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 variability 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 factor 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 \( Y_l(k) = X_l(k) + N_l(k), k = 1, \ldots, K \). 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’)

\[
\gamma_l(k) = \frac{|Y_l(k)|^2}{\lambda_{N_l(k)}}. \tag{1}
\]

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

\[
\xi_l(k) = \beta \frac{|\tilde{X}_{l-1}(k)|^2}{\lambda_{N_{l-1}(k)}} + (1 - \beta)P[\gamma_l(k) - 1], \tag{2}
\]

with half-wave rectification function \( P[.] \) and the parameters \( \beta \) and \( \xi_{\text{min}} \) being set to 0.98 and -15 dB, respectively.

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

\[
\beta_l(k) = \left[ 1 + \left( \frac{\xi_l(k) - \xi_{l-1}(k)}{\xi_l(k) + 1} \right)^2 \right]^{-1}. \tag{3}
\]
with $xi(k) \approx P[\gamma_1(k) - 1]$ and $\hat{xi}_1(k) = \frac{|X_i(k)|^2}{\lambda_{N_i}(k)}$. When the signal is relatively constant, the DD weighting factor $\beta_i(k)$ 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 $\beta_i(k)$ 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’)

$$G_i(k) = \frac{\xi_i(k)}{\xi_i(k) + 1}$$

(4)

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

$$G_i(k) = u + \frac{v}{2\gamma(k)}$$
$$u = \frac{1}{2} \left( 1 - \frac{w}{4\sqrt{\gamma(k)}\xi_i(k)} \right)$$

(5)

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 Cappe’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 smooths the a priori SNR values over time that helps to suppress the musical tone artefacts. 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 $\beta_i(k)$. The parameter $\beta_i(k)$ 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 $\beta_i(k)$ leads to poorer noise attenuation in speech pause, although it successfully reduces the transient distortion. In this case, a noise-oversaturation 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 $\beta$ during transitional segments. To keep the smoothness of the a priori SNR, we subsequently compensate the decrease of the parameter $\beta$ 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

$$\gamma_i(k) = \gamma_i(k) + 1 + 2\gamma_i(k)\frac{|X_i(k)|^2}{\lambda_{N_i}(k)} - \frac{|X_i(k)|^2}{\lambda_{N_i}(k)} + P[\gamma_i(k) - 1]$$

(6)

If we ensure that the second summand in (6) is positive, we can interprete it as some kind of a priori SNR. Now we follow the 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 $|X_i(k)|^2 \approx G_i(k)\gamma_i(k)$, the new a posteriori SNR will be formulated as (‘Smoothed-SNRpost’)

$$\gamma'_i(k) = 1 + \alpha_i(k)P[\gamma_i(k) - 1] + \frac{1 - \alpha_i(k)}{\lambda_{N_i}(k)}G_i(k)\gamma_i(k),$$

(7)

where $\gamma_i(k)$ 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 $G_i(k)$ to smooth the fluctuation of the classical a posteriori SNR.

To achieve our aims, the weighting factor $\alpha_i(k)$ 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

$$\alpha_i(k) = \frac{P[\gamma_i(k) - \gamma_{\min}]}{1 + P[\gamma_i(k) - \gamma_{\min}]}$$

(8)

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

to compute the weighting factor $\alpha_i(k)$. In Fig. 1, a plot of the $\alpha_i(k)$ function is depicted with $\gamma_{\min} = 5$ dB. The parameter $\gamma_{\min}$ 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

$$\hat{xi}(k) \approx \beta_i(k)\frac{|\hat{XY}_{i-1}(k)|^2}{\lambda_{N_i}(k)} + (1 - \beta_i(k))\frac{|\hat{XY}_{i-1}(k)|^2}{\lambda_{N_i}(k)}$$

(9)

with

$$\beta_i(k) = 1 - (1 - \beta_i(k))\alpha_i(k)$$

(10)

where $\beta_i(k)$ is computed according to (3). In speech pause the value of $\beta_i(k)$ mostly equals to one considering that $\alpha_i(k) \approx 0$. The residual noise thus is greatly suppressed in speech absence disregarding the value of $\beta_i(k)$. In contrary, during transitional segments, due to $\alpha_i(k) \approx 1$ the value of $\beta_i(k)$ becomes very low (even sometimes close to 0) and the transient distortion is correspondingly reduced. Finally, in speech presence, as $0 < \beta_i(k) < 1$ and $\alpha_i(k) = 1$, the a posteriori SNR is kept updated greatly based on the observation data.
\[ \beta \text{ scenario A, or with the time-varying DD weighting factor a posteriori compute the} \]
\[ \beta \text{ scenario B, however, our proposed approach (7) is employed to computed according to (3) for scenario B. Scenario C equals approach (2) with constant DD weighting factor } \beta \text{ a priori rated with three different SNR computations: In the scenari os A} \]
\[ \text{enhanced signals after applying a testing weighting rule inco rpor-}
\[ \text{frame length of 200 samples and a frame shift of 160 samples. This makes a}
\[ \text{Wiener filter along with modulus of error signal } |e(n)| \text{ between clean speech signal } x(n) \text{ and filtered clean speech signal } \tilde{x}(n) \text{ subject to the obtained spectral magnitude gain values.} \]

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 statis-
\[ \beta \text{ set to } 5 \text{ dB.} \]
\[ \text{We evaluate the performance of the proposed approach in car}
\[ \text{utterances and conduct informal listening tests. The } \beta \text{ with }
\[ \beta > 0.9, \text{ it turns out that it distorts the}
\[ \text{Next, we conduct informal listening tests for all scenarios. Scenario A with } \beta = 0.98 \text{ indeed reduces the musical tones significantly, yet it leads to the transient distortion. After applying the time-varying DD weighting factor, the speech distortion (especially the transient distortion) can be decreased. As our expectation, in scenario C the noise (and musical tones) attenuation level in speech pause is still preserved relatively high, while scenario B results in more residual noise.} \]
\[ \text{In Fig. 2, the enhanced signals for each scenario are shown, along with the clean speech and noisy speech signals. To ob-
\[ \text{In our second experiment, we objectively assess the perfor-
\[ \text{In the first experiment, we test our algorithm on different ut-}
\[ \text{As the parameter } \beta \text{ is reduced to a value closer to 0, it}
\[ \text{shows that our proposed algorithm still can maintain the noise a-}
\[ \text{In the second experiment, we objectively assess the perfor-} \]
performance of our proposed algorithm measured in terms of segmental noise attenuation (segmental NA) and segmental speech-to-speech distortion ratio (segmental SSDR). For both quantities, the filtered clean speech signal \( \tilde{x}(n) \) and the filtered noise signal \( \hat{n}(n) \) are to be computed based on the obtained spectral magnitude gain values. The segmental NA is then computed as

\[
NA(l) = \frac{\sum_{\nu=0}^{N-1} n^2(\nu+lN)}{\sum_{\nu=0}^{N-1} e^2(\nu+lN)}
\]

with \( L \) being the total number of frames. In analogy, the segmental SSDR is computed as

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

Please note that only those frames with 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 NTU-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 \textit{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 \textit{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 \textit{a posteriori} SNR computation to improve the performance of the decision-directed approach in computing the \textit{a priori} SNR. Along with the recently proposed time-varying decision-directed weighting factor, we propose to perform an additional weighting in the \textit{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