



Speech Enhancement with Improved A Posteriori SNR Computation

Suhadi Suhadi, Tim Fingscheidt

Institute for Communications Technology, Braunschweig Technical University,
Schleinitzstr. 22, D – 38106 Braunschweig, Germany

{s.suhadi, t.fingscheidt}@tu-bs.de

Abstract

In speech enhancement, the decision-directed (DD) approach to compute the *a priori* SNR is often used to reduce the musical tones. However, the constant DD weighting factor very close to one results in more speech distortion during transitional speech segments. Contrarily, a time-varying weighting factor gives less speech distortion but with more residual noise in speech pause. In this contribution we present a new *a posteriori* SNR computation to relax the dependence on the decision-directed weighting factor. By computing the *a posteriori* SNR with a time-varying weighting factor, we actually derive a correction factor to the time-varying DD weighting factor resulting in less speech distortion during transitions, *as well as* less residual noise in speech pause.

Index Terms: speech enhancement, decision-directed approach

1. Introduction

In speech enhancement, one aims at reducing background noise to a desired level while preserving speech content as much as possible (i.e., low speech distortion). However, because of the imperfect noise estimation, it is difficult to achieve both goals simultaneously. This situation is occasionally worsened as the noise spectra are often *locally* not well-suppressed (i.e., too low *local* noise estimate or not enough noise suppression from the weighting rule at low SNR) leading to the occurrence of unnatural sounding musical tones, which mostly distracts listeners' attention.

As a practical solution, several ways have been proposed to find an optimal trade-off between the residual noise and the speech distortion, as well as to reduce the musical tones as much as possible. Berouti et al. [1] introduced two parameters, i.e., noise-overestimation and spectral flooring, to aid attaining the optimal trade-off. Along with it, the decision-directed (DD) approach [2] to *a priori* SNR computation is commonly applied to reduce the musical tones.

In the DD approach, the *a priori* SNR evolves smoothly over time. A constant weighting factor very close to one, i.e., in the range of 0.96-0.99, is selected. Accordingly, the variance of the *a priori* SNR is greatly reduced, which helps to reduce the musical tones significantly [3]. Unfortunately, this approach leads to an unexpected transient distortion to the enhanced signal. A time-varying DD weighting factor as a function of the spectral change [4] or of the *a priori* SNR change [5] has been proposed to overcome this problem. Although these approaches reduce the transient distortion, they do not yield a sufficient noise attenuation in speech pauses.

In this paper we propose a different possibility to obtain a high noise (and musical tones) attenuation level while preserving the speech, especially during transition. Hence, we apply the DD approach with the time-variant DD weighting fac-

tor to reduce the transient distortion. Along with it, we propose a new computation of the *a posteriori* SNR with a time-varying weighting factor. The time-varying weighting factor is selected so that the *a posteriori* SNR will be highly smoothed in speech absence to keep the residual noise reduced. Meanwhile, in speech presence the *a posteriori* SNR tends towards the classical formulation.

As an evaluation framework, we employ two different weighting rules, i.e., the Wiener filter and the joint maximum a posteriori (JMAP) estimator based on supergaussian speech modelling. The first weighting rule is computed based on the *a priori* SNR only [6], while the latter is computed based on both *a priori* and *a posteriori* SNRs [7]. These weighting rules realized with the classical SNR computations will be shortly recapitulated in section 2. An outline of our approach follows then in section 3. Finally, in section 4 we present the performance of the proposed algorithm in informative auditive and objective measurement tests.

2. Speech Enhancement in General

After short-time Fourier transform of length K , the clean speech spectrum subject to additive noise at frame l and frequency bin k can be expressed as $Y_l(k) = X_l(k) + N_l(k)$, $k = 1, \dots, K$. The time domain signals of noisy speech, clean speech, and noise are $y(n)$, $x(n)$, and $n(n)$, respectively.

2.1. Classical SNR Computation

Following the noise estimation, the resulting noise spectral variance $\lambda_{N_l(k)}$ and the noisy speech spectrum $Y_l(k)$ are used to compute the *a posteriori* SNR ('Classical-SNRpost')

$$\gamma_l(k) = \frac{|Y_l(k)|^2}{\lambda_{N_l(k)}} \quad (1)$$

According to the decision-directed (DD) approach of Ephraim and Malah [2], the *a priori* SNR $\xi_l(k)$ then is estimated ('Classical-DD')

$$\begin{aligned} \xi'_l(k) &= \beta \frac{|\hat{X}_{l-1}(k)|^2}{\lambda_{N_{l-1}(k)}} + (1 - \beta)P[\gamma_l(k) - 1], \\ \xi_l(k) &= \max \{ \xi'_l(k), \xi_{\min} \}, \end{aligned} \quad (2)$$

with half-wave rectification function $P[\cdot]$ and the parameters β and ξ_{\min} being set to 0.98 and -15 dB, respectively.

Alternatively, the DD weighting factor β could also be a time-varying parameter, i.e., as a function of the *a priori* SNR change [5] ('Time-varying-DD')

$$\beta_l(k) = \left[1 + \left(\frac{\xi_l(k) - \tilde{\xi}_{l-1}(k)}{\xi_l(k) + 1} \right)^2 \right]^{-1} \quad (3)$$

with $\xi_l(k) \approx P[\gamma_l(k) - 1]$ and $\tilde{\xi}_{l-1}(k) = \frac{|\hat{X}_{l-1}(k)|^2}{\lambda_{N_{l-1}(k)}}$. When the signal is relatively constant, the DD weighting factor $\beta_l(k)$ will attain a value close to 1 and therefore any weighting process will not severely distort the signal. In contrary, if there is an abrupt change (which is most likely a speech transition), the parameter $\beta_l(k)$ is lowered to a value close to 0, and accordingly the signal will be less distorted.

2.2. Weighting Rule Computation

By minimizing the mean square error (MMSE) between clean speech spectrum and its estimate, the Wiener filter can be computed by using the *a priori* SNR [6] ('WF-DD')

$$G_l(k) = \frac{\xi_l(k)}{\xi_l(k) + 1} \quad (4)$$

Meanwhile, by modelling the pdf of clean speech spectral amplitudes $|X_l(k)|$ as supergaussian, the joint-MAP (JMAP) estimator can be obtained as follows [7] ('SG-JMAP')

$$G_l(k) = u + \sqrt{u^2 + \frac{v}{2\gamma_l(k)}} \quad (5)$$

$$u = \frac{1}{2} - \frac{w}{4\sqrt{\gamma_l(k)\xi_l(k)}}.$$

The parameters w and v determine the shape of the supergaussian pdf. Their optimal values for clean speech spectral amplitudes are reported as 1.74 and 0.126, respectively [7].

3. New *A posteriori* SNR Computation

According to Cappé's analysis [3], the selection of $\beta = 0.98$ in (2) aims at limiting the effect of the highly fluctuating *a posteriori* SNR. In this manner, it smoothes the *a priori* SNR values over time that helps to suppress the musical tone artifacts. Nevertheless, the high dependence on the previous frame distorts the speech severely in the transition from speech absence to speech presence, or vice versa (transient distortion).

This transient distortion can be minimized by applying a time-varying DD weighting factor $\beta_l(k)$. The parameter $\beta_l(k)$ is normally computed based on the spectral change [4] or on the *a priori* SNR change [5], which unfortunately characterizes the residual noise in speech absence as well. As the consequence, the time-varying $\beta_l(k)$ leads to poorer noise attenuation in speech pause, although it successfully reduces the transient distortion. In this case, a noise-overestimation can be applied to reduce the increased residual noise, but it again results in higher speech distortion.

In this paper, we are aiming at reducing the transient distortion while preserving the noise (and musical tones) attenuation level in speech absence. Hence, we avoid the highly direct-weighting of the *a priori* SNR, i.e., by applying (3) to decrease the parameter β during transitional speech segments. To keep the smoothness of the *a priori* SNR, we subsequently compensate the decrease of the parameter β by employing an additional weighting on the *a posteriori* SNR with time-varying weighting factor.

Still identical to (1), we formulate the new *a posteriori* SNR as follows

$$\gamma'_l(k) = \gamma_l(k) = 1 + \frac{|Y_l(k)|^2 - \lambda_{N_l(k)}}{\lambda_{N_l(k)}}. \quad (6)$$

If we ensure that the second summand in (6) is positive, we can interpret it as some kind of *a priori* SNR. Now we follow the

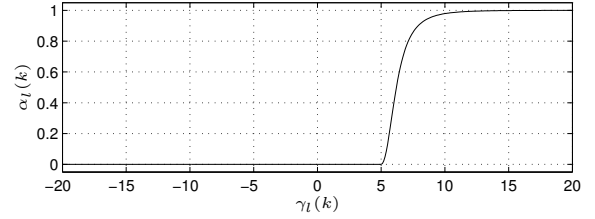


Figure 1: An example of $\alpha_l(k)$ function with $\gamma_{min} = 5$ dB.

conceptual design of the DD approach for the *a priori* SNR and apply it to the *a posteriori* SNR in (6). With the coarse approximation $|\hat{X}_{l-1}(k)|^2 \approx G_{l-1}^2(k)|\hat{Y}_l(k)|^2$, the new *a posteriori* SNR shall be written as ('Smoothed-SNRpost')

$$\gamma'_l(k) = 1 + \alpha_l(k) \frac{P[|Y_l(k)|^2 - \lambda_{N_l(k)}]}{\lambda_{N_l(k)}} + (1 - \alpha_l(k)) \frac{|\hat{X}_{l-1}(k)|^2}{\lambda_{N_l(k)}} \approx 1 + \alpha_l(k) P[\gamma_l(k) - 1] + (1 - \alpha_l(k)) G_{l-1}^2(k) \gamma_l(k), \quad (7)$$

where $\gamma_l(k)$ is computed via (1). Instead of using the *a priori* SNR of the previous frame, we employ the squared weighting rule of the previous frame $G_{l-1}^2(k)$ to smooth the fluctuation of the classical *a posteriori* SNR.

To achieve our aims, the weighting factor $\alpha_l(k)$ must be selected carefully. In speech pauses it must be kept as low as possible such that the smoothed *a posteriori* SNR suppresses the variance of the *a priori* SNR leading to strong noise attenuation. On the other hand, it must be as high as possible during the speech transition and presence segments in order to reduce the transient distortion. Hence, we propose the following function

$$\alpha_l(k) = \frac{P^2[\gamma_l(k) - \gamma_{min}]}{1 + P^2[\gamma_l(k) - \gamma_{min}]} \quad (8)$$

to compute the weighting factor $\alpha_l(k)$. In Fig. 1, a plot of the $\alpha_l(k)$ function is depicted with $\gamma_{min} = 5$ dB. The parameter γ_{min} principally acts as an instantaneous noise-overestimation so that only observation data with high local SNR (i.e., in speech presence) will be selected to update the *a posteriori* SNR. Otherwise, the *a posteriori* SNR will be highly smoothed by the weighting rule of the previous frame and accordingly the residual noise in speech absence is suppressed.

Let us observe the impact of the new *a posteriori* SNR computation to the DD approach. By inserting (7) into (2), we will easily obtain the following approximation

$$\xi'_l(k) \approx \beta'_l(k) \frac{|\hat{X}_{l-1}(k)|^2}{\lambda_{N_{l-1}(k)}} + (1 - \beta'_l(k)) P[\gamma_l(k) - 1] \quad (9)$$

with

$$\beta'_l(k) = 1 - (1 - \beta_l(k)) \alpha_l(k) \quad (10)$$

where $\beta_l(k)$ is computed according to (3). In speech pause the value of $\beta'_l(k)$ mostly equals to one considering that $\alpha_l(k) \approx 0$. The residual noise thus is greatly suppressed in speech absence disregarding the value of $\beta_l(k)$. In contrary, during transitional segments, due to $\alpha_l(k) \lesssim 1$ the value of $\beta'_l(k) \approx \beta_l(k)$ becomes very low (even sometimes close to 0) and the transient distortion is correspondingly reduced. Finally, in speech presence, as $0 < \beta_l(k) < 1$ and $\alpha_l(k) = 1$, the *a posteriori* SNR is kept updated greatly based on the observation data.

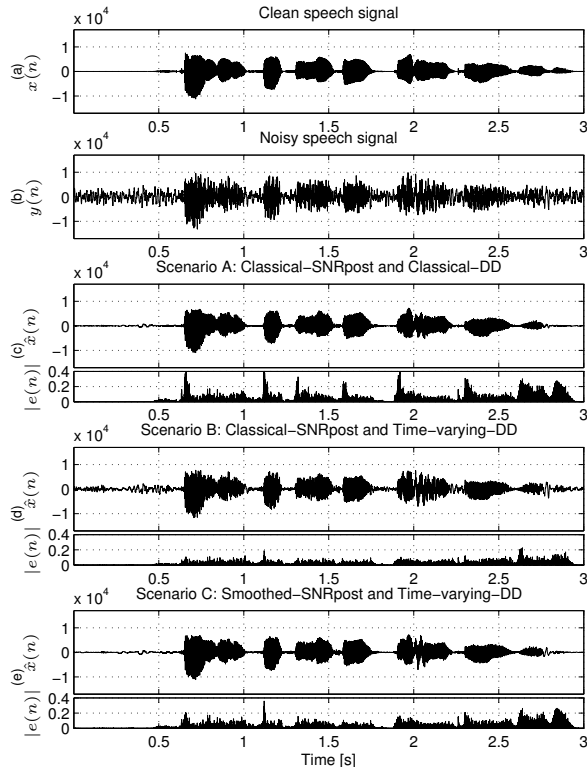


Figure 2: (a) Clean speech signal, (b) noisy speech signal, and (c)-(e) enhanced signals $\hat{x}(n)$ by the *a priori* SNR-driven Wiener filter along with modulus of error signal $|e(n)|$ between clean speech signal $x(n)$ and filtered clean speech signal $\hat{x}(n)$ subject to the obtained spectral magnitude gain values.

4. Experimental Result

We evaluate the performance of the proposed approach in car noise with the following experimental setting: Noise estimation for all compared approaches is performed via minimum statistics [8]. The DFT length is $K = 256$, and the window function is a flat-top Hann window with 40 samples rising edge, 120 samples being 1, and 40 samples trailing edge. This makes a frame length of 200 samples and a frame shift of 160 samples.

The performance assessment is performed based on enhanced signals after applying a testing weighting rule incorporated with three different SNR computations: In the scenarios A and B, we employ the classical *a posteriori* SNR computation (1). The *a priori* SNR computation is performed via the DD approach (2) with constant DD weighting factor $\beta = 0.98$ for scenario A, or with the time-varying DD weighting factor $\beta_l(k)$ computed according to (3) for scenario B. Scenario C equals scenario B, however, our proposed approach (7) is employed to compute the *a posteriori* SNR with the parameter γ_{min} being set to 5 dB.

In the first experiment, we test our algorithm on different utterances and conduct informal listening tests. The *a priori* SNR-driven Wiener filter (4) is chosen as the testing weighting rule. For a preliminary experiment, we firstly investigate the performance of the proposed algorithm being combined with the DD approach (2) using different constant values of β , $0 \leq \beta < 1$. As the parameter β is reduced to a value closer to 0, it turns out that our proposed algorithm still can give a sufficient noise attenuation level in speech pause even in case of $\beta =$

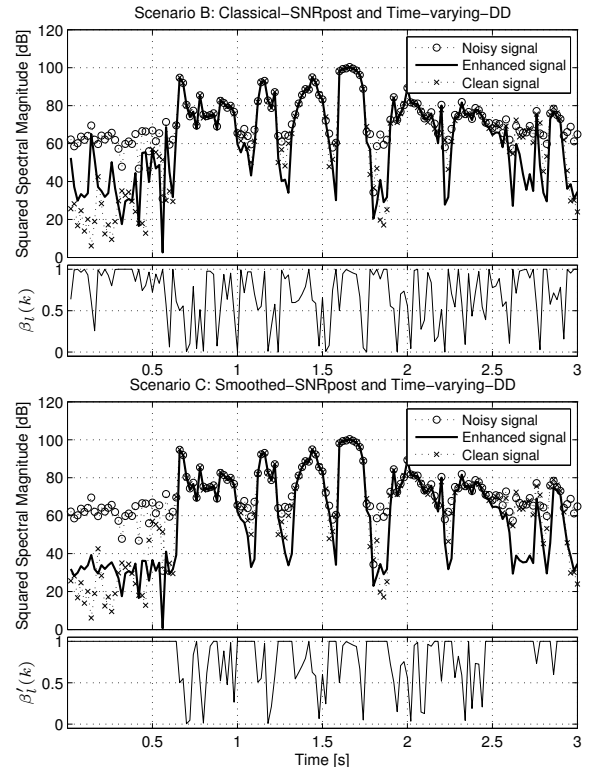


Figure 3: Temporal variation of the DD weighting factor of signal spectra for scenario B (upper) and scenario C (lower) at frequency $f = 500$ Hz.

0. Unfortunately, if $\beta > 0.9$, it turns out that it distorts the speech more than the system with the classical *a posteriori* SNR computation does.

Next, we conduct informal listening tests for all scenarios. Scenario A with $\beta = 0.98$ indeed reduces the musical tones significantly, yet it leads to the transient distortion. After applying the time-varying DD weighting factor, the speech distortion (especially the transient distortion) can be decreased. As our expectation, in scenario C the noise (and musical tones) attenuation level in speech pause is still preserved relatively high, while scenario B results in more residual noise.

In Fig. 2, the enhanced signals for each scenario are shown, along with the clean speech and noisy speech signals. To observe the speech distortion, the magnitude of the corresponding error signals $e(n) = x(n) - \hat{x}(n)$ is intentionally plotted, where $x(n)$ and $\hat{x}(n)$ denote the clean speech signal and the filtered clean speech signal component subject to the obtained spectral magnitude gain values, respectively. It clearly shows that our proposed approach (Fig. 2(e)) still can maintain the noise attenuation level as good as that of scenario A (Fig. 2(c)). Apart from it, it gives a comparable performance to that of scenario B (Fig. 2(d)) in reducing the transient distortion that always occurs at the beginning of the utterance. The time-varying $\alpha_l(k)$ in (7) indirectly acts as a correction factor to improve the computation of the time-varying DD weighting factor $\beta_l(k)$ in reducing the residual noise in speech absence. This is clearly seen at the beginning of the utterance when comparing the associated temporal variation of the DD weighting factor computed via (3) for scenario B and via (9) for scenario C, as it is shown in Fig. 3.

In our second experiment, we objectively assess the perfor-

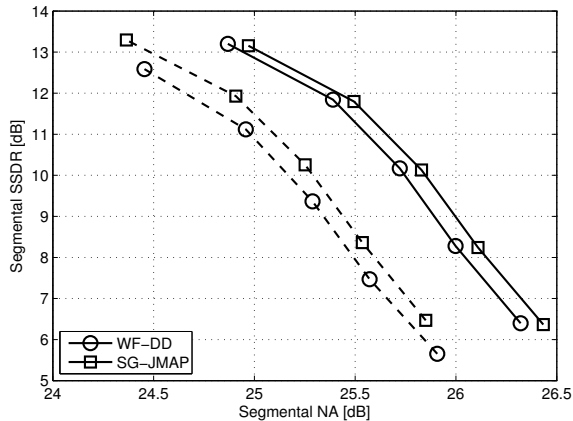


Figure 4: Segmental SSSDR vs. segmental NA for the whole utterance: Scenario A (dashed) and scenario C (solid)

mance of our proposed algorithm measured in terms of segmental noise attenuation (segmental NA) and segmental speech-to-speech distortion ratio (segmental SSSDR). For both quantities, the filtered clean speech signal $\tilde{x}(n)$ and the filtered noise signal $\tilde{n}(n)$ are to be computed based on the obtained spectral magnitude gain values. The segmental NA is then computed as

$$\text{NA}_{\text{seg}} = 10 \log_{10} \left[\frac{1}{L} \sum_l \text{NA}(l) \right]$$

$$\text{NA}(l) = \frac{\sum_{\nu=0}^{N-1} n^2(\nu+lN)}{\sum_{\nu=0}^{N-1} \tilde{n}^2(\nu+lN)}, \quad (11)$$

with L being the total number of frames. In analogy, the segmental SSSDR is computed as

$$\text{SSDR}_{\text{seg}} = \frac{1}{L} \sum_l \text{SSDR}(l)$$

$$\text{SSDR}(l) = \min \left\{ 10 \log_{10} \left[\frac{\sum_{\nu=0}^{N-1} x^2(\nu+lN)}{\sum_{\nu=0}^{N-1} e^2(\nu+lN)} \right], 30 \text{ dB} \right\}$$

$$e(\nu+lN) = \tilde{x}(\nu+lN) - x(\nu+lN). \quad (12)$$

Please note that only those frames with $\text{SSDR}(l) > -10$ dB are considered in (12) to avoid any frames with extremely low speech energy being involved.

For this objective evaluation, 20 different utterances comprising of 4 different speakers (2 male and 2 female) and 42 car noise signals are taken from the NTT-AT speech and noise databases. After combination, $20 \times 42 = 840$ noisy speech utterances at $f_s = 8$ kHz are obtained as testing signals. As the testing weighting rules, we employ the *a priori* SNR-driven Wiener filter (4) ('WF-DD'), and the JMAP estimator based on supergaussian speech modelling (5) ('SG-JMAP').

Fig. 4 shows the performance of the testing weighting rules in scenarios A (dashed line) and C (solid line). The more a curve is located in the upper right of the figure, the less speech distortion and residual noise remain, and the better the algorithm performs. For both testing weighting rules, our proposed algorithm can generally give better performance than the reference system, which employs the classical *a posteriori* SNR computation. Our proposed approach yields 1 dB less in speech distortion for the Wiener filter. With this improvement, the Wiener filter can achieve about the same speech preservation level as the SG-JMAP estimator. Meanwhile, for the SG-JMAP estimator, the proposed algorithm results in about the same speech

preservation, but it can improve the noise attenuation performance. Scenario B can indeed give about 1 dB less speech distortion than our proposed algorithm, however it severely fails to suppress the residual noise, which results in 6 dB less noise attenuation than the proposed algorithm. Therefore we have not included it in Fig. 4.

5. Conclusion

In this paper we address a new *a posteriori* SNR computation to improve the performance of the decision-directed approach in computing the *a priori* SNR. Along with the recently proposed time-varying decision-directed weighting factor, we propose to perform an additional weighting in the *a posteriori* SNR computation also with a time-varying weighting factor. This additional weighting effectively results in a correction factor to the decision-directed approach yielding less speech distortion during transitional speech segments and less residual noise in speech absence.

6. References

- [1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted By Acoustic Noise," in *Proc. of ICASSP'79*, Apr 1979, pp. 208–211.
- [2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
- [3] O. Cappé, "Elimination of the Musical Noise Phenomenon With the Ephraim and Malah Noise Suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
- [4] C. Beaugeant and P. Scalart, "Noise Reduction Using Perceptual Spectral Change," in *Proc. of EUROSPEECH'99*, Budapest, Hungary, Sept. 1999, pp. 2543–2546.
- [5] M.K. Hasan, S. Salahuddin, and M.R. Khan, "A Modified A Priori SNR for Speech Enhancement Using Spectral Subtraction Rules," *IEEE Signal Processing Letters*, vol. 11, no. 4, pp. 450–453, Apr. 2004.
- [6] P. Scalart and J.V. Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP'96*, Atlanta, GA, May 1996, pp. 629–632.
- [7] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 1110–1126, 2005.
- [8] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 9, no. 5, pp. 504–512, July 2001.