Speech enhancement based on novel two-step \textit{a priori} SNR estimators

Md. Jahangir Alam$^1$, Douglas O’Shaughnessy $^1$, Sid-Ahmed Selouani $^2$

$^1$INRS-Energie-Matériaux-Télécommunications, Université du Québec, Montréal, Canada $^2$Université de Moncton, Campus de Shippagan, NB, Canada

alam@emt.inrs.ca, dougo@emt.inrs.ca, selouani@umcs.ca

Abstract

A widely used method to determine the \textit{a priori} SNR from noisy speech is the decision directed (DD) approach, but the \textit{a priori} SNR follows the \textit{a posteriori} SNR with a delay of one frame in speech frames. As a consequence, the performance of the noise reduction system degrades. In order to overcome this artifact, we propose three computationally simple and efficient two-step methods based on the minimum mean square error (MMSE), the maximum \textit{a posteriori} (MAP) and the joint MAP criteria for the estimation of the \textit{a priori} SNR for speech enhancement. The proposed methods avoid the delay problem while keeping the advantages of the DD method. The performance of the proposed \textit{a priori} SNR estimation methods are evaluated and compared with the conventional DD method by extensive objective quality measures and yield better results than the DD approach-based speech enhancement system.

Index Terms: decision-directed, \textit{a priori} SNR, speech enhancement, joint MAP, MMSE

1. Introduction

The reduction of acoustical background noise using a single microphone is an important subject to improve the quality of speech communication systems in the context of digital hearing aids, speech recognition, hands-free telephony, or teleconferencing. Although single-microphone speech enhancement has been a research topic for decades, the estimation of a clean speech signal from its noisy observation remains a challenging task, especially due to the wide variety of environmental noises.

A unified view of the main single-microphone speech enhancement system is presented in [1], where the enhancement process depends on the estimation of the spectral gain, which is a function of the \textit{a priori} SNR and/or \textit{a posteriori} SNR. A well known method for the estimation of the \textit{a priori} SNR is the DD approach [2], as it leads to reduced musical noise, which comes from the residual noise composed of sinusoidal components randomly distributed over successive frames and annoying to the listener. However, this method has a serious drawback that the estimated \textit{a priori} SNR follows the shape of the \textit{a posteriori} SNR with a simple delay of one short time frame [3]. This bias is due to the use of the speech spectrum estimated in the previous frame to compute the current \textit{a priori} SNR. Therefore the spectral gain of the DD approach-based techniques matches the previous frame rather than the current one and degrades the speech enhancement performance. To overcome this problem while maintaining the advantages of the DD approach we propose here three novel two-step \textit{a priori} SNR estimation methods for the enhancement of the corrupted speech signals. The proposed methods yield improved performance over the conventional DD approach for a number of objective quality measures.

The remaining part of this paper is organized as follows: section 2 gives an overview of the baseline speech enhancement system. Estimation of the \textit{a priori} SNR using the DD approach and the proposed approaches are described in section 3. Section 4 provides description on experimental results and the conclusion is made in section 5.

2. Baseline speech enhancement method

Let the distorted signal be expressed as
\[ y(n) = x(n) + d(n), \]
where \( x(n) \) is the clean signal and \( d(n) \) is the additive random noise signal, uncorrelated with the original signal. If at the \( m \)-th frame and \( k \)-th frequency bin, \( y(m,k), X(m,k) \) and \( D(m,k) \) represent the spectral component of \( y(n), x(n) \) and \( d(n) \), respectively, then the distorted signal in the transformed domain is
\[ Y(m,k) = X(m,k) + D(m,k). \]
An estimate \( \hat{X}(m,k) \) of \( X(m,k) \) is given by
\[ \hat{X}(m,k) = H(m,k)Y(m,k), \]
where \( H(m,k) \) is the noise suppression gain (denoising filter), which is a function of \textit{a priori} SNR and/or \textit{a posteriori} SNR, given by
\[ H(m,k) = \left( \frac{\xi(m,k)}{1 + \xi(m,k)} \right) \Gamma_{\nu}(m,k). \]
The first parameter of the noise suppression rule is the \textit{a posteriori} SNR given by
\[ \gamma(m,k) = \frac{\left\| Y(m,k) \right\|^2}{\Gamma_{\nu}(m,k)}, \]
where \( \Gamma_{\nu}(m,k) = E \{ |D(m,k)|^2 \} \) is the noise power spectrum estimated during speech pauses. For the estimation of the noise power spectrum we have used weighted noise estimation method described in [4]. The \textit{a priori} SNR, which is the second parameter of the noise suppression rule, is expressed as
\[ \xi(m,k) = \frac{\Gamma_{\nu}(m,k)}{\Gamma_{\nu}(m,k)}, \]
where \( \Gamma_{\nu}(m,k) = E \left\{ |X(m,k)|^2 \right\} \).

The instantaneous SNR can be defined as
\[ \phi(m,k) = \frac{\left\| Y(m,k) \right\|^2}{\Gamma_{\nu}(m,k)} - 1. \]
The temporal-domain denoised speech is obtained with the following relation
\[ \hat{x}(n) = \text{IFFT}\left\{ \hat{X}(m,k) e^{j\phi(m,k)} \right\}. \]
3. Estimation of a priori SNR

An important parameter of numerous speech enhancement techniques is the a priori SNR. Although most speech enhancement techniques improve speech quality, they often suffer from an annoying artifact called musical noise caused by randomly spaced spectral peaks that come and go in each frame, and at random frequencies. The randomly spaced peaks are due to the inaccurate estimate of the a priori SNR. Therefore, an accurate estimate of the a priori SNR is critical for eliminating musical noise.

3.1 Decision-directed approach

A widely used method to determine the a priori SNR from distorted speech is the decision-directed (DD) approach. In [2] the DD approach was defined as a linear combination of (7) and (8). With a weighting parameter \( \alpha \) that is constrained to be \( 0 < \alpha < 1 \), the linear combination results in

\[
\hat{x}(m,k) = E\left[ \alpha X(m,k) + (1-\alpha) \hat{p}(m,k) \right].
\]

However, as this expression is hard to implement in practice, approximations were made. This led to the following equation

\[
\tilde{\alpha}(m,k) = \max \left\{ \frac{H_{JMAP}(m-1,k)Y(m-1,k)}{\Gamma_d(m,k)} + \ldots, \right. \]

\[
(1-\alpha)P'[\hat{\theta}(m,k),x_{\text{min}}] \}
\]

(10)

where \( P'[x]=x \) if \( x \geq 0 \) and \( P'[x]=0 \) otherwise. In this paper we have chosen \( \alpha = 0.98 \) and \( x_{\text{min}} = 0.0032 \) (i.e., \(-25 \) dB) by the simulations and informal listening tests. The multiplicative gain function for this approach becomes

\[
H_{JMAP}(m,k) = \frac{\delta_{JMAP}(m,k)}{\Gamma_d(m,k)}. \]

Then the enhanced speech spectrum is obtained using (3). This method results in a significant elimination of musical noise.

3.2 The proposed a priori SNR estimators

In the well-known decision-directed approach, the a priori SNR depends on the speech spectrum estimation in the previous frame, which results in degradation of the speech enhancement performance. In order to remove this problem while keeping its benefits we have proposed three two-step a priori SNR estimation approaches based on the Joint Maximum A posteriori (JMAP), the Maximum A posteriori (MAP), the Minimum Mean Square Error (MMSE) criteria.

3.2.1 Joint MAP-based a priori SNR estimator

The joint MAP estimator gives the speech spectral amplitude \( \hat{X}(m,k) \) that maximizes \( p(X(m,k),\theta(m,k)|Y(m,k)) \) as follows [5]; for simplicity the frame index \( m \) and frequency bin index \( k \) are omitted:

\[
\hat{X} = \arg \max_x p(X,\theta|Y)
\]

\[
= \arg \max_x p(Y|X,\theta)p(X,\theta) / p(Y)
\]

(12)

where

\[
p(Y|X,\theta) = \frac{1}{\pi \Gamma_d} e^{-\frac{2Y^2}{\Gamma_d}}.
\]

(13)

It is assumed that \( p(X,\theta|Y) \) is Gaussian and that \( p(X) \) and \( p(\theta) \) are statistically independent. Since \( p(Y) \) is independent of \( X \) we need to maximize only \( p(Y|X,\theta)p(X,\theta) \). Therefore

\[
p(X,\theta|Y) = \frac{X}{\pi \Gamma_d} e^{-\frac{2Y^2}{\Gamma_d}}.
\]

(14)

Differentiating (15) with respect to the phase \( \theta \) yields

\[
\frac{\partial}{\partial \theta} \ln p(X,\theta|Y) = \frac{Y - Xe^{i\theta}}{\Gamma_d} - \frac{X^2}{\Gamma_d} + \ln X + \text{constant}.
\]

(15)

Insertion of (3), (5) and (6) into (18) gives

\[
priorit\text{MMSE} = \hat{X}^2 = \frac{0.5\xi}{1 + \xi} (1 + 2HY).
\]

(19)

3.2.2 MAP-based a priori SNR estimator

The pdf of the noisy spectrum \( Y \) conditioned on the speech amplitude \( X \) and phase \( \theta \) can be written as a joint Gaussian given by (13). A Ricci pdf is obtained for the density of the noisy spectrum \( Y \) and the decision-directed approach was defined as a linear combination of (7) and (8). With a weighting parameter \( \alpha \) that is constrained to be \( 0 < \alpha < 1 \), the linear combination results in

\[
\hat{X} = \arg \max_{x} p(X|A)
\]

\[
= \arg \max_{x} \frac{p(A|X)p(X)}{p(A)}
\]

(21)

We need to maximize \( p(A|X)p(X) \) as \( p(A) \) is independent of \( X \). Therefore

\[
p(X|A) = \left( \frac{2A e^{-\frac{X^2+\theta^2}{\Gamma_d}}} {\pi \Gamma_d} \right) I_0\left( \frac{2AX}{\Gamma_d} \right) \left( \frac{X^2}{\Gamma_d} \right).
\]

(22)

A closed form solution can be found if the modified Bessel function \( I_0 \) is considered asymptotically with
\[
I_0(x) = \frac{1}{\sqrt{2\pi x}} e^{-x}. \quad (23)
\]

Using (23) in (22) we get
\[
p(X|A) = \frac{2\sqrt{AX}}{\sqrt{2\pi a}} \left( \frac{A - X}{\sqrt{2\pi a}} \right) e^{-\frac{(A - X)^2}{2a}}
\]
\[
\ln p(X|A) = -\frac{(A - X)^2}{\Gamma_d} - \frac{X^2}{\Gamma_v} + \ln A + \ln X + \text{constant} \quad (24)
\]

Differentiating (24) with respect to the speech amplitude \(X\) and setting to zero yields
\[
1 - \frac{\hat{X}^2}{\Gamma_v} + \frac{A \hat{X}}{\Gamma_d} = 0
\]
\[
(25)
\]

Using (3), (5), and (6) in (25) we get
\[
priori_{\text{MAP}} = \frac{\hat{X}^2}{\Gamma_d} = \frac{\xi}{1 + \xi} (1 + Hy).
\]
\[
(26)
\]

### 3.2.2 MMSE based \textit{a priori} SNR estimator

The MMSE estimator for the power spectral density \(X^2\) can be given by the conditional expectation as follows:
\[
\hat{X}^2 = E\{X^2|Y\} = \frac{\int_{-\infty}^{\infty} X^2 P(Y|X)P(X) dX}{\int_{-\infty}^{\infty} P(Y|X)P(X) dX} \quad (27)
\]

Now following the same procedure as described in [6] we get from (27)
\[
\hat{X}^2 = \left( \frac{\Gamma_v}{\Gamma_v + \Gamma_d} \right)^2 A^2 + \frac{\Gamma_d}{\Gamma_v + \Gamma_d}, \quad (28)
\]

where \(A = |Y|\). Using (5), and (6) in (28) we get
\[
priori_{\text{MMSE}} = \frac{\hat{X}^2}{\Gamma_d} = \frac{\xi}{1 + \xi} \left( 1 + \frac{\xi}{1 + \xi} Y \right). \quad (29)
\]

In (19), (26), and (28) the estimation of \(\xi\) is given by (10) i.e., the first step of the proposed methods is the well known DD approach whereas the second step, equations (19), (26) and (28) are used to refine the estimated \textit{a priori} SNR of the DD approach.

Figure 1 represents the variations of the \textit{a priori} SNR estimated using the proposed approaches and the DD approach with the \textit{a posteriori} SNR, noisy signal includes WGN with SNR = 10 dB. It is observed that the proposed two-step \textit{a priori} SNR estimation approaches solve the delay problem while maintaining the advantages of the DD approach.

### 4. Performance Evaluation and Discussion

In order to evaluate the performance of the proposed \textit{a priori} SNR approaches, we conducted extensive objective quality tests [8] under various noisy environments. The frame sizes were chosen to be 256 samples (32 msec) long with 40% overlap; a sampling frequency of 8 kHz and a hamming window were applied. To evaluate and compare the performance of the \textit{a priori} SNR estimators, we carried out simulations with the \textit{TEST A} database of Aurora [9]. Speech signals were degraded with four types of noise at global SNR levels of 0 dB, 5 dB, and 10 dB. The noises were N1 (Subway noise), N2 (Babble Noise), N3 (Car Noise), and WGN (White Gaussian Noise).

Table 1 shows the average segmental SNR for the enhanced speech signals in various types of noisy environments. It is observed that the proposed approaches give better segmental SNR (SegSNR) than that of the conventional DD \textit{a priori} SNR-based speech enhancement technique under all tested noisy environments. Table 2 shows the Log Spectral Distance (LSD) measure for the enhanced speech signals in various types of noise corruption. The proposed two-step \textit{a priori} SNR estimators exhibit lower values of LSD for almost all noisy environments compared to those obtained by the DD approach based enhancement technique.

Figure 2 represents the spectrograms of the clean speech signal, noisy signal and enhanced speech signals obtained with different \textit{a priori} SNR estimation approaches. The speech spectrograms provide more accurate information about the residual noise and speech distortion than the corresponding time domain waveforms. We compared the spectrograms for each of the methods and confirmed a reduction of the residual noise and speech distortion. Speech spectrograms presented in Figure 2 use a Hamming window of 256 samples with 50% overlap and the noisy signals include N3 (Car Noise) with SNR = 0 dB. It is seen that the musical noise is removed for the most part in figures 2(d), 2(e) and 2(f).

In this paper we have used only objective tests for performance evaluation of the proposed approaches. In the future we will evaluate the subjective performances of the proposed approaches.
5. Conclusion

An improved speech enhancement scheme is proposed based on three low complexity and efficient two-step \textit{a priori} SNR estimation methods. Although the \textit{a priori} SNR estimated in the first step provides interesting properties, it suffers from a delay of one frame which degrades speech enhancement performance. This delay problem is removed by the second step of the proposed approaches and results in significant improvement in speech enhancement performance in terms of objective quality measures.

6. Reference


