}$ is used for calculating $\hat{\gamma}_\ell$. The depicted time span can be divided in 3 distinctive parts: $\hat{\gamma}_\ell < 1$, $\hat{\gamma}_\ell \gg 1$ and $\hat{\gamma}_\ell \approx 1$. The main advantage of the proposed smoothed noisy speech PSD trajectories $\tilde{\lambda}_{N,\ell}^{\text{MAP}}$ and $\tilde{\lambda}_{N,\ell}^{\text{MMSE}}$ is their ability to follow the changes of the noise power level for $\hat{\gamma}_\ell < 1$. It is due to the small values of α_ℓ^{MAP} and $\alpha_\ell^{\text{MMSE}}$, which can be seen in the bottom picture. This property can be used to better estimate the noise power floor in the MS approach. For $\hat{\gamma}_\ell \gg 1$ (power push in noise signal) all α_ℓ drop to their minimum values and try to follow the rapid changes in noise power. Bacause of the relative high minimum value of $\alpha_{\text{min}}^{\text{MAP}} = 0.5$, $\tilde{\lambda}_{N,\ell}^{\text{MAP}}$ performs not well enough in such time segments. In the last time interval with $\hat{\gamma}_\ell \approx 1$ the dead lock problem of $\alpha_\ell^{\text{MS-opt}}$ can be observed: $\tilde{\lambda}_{N,\ell}^{\text{MS-opt}}$ hardly follows the noise trajectory.

4. Experimental results

The performance of the proposed smoothing parameter $\alpha_\ell^{\text{MMSE}}$ versus conventional α_ℓ^{MS} is experimentally evaluated employing speech degraded by various noise types and at global SNR values varied from -10 dB to 25 dB in steps of 7 dB. The clean speech signals are generated by concatenating utterances of male and female speakers from the TIMIT database [11], having a total length of 3 minutes. To them the noise signals of

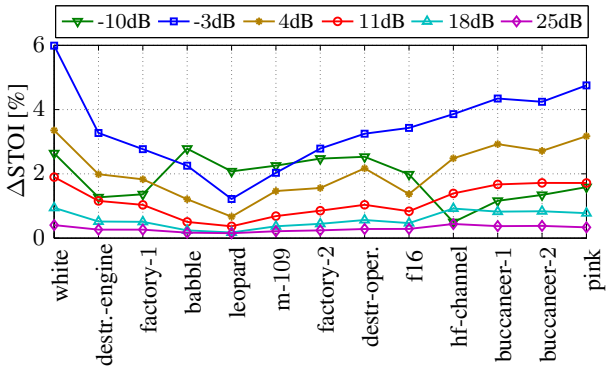

Figure 3: *Improvement of the short-time objectve intelligibility (STOI) measure* $\Delta\text{STOI}=\text{STOI}^{\text{MMSE}}-\text{STOI}^{\text{MS}}$ *of the enhanced signals obtained by using the MS approach with α_l^{MMSE}-based versus with α_l^{MS}-based smoothing for NOISEX-92 database.*

13 different noise types, which were taken from the NOISEX-92 database [12], were artificially added. All signals are sampled at 16 kHz. The STFT spectral analysis used a Hann window of 1024 samples length with a frame overlap of 50%.

Since the implementation of the MS approach of [1] was available, the proposed function $\alpha_\ell^{\text{MMSE}}$ from (21) is simply integrated in the MS implementation. The length of the MS window for minimum search is set to $D = U \cdot V = 96$ frames devided into $U = 8$ subwindows of the length of $V = 12$ frames. The performance of the MS-based noise tracking with either α_ℓ^{MS}-based or $\alpha_\ell^{\text{MMSE}}$-based smoothing of the noisy speech PSD is given in Fig. 2 in terms of the spectral distance measure (SDM) [13] and logarithmic error variance (LEV) [14], where the true noise periodogram $|N_{\ell,k}|^2$ is used as a reference noise PSD. We observed that both the estimator errors measured by SDM and the estimator variances measured by LEV reduced for $\alpha_\ell^{\text{MMSE}}$-based smoothing versus the α_ℓ^{MS}-based smoothing for all SNR values and for all considered noise types of the NOISEX-92 database, of which only 4 representatives are depicted in Fig. 2. The absolute values and the reductions given in % of SDM and LEV are the averages over all SNR values.

Further both noise trackers are combined with the optimally-modified log-spectral amplitude (OM-LSA) estimator to obtain the enhanced speech signals [15]. Fig. 3 shows the improvement of the short-time objectve intelligibility (STOI) measure $\Delta\text{STOI} = \text{STOI}^{\text{MMSE}} - \text{STOI}^{\text{MS}}$ of enhanced signals [16] obtained by using the MS approach with α_l^{MMSE}-based versus with α_l^{MS}-based smoothing. The improvements for the NOISEX-92 database are small, however consistent over all noise types and SNR values. As expected the best over all SNR values averaged improvements of 2.5% and 2% are achieved for 'white' and 'pink' noise types, respectively.

5. Conclusions

In this contribution we proposed an alternative computation of the smoothing parameter, which is used to smooth the PSD of the microphone signal at the input of the Minimum Statistics based noise tracking. The smoothing parameter is controlled by a degree of nonstationary measure. Its single parameter, $\Delta\nu$, is chosen to realize an MMSE estimate of the noisy speech PSD. The experimental evaluation showed that the performance of the MS-based noise tracking was improved for all noise types of the NOISEX-92 database and all tested SNR values, compared to the use of the original smoothing parameter as proposed in [1].

6. References

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

[2] ——, "Spectral subtraction based on minimum statistics," *In Proc. of the European Signal Processing Conference (EUSIPCO)*, pp. 1182–1185, Sept 1994.

[3] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 153–156, May 1995.

[4] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," *In Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1513–1516, Sept 1995.

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

[6] 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, Jan 2002.

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

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

[9] A. Chinaev, A. Krueger, D. H. T. 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.

[10] M.-S. Choi and H.-G. Kang, "A two-channel noise estimator for speech enhancement in a highly nonstationary environment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 905–915, May 2011.

[11] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[12] 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," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[13] B. Iser and G. Schmidt, "Bewertung verschiedener Methoden zur Erzeugung des Anregungssignals innerhalb eines Algorithmus zur Bandbreitenerweiterung," Kiel (Germany), April 2006. [Online]. Available: http://it.e-technik.uni-ulm.de/World/Research.DS/publications/2006bi01.pdf

[14] 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," *In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4640–4643, May 2011.

[15] I. Cohen, "On speech enhancement under signal presence uncertainty," *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 661–664, May 2001.

[16] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.