An Iterative Speech Model-Based A Priori SNR Estimator

Samy Elshamy1, Nilesh Madhu2, Wouter Tirry2, Tim Fingscheidt1

1Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, D–38106 Braunschweig, Germany
2NXP Software
Interleuvenlaan 80, B–3001 Leuven, Belgium

{s.elshamy,t.fingscheidt}@tu-bs.de, {nilesh.madhu,wouter.tirry}@nxp.com

Abstract

In this contribution we propose an a priori signal-to-noise ratio (SNR) estimator based on a probabilistic speech model. Since the a priori SNR is an important means for speech enhancement algorithms, such as weighting rule calculation for noise reduction or speech presence probability computation, its diligent estimation is of wide interest. As a basis for this estimator a Gaussian mixture model (GMM) is trained on clean speech amplitudes and by finding the maximum likelihood (ML) clean speech estimate of the corresponding observed frame the a priori SNR can easily be calculated. Additionally, an iterative scheme is applied to consequently enhance the estimate by repetitively evaluating the GMM. This technique allows to accomplish noise reduction free of musical tones even in non-stationary noise environments and exceeds the quality of the classical decision-directed (DD) approach for typical spectral weighting rules.

Index Terms: speech enhancement, a priori SNR estimation

1. Introduction

The estimation of the a priori SNR has been subject to a number of publications in the past, e.g., [1, 2, 3, 4, 5]. It is an important entity for most speech enhancement applications such as spectral weighting rules for noise reduction algorithms [1, 2, 6, 7], means of speech presence probability estimation [8], and voice activity detection [9], for example. Since most of the weighting rules are functions of the a priori SNR and a posteriori SNR, with the a priori SNR having the stronger impact, it is important to have a good a priori SNR estimate at hand. The quality of the estimator influences the amount of introduced musical tones, speech distortion, and the degree of achieved noise suppression.

The most common and famous way is the DD approach by Ephraim and Malah [1] where a weighted sum of two components yields the desired estimate. Both components are representing a priori SNR-like entities where the first depends on previous estimates and the other on the current observation. Since the weights add up to one and the weight for the first component is mostly close to unity, the a priori SNR estimate is strongly influenced by the previous frame. As a consequence the DD approach deteriorates once sudden changes of the instantaneous a priori SNR occur [10, 11].

Martin et al. [3] propose a smoothing algorithm in the cepstral domain where the cepstrum of the ML a priori SNR is smoothed with an adaptive frequency bin-dependent smoothing factor which is adjusted depending on where the spectral envelope and the excitation is expected. Therefore the algorithm is relying on a fundamental frequency estimator. The authors are able to show that cepstral smoothing is superior to temporal smoothing since the cepstrum offers better abilities to individually smooth coefficients related to noise (stronger smoothing applied) and coefficients representing speech (weaker smoothing applied), and thus maintaining a better speech component while achieving high noise attenuation simultaneously.

Two data-driven approaches are introduced by Fingscheidt et al. [4] in 2011 where one utilizes two neural networks in order to estimate the a priori SNR on behalf of the two smoothing components of the DD approach. One network is trained under the hypothesis of speech presence and the other under speech absence. The outputs of the neural networks are finally smoothed with a smoothing factor based on an internal speech absence probability yielding the final a priori SNR estimate.

Our proposed approach is a continuation of the work of Mowlace et al. [12] which can be interpreted as a noise reduction algorithm that we modify to provide an a priori SNR estimate. Our algorithm is designed as a two-step procedure where the first stage is a classical noise reduction composed of a noise power estimator, e.g., [13, 14, 15], an a priori SNR estimator, e.g., [1, 3], and finally one of the spectral weighting rules as already mentioned. Throughout the paper this preprocessing stage is denoted as preliminary noise reduction. As a second stage the ML clean speech amplitude estimate is selected by evaluating the preliminary enhanced signal with the appropriate metric against a GMM which has been trained on clean speech amplitudes. While Mowlace et al. [12] directly deduce a simple spectral subtraction-related weighting rule from the ML clean speech estimate, we utilize the GMM for sole a priori SNR estimation and thus are independent from the employed weighting rule: We are able to use any a priori SNR-driven weighting rule to retrieve the clean speech estimate from the microphone signal. As an important improvement we introduce an iterative scheme where the GMM is repeatedly evaluated in order to retrieve a better ML clean speech amplitude for subsequent a priori SNR estimation.

The remainder of the publication is structured as follows: In Sec. 2 we introduce mathematical notations and briefly revisit the approach of Mowlace et al. as we understand it. Along comes our generalized derivation of the a priori SNR by the GMM and the new iterative scheme. Next in Sec. 3 an evaluation of the a priori SNR estimator in combination with different spectral weighting rules is presented, and finally we draw our conclusions in Sec. 4.
2. Proposed A Priori SNR Estimation

Since our approach is derived from a paper describing a noise reduction we embed our contribution in a speech enhancement context as well. For completeness and better understanding we sketch the whole procedure of Mowlaee et al. in a slightly different manner. Note that the final spectral gain calculation and application can be omitted once we will use it for sole a priori SNR estimation.

2.1. Notations and Assumptions

We assume an additive noise model which can be expressed in the time domain by \( y(n) = s(n) + d(n) \), with \( y(n) \) being the microphone noisy speech signal, \( s(n) \) the clean speech, and \( d(n) \) the additive noise, while \( n \) denotes the discrete-time sample index. Applying the discrete Fourier transform (DFT) we obtain \( Y_\ell(k) = S_\ell(k) + D_\ell(k) \), with frame index \( \ell \) and frequency bin index \( k \) with \( 0 \leq k \leq K-1 \). Additionally, we understand that the noise signal \( d(n) \) can be split up into a stationary (ST) and a non-stationary (NST) component. Thus we define the noise signal as

\[
d(n) = d_{\text{NST}}(n) + d_{\text{ST}}(n),
\]

leading to the frequency domain representation

\[
Y_\ell(k) - S_\ell(k) = D_\ell(k) = \hat{D}_\ell(k)_{\text{ST}} + \hat{D}_\ell(k)_{\text{ST}}.
\]

Assuming statistical independence of speech and noise, as well as of the stationary and the non-stationary noise components, respectively, we obtain the power spectrum representation

\[
|Y_\ell(k)|^2 - \sigma^2_{\text{ST}}(\ell, k) = \sigma^2_S(\ell, k) + \sigma^2_{\text{ST}}(\ell, k)^2 + \sigma^2_{\text{ST}}(\ell, k)^2,
\]

where the variance of the microphone signal has been replaced by its instantaneous estimate \( \hat{\sigma}^2_S(\ell, k) = |Y_\ell(k)|^2 \).

2.2. Model-Based Noise Reduction after Mowlaee

In the following we sketch the model-based speech enhancement approach of Mowlaee et al. [12] as we will partially adopt his algorithm.

A preliminary enhanced speech signal is obtained by employing a conventional noise reduction, here by means of the noise power estimator presented in [15], a priori SNR estimation by the DD approach from [1], and the minimum mean-square error log-spectral amplitude estimator (MMSE-LSA) [2]. Thus we can define the preliminary enhanced speech signal by

\[
\hat{Y}_\ell(k) = Y_\ell(k) \cdot \hat{G}_\ell(k),
\]

where \( \hat{G}_\ell(k) \) represents the MMSE-LSA spectral gains [2], which are depending on the \( a \) priori SNR \( \xi_\ell(k) = \sigma^2_S(\ell, k)^2 \), and the \( a \) posteriori SNR \( \gamma_\ell(k) = \frac{|Y_\ell(k)|^2}{\sigma^2_S(\ell, k)^2} \). Since some entities are not available as such, we need to rely on estimated quantities, whereby \( \xi_\ell(k) \) is provided by the DD a priori SNR estimator [1] and \( \sigma^2_{\text{ST}}(\ell, k)^2 \) is the estimated noise power by [15]. Please note that the utilization of estimated noise power estimate is not perfectly capable of comprising the total noise power and thus a residual non-stationary noise component \( \sigma^2_{\text{ST}}(\ell, k)^2 \) still requires attention. Moreover the estimate of [15] is taken as \( \sigma^2_{\text{ST}}(\ell, k)^2 \) even though it is supposed to cover non-stationary noise as well at least to some extent.

As a next step a binary mask is applied to the pre-enhanced signal in order to identify bins which contain active speech, yielding

\[
\hat{Y}_\ell(k) = \hat{Y}_\ell(k) \cdot B_\ell(k)
\]

with

\[
B_\ell(k) = \begin{cases} 1, & \text{if } \sqrt{\sigma^2_{\text{ST}}(\ell, k)^2} < |\hat{Y}_\ell(k)| \\ 0, & \text{otherwise} \end{cases}
\]

Next we define a Gaussian mixture model (GMM) probability density function (PDF)

\[
p(X) = \sum_{m=1}^{M} c_m \cdot \mathcal{N}(X; \mu_m, \Sigma_m)
\]

with \( M \) distinct modes, each representing a normal distribution \( \mathcal{N}(X; \mu_m, \Sigma_m) \) with mean vector

\[
\mu_m = (\mu_m(0), \mu_m(1), \ldots, \mu_m(K-1))^T
\]

main-diagonal covariance matrix

\[
\Sigma_m = \text{diag} \left( \left( \sigma^2_{\text{ST}}(1), \sigma^2_{\text{ST}}(1), \ldots, \sigma^2_{\text{ST}}(K-1) \right) \right),
\]

and mixture weight \( c_m \) following the constraint \( \sum_{m=1}^{M} c_m = 1 \).

In a training step, the vector \( X \) represents clean speech DFT amplitudes according to \( X = (|S_\ell(0)|, |S_\ell(1)|, \ldots, |S_\ell(K-1)|)^T \). The GMM is trained by performing ten iterations with enforced main-diagonal covariance matrices of the expectation-maximization (EM) algorithm [16] on clean speech amplitudes.

The next step of the baseline approach under consideration is applying the pre-enhanced masked amplitudes (5)

\[
X = (|Y_\ell(0)|, |Y_\ell(1)|, \ldots, |Y_\ell(K-1)|)^T = \hat{X}_\ell
\]

to each of the weighted modes in the GMM (7) and searching for the maximum likelihood (ML) clean speech amplitude estimate

\[
\hat{X}_{\ell}^{\text{ML}} = (|S_\ell(0)|, |S_\ell(1)|, \ldots, |S_\ell(K-1)|)^T = \mu_{m*},
\]

which is the mean vector of the Gaussian mode with index

\[
m^* = \arg \max_m c_m \cdot \mathcal{N}(X; \mu_m, \Sigma_m) = \arg \min_m \sum_{k=0}^{K-1} \left( \frac{|\hat{Y}_\ell(k)|^2 - \mu_m(k)^2}{2\sigma_m^2(k)} - \ln \left( \frac{c_m}{\sqrt{2\pi\sigma_m(k)}} \right) \right).
\]

Now the standard deviation \( \sigma^2_{\text{ST}}(\ell, k) \) of the non-stationary noise power is estimated by a simple spectral power subtraction approach motivated by (3) and as presented in [12], where a Wiener filter-like solution is proposed to obtain the noise power estimate from the microphone signal \( Y_\ell(k) \) by

\[
\hat{\sigma}^2_{\text{ST}}(\ell, k) = |Y_\ell(k)| \cdot G_{\ell}^{\text{NST}}(k)
\]

with

\[
G_{\ell}^{\text{NST}}(k) = \frac{|Y_\ell(k)|^2 - |S_\ell(k)|^2 - \hat{\sigma}^2_{\text{ST}}(\ell, k)^2}{|Y_\ell(k)|^2}
\]

which can be interpreted as (compare to (2)) \( G_{\ell}^{\text{NST}}(k) = \frac{\hat{\sigma}^2_{\text{ST}}(\ell, k)^2}{\sigma^2_{\text{ST}}(\ell, k)^2} \). The final spectral gain is calculated on behalf of the maximum likelihood clean speech amplitude estimate (11), the stationary, and the non-stationary noise power estimates. The estimated frequency-domain clean speech signal \( \hat{S}_\ell(k) \) is then
obtained by
\[ \tilde{S}_\ell(k) = Y_\ell(k) \cdot G_\ell(k) \] (15)
with \( G_\ell(k) = \left\{ \begin{array}{ll} \frac{|\tilde{S}_\ell^{\text{ML}}(k)|}{\sqrt{|\tilde{S}_\ell^{\text{ML}}(k)|^2 + \max\left(\sigma_D^2(\ell,k)^2,\sigma_{\tilde{D}}^2(\ell,k)^2\right)}} & \text{if } \sqrt{|\tilde{S}_\ell^{\text{ML}}(k)|^2 + \max\left(\sigma_D^2(\ell,k)^2,\sigma_{\tilde{D}}^2(\ell,k)^2\right)} < |\tilde{S}_\ell^{\text{ML}}(k)| \\ G_{\text{min}} & \text{otherwise} \end{array} \right. \] (16)

where \( G_{\text{min}} \) denotes a typical gain floor of -15 dB.

2.3. New Model-Based A Priori SNR Estimation

In (16) the model-based clean speech amplitude estimate is used to form a specific gain function. However, considering the existing algorithm pipeline, it is easy to see that one can influence a key factor: the a priori SNR. The a priori SNR is not only important for noise reduction applications but also for speech processing in general. Instead of computing a particular weighting rule directly as done in (16) the sole a priori SNR widens the scope of application in speech enhancement, e.g., for speech presence detection. The block diagram of our proposed architecture is depicted in Fig. 1.

We introduce the new estimate for the a priori SNR based on the ML clean speech amplitude estimate as
\[ \xi_\ell^{\text{ML}}(k) = \frac{|\tilde{S}_\ell^{\text{ML}}(k)|^2}{\sigma_D^2(\ell,k)^2} \] (17)

and also adopt the estimated a posteriori SNR
\[ \gamma_\ell(k) = \frac{|Y_\ell(k)|^2}{\sigma_D^2(\ell,k)^2}. \] (18)

Next, we consider the application of the gain function in (15) and rewrite (16) as
\[ G_\ell(k) = \left\{ \begin{array}{ll} \frac{\xi_\ell^{\text{ML}}(k)}{1+\xi_\ell^{\text{ML}}(k)} & \text{if } \xi_\ell^{\text{ML}}(k) > 1 \\ G_{\text{min}} & \text{otherwise} \end{array} \right. \] (19)

Please note that in the derivation of (19) from (16) we disregard the max operation in the denominator of (16) for the moment and simply use \( \sigma_D^2(\ell,k)^2 \) instead.

Having defined \( \xi_\ell^{\text{ML}}(k) \) and \( \gamma_\ell(k) \) as in (17) and (18), respectively, we are now able to substitute the gain function \( G_\ell(k) \) by any given weighting rule that depends on \( \xi_\ell^{\text{ML}}(k) \) and/or \( \gamma_\ell(k) \), e.g., by the MMSE-LSA [2].

2.4. New Iterative Approach

Our iterative approach is based on the observation that the repetitive application of \( G_\ell(k) \) to the preliminary enhanced signal \( \tilde{Y}_\ell(k) \) yields a more suitable signal w.r.t. a priori SNR estimation. This does not necessarily imply that the obtained intermediate clean speech estimates improve w.r.t. speech quality.

First, we define the iterative clean speech amplitude estimate based on (11) by introducing the iteration index \( i \), yielding
\[ \tilde{S}_\ell^{\text{ML}}(k) = (|\tilde{S}_\ell^{\text{ML}}(0)|,|\tilde{S}_\ell^{\text{ML}}(1)|,\ldots,|\tilde{S}_\ell^{\text{ML}}(K-1)|)^T = \mu_{m^*_i} \]
with Gaussian mode index \( m^*_i \)
\[ \arg \min_m \sum_{k=0}^{K-1} \left[ \frac{|\tilde{S}_{\ell,i-1}(k)| - \mu_m(k)}{2\sigma_m^2(k)} \right]^2 - \ln \left( \frac{\epsilon_m}{\sqrt{2\pi\sigma_m(k)}} \right) \] (21)

for \( i = 1, \ldots, I \), where \( I \) denotes the maximum number of performed iterations. We initialize \( \tilde{S}_{\ell,0}(k) = \tilde{Y}_\ell(k) \) and define an update rule for the estimated frequency-domain clean speech signal as follows
\[ \tilde{S}_{\ell,i}(k) = \tilde{Y}_\ell(k) \cdot G_{\ell,i}(k) \] (22)

with \( G_{\ell,i}(k) \) being the MMSE-LSA spectral weighting rule [2] depending on the iterative a priori SNR
\[ \xi_{\ell,i}(k) = \frac{|\tilde{S}_{\ell,i}(k)|^2}{\sigma_D^2(\ell,k)^2} \] (23)

and the fixed a posteriori SNR,
\[ \gamma_{\ell,i}(k) = \gamma_{\ell}(k). \] (24)

Since the iterative weighting rule is applied to the same signal \( \tilde{Y}_\ell(k) \) each time (22), we define a convergence criterion being that the same mode from the GMM is consecutively chosen: \( m^*_i = m^*_{i-1} \) or a maximum amount of iterations reached. This implies that the gain function \( G_{\ell,i}(k) \) does not change once the same amplitude mean has been selected by (21).

3. Experimental Results

3.1. Experimental Setup

Our framing throughout the whole training and testing process is the following: We operate at a sample rate of 8 kHz with a frame size of 256 samples and a frame shift of 64 samples. The employed window for analysis and synthesis is a periodic square root Hann window. Since the core of the approach is a GMM trained on clean speech amplitudes we need a significant amount of speech data. We chose the GRID corpus [17], downsampled to 8 kHz and divided it into a test set containing speakers 2, 4, 6, 7, and a training set comprising the remainder of the speakers. Thus we have two male and two female speakers in our test set and our test results are averaged over 100 sentences per speaker, i.e., 400 total. Since the files are alphabetically sorted and follow a certain grammar we designate every 10th file of every speaker to be in the test set in order to achieve some variance. The training of a speaker-independent GMM is based on the remaining 30 speakers and 50 files per speaker amounting to a total of 1500 files. Next we extract the clean

---

Figure 1: Block diagram of the proposed iterative model-based a priori SNR estimator.
Figure 2: Comparing the new a priori SNR estimator (MB) to the DD approach in combination with three different weighting rules by means of segmental NA and PESQ MOS of the speech component.

Speech amplitudes of the training pool by employing a simple 3-state voice activity detection as proposed in [18], please note that we do not perform any normalization on the source files. Subsequently the EM algorithm is employed with forced main-diagonal covariance matrices, a maximum of 100 iterations, and the desired order of $M = 512$. Hence the computed GMM contains 512 modes representing clean speech amplitudes, each being a vector with dimension corresponding to the DFT size $K = 256$.

For the simulation we make use of the ETSI background noise database [19], downsampled to 8 kHz, and employ the left channel noise files recorded in a full-size car driving at 80 km/h, in a call center, and at crossroads. Subsequently we measure the active speech level [20], adjust the level of the noise accordingly to obtain the desired SNR ranging from $-5$ dB to $15$ dB in $5$ dB steps, and superimpose speech and noise.

Evaluation is performed by employing a standard noise reduction using minimum statistics (MS) [14] as noise power estimator adjusted with a fixed overestimation factor for each weighting rule such that a comparable PESQ score is met at $-5$ dB SNR. We apply the new and the DD a priori SNR estimator, and three different weighting rules, namely MMSE-LSA [2], the Wiener filter (WF) [7], and Loizou’s super-Gaussian joint maximum a posteriori (SG-jMAP) estimator [6]. The reference algorithms are operated with the optimal parameters1 as proposed in [21]. Our setup of the preliminary noise reduction of the iterative a priori SNR estimator as described in Sec. 2.2 is as follows: The noise power is estimated by employing the (MS) algorithm, the smoothing factor for the DD a priori SNR estimation is set to $\beta_{\text{DD}} = 0.985$, and the minimum for the a priori SNR is $\xi_{\text{min}} = -15$ dB as is the gain floor $G_{\text{min}}$ for the MMSE-LSA weighting rule.

As we assume a linear model in Sec. 2.1 the components of the noisy signal can be separately processed by applying the respective gain function $G_{l}(k)$ to the individual signals $s(n)$ and $d(n)$ in the frequency domain. This so-called white-box approach [22] yields the filtered clean speech $\hat{s}(n)$ and the filtered noise $\tilde{d}(n)$, respectively.

For measures of quality we are taking into account the segmental noise attenuation (NA) computed as [23]

$$\text{NA}_{\text{seg}} = 10 \log_{10} \left[ \frac{1}{L} \sum_{\ell \in L} \text{NA}(\ell) \right],$$

with

$$\text{NA}(\ell) = \frac{\sum_{\nu=0}^{N-1} d_{\nu}^2(\nu + \ell N)}{\sum_{\nu=0}^{N-1} d_{\nu}^2(\nu + \ell N + \delta)},$$

where $\ell$ defines a segment of length $N = 256$, $\delta$ is compensating the sample delay of the filtered signals, and $\frac{1}{L}$ is a normalization factor since $|L|$ represents the cardinality of set $L$, containing all frames. Besides, we employ the PESQ MOS-LQO score [24] to evaluate the quality of the filtered clean speech component $\hat{s}(n)$, not of the total enhanced speech $\hat{s}(n)$.

3.2. Experimental Evaluation

Fig. 2 shows the results of our simulations averaged for the three different noise environments. We plot the PESQ MOS scores over the segmental NA. Every marker denotes a different SNR condition, starting with $15$ dB at the top to $-5$ dB at the bottom in steps of $5$ dB. The dashed lines represent the reference a priori SNR estimator (DD) while the solid ones depict the proposed approach. Each a priori SNR estimator is evaluated in combination with the different weighting rules, MMSE-LSA, SG-jMAP, and WF, which are distinguished by the different markers. In general, the further to the top and to the right a curve is located in the plot, the lower the speech distortion and the amount of perceived residual noise in the enhanced speech signal. As we tuned all the different setups to achieve a similar PESQ score at $-5$ dB to facilitate a fair comparison, the curves only differ in the achieved level of NA while the quality of the speech component remains similar in terms of PESQ scores.

The figure shows that the proposed a priori SNR estimator allows to achieve a much higher segmental NA while maintaining comparable speech distortion (speech component PESQ scores). Especially in low SNR conditions the new approach outperforms the DD estimation by up to more than $4$ dB segmental NA. Interestingly the model-based estimation leads to a far stronger decrease of segmental NA in higher SNR conditions but still exceeds the reference algorithms in all three cases. It is also observable that the performance of the different weighting rules seems to be less varying with the new approach since the curves show very similar behavior and exhibit less variance in the results when compared to the reference approach. In formal listening tests have also shown that when musical tones are present in an enhanced file processed with the DD estimator they vanish when alternative a priori SNR estimation is applied independent of the utilized weighting rule.

4. Conclusion

In this contribution we presented a new a priori SNR estimator based on a GMM providing a ML clean speech amplitude estimate. We have evaluated our approach against Ephraim/Mahla’s DD estimator as reference, three different weighting rules, and three different noise environments on the GRID corpus. Our experiments have shown that the proposed estimator outperforms the DD approach, not only, but especially in low SNR conditions in terms of segmental NA while maintaining a comparable quality of the speech component.
5. References


