LETTER

A Novel Approach to a Robust a Priori SNR Estimator in Speech Enhancement

Yun-Sik PARK†, Nonmember and Joon-Hyuk CHANG†(a), Member

SUMMARY This paper presents a novel approach to single channel speech enhancement in noisy environments. Widely adopted noise reduction techniques based on the spectral subtraction are generally expressed as a spectral gain depending on the signal-to-noise ratio (SNR) [1]–[4]. As the estimation method of the SNR, the well-known decision-directed (DD) estimator of Ephraim and Malah efficiently is known to reduces musical noise in noise frames, but the a priori SNR, which is a crucial parameter of the spectral gain, follows the a posteriori SNR with a delay of one frame in speech frames [5]. Therefore, the noise suppression gain using the delayed a priori SNR, which is estimated by the DD algorithm matches the previous frame rather than the current one, so after noise suppression, this degrades the performance of a noise reduction during abrupt transient parts. To overcome this artifact, we propose a computationally simple but effective speech enhancement technique based on the sigmoid type function to adaptively determine the weighting factor of the DD algorithm. Actually, the proposed approach avoids the delay problem of the a priori SNR while maintaining the advantage of the DD algorithm. The performance of the proposed enhancement algorithm is evaluated by the objective and subjective test under various environments and yields better results compared with the conventional DD scheme based approach.

key words: a priori SNR, decision-directed, speech enhancement, sigmoid type

1. Introduction

Since the demand for speech communication systems in mobile environments is increasing, effective speech enhancement is seen as an indispensable speech processing tool [1]–[4]. Relevant speech enhancement techniques can be expressed as a spectral noise suppression gain, based on the signal-to-noise ratio (SNR) [1]–[4]. However, the application of the spectral gain results in the artifact which is known as the “musical noise” during noise frames. The musical noise comes from the residual noise composed of sinusoidal components randomly distributed over successive frames and sounds disturbing to the listener. A method for significant elimination of the musical noise is the decision-directed (DD) estimation approach originally proposed by Ephraim and Malah [3]. Cappé analyzed the performance of the DD estimation and demonstrated that the musical noise is strongly reduced by the a priori SNR corresponding to a highly smoothed version of the a posteriori SNR in noise frames, while the a priori SNR follows the a posteriori SNR with a delay of one frame in speech frames. Since the noise suppression gain utilizing a minimum mean-square error (MMSE) mainly depends on the a priori SNR in the case of low SNR, the noise suppression gain using the delayed a priori SNR, which is estimated by the conventional DD scheme matches the previous frame rather than the current version. For this reason, the distortion of the noise suppression gain degrades the quality of the enhanced speech signal, especially in abrupt transient parts [5].

In this letter, we present a novel and computationally simple a priori estimation technique in noisy speech enhancement. In contrast to the DD estimator, having the fixed weighting factor, the proposed DD estimator utilizes the sigmoid type function for the adaptive weighting factor, where the output value of sigmoid type function is determined by the deviation of the a posteriori SNR. The performance of the proposed algorithm is evaluated by the Perceptual Evaluation of Speech Quality scores (PESQ, ITU-T P.862), the mean opinion score (MOS) and the speech spectrograms under various noise conditions and shown better results than those of the DD estimator using the fixed weighting factor [7]–[9].

2. Spectral Gain Function for Speech Enhancement

Let \( x(t) \) and \( d(t) \) denote the speech and the noise signal, respectively. The noisy speech signal \( y(t) \) is given by

\[
y(t) = x(t) + d(t).
\]

Let \( X(k) \), \( D(k) \) and \( Y(k) \) denote the \( k \)th spectral component of the speech \( x(t) \), the noise \( d(t) \) and the noisy speech \( y(t) \), respectively. The spectral noise suppression gain depends on two parameters, a posteriori SNR and a priori SNR, defined by \( \gamma(k) \) and \( \xi(k) \), respectively [3].

\[
\gamma(k) \equiv \frac{|Y(k)|^2}{E[|D(k)|^2]}
\]

(2)

\[
\xi(k) \equiv \frac{E[|X(k)|^2]}{E[|D(k)|^2]}
\]

(3)

where \( E[\cdot] \) denotes the expectation operator. Under the assumption that the spectral components of the noisy speech signal are assumed statistically independent, the MMSE amplitude estimator can be derived from \( Y(k) \) only. As a result, the MMSE estimator \( \hat{X}(k) \) of \( |X(k)| \) is obtained as follows:

\[
|\hat{X}(k)| = E[|X(k)| \mid y(t), 0 \leq t \leq T] \quad (4)
\]

\[
= E[|X(k)| \mid Y(0), Y(1), \cdots]
\]

Manuscript received March 6, 2007.
†The authors are with the Department of Electronic Engineering, Inha University, Incheon, Korea.
a)E-mail: changjh@inha.ac.kr
DOI: 10.1093/ietcom/e90-b.8.2182
It is useful to consider the estimated speech spectrum $\hat{X}(k)$ as being achieved from $Y(k)$ by a multiplicative gain function $G(\hat{x}(k), \gamma(k))$, as given below:

$$\hat{X}(k) = G(\hat{x}(k), \gamma(k))Y(k).$$

(5)

Noise suppression gain $G(\cdot, \cdot)$ which is achieved by parameters defined by (2) and (3) can be computed as follow [3]:

$$G(\xi(k), \gamma(k)) = \frac{\sqrt{\pi}v(k)}{2\gamma(k)} \exp\left(-\frac{v(k)}{2}\right) \left[1 + v(k)I_0\left(\frac{v(k)}{2}\right) + v(k)I_1\left(\frac{v(k)}{2}\right)\right]$$

(6)

where $v(k)$ is defined by

$$v(k) = \frac{\xi(k)}{1 + \xi(k)}\gamma(k)$$

(7)

where $I_0$ and $I_1$ are the modified Bessel functions of zero and first order, respectively [3]. Also, $G(\cdot, \cdot)$ shown in (6) represents the noise suppression gain function, which depends only on the a priori and the a posteriori SNR, $\xi(k)$ and $\gamma(k)$, respectively. The mean square value of the noise $D(k)$ is obtained through a noise power spectrum estimate, which is updated using a voice activity detector (VAD) during the absence of speech [4, 6]. The a posteriori SNR $\gamma(k)$ defined by (2) is estimated by using the observed signal $Y(k)$ whereas the a priori SNR $\xi(k)$ is estimated by using the speech signal estimate $\hat{X}(k)$, refined by the noise suppression rule.

In order to estimate the a priori SNR, Ephraim and Malah proposed the decision-directed (DD) approach, as given below:

$$\hat{\xi}(k, n) = \alpha \frac{[\hat{X}(k, n-1)]^2}{E[D(k)^2]} + (1 - \alpha)P[\gamma(k, n) - 1], \quad 0 \leq \alpha < 1$$

(8)

where $P[x] = x$ if $x \geq 0$, and $P[x] = 0$ otherwise. The operator $P[\cdot]$ is used to ensure the positivity of $\gamma(k, n) - 1$. Also, $\hat{X}(k, n-1)$ represents the estimated speech spectrum of the $k$th signal spectral component in the $(n-1)$th frame, and $\alpha$ is the weighting factor. Recently, Cappé presented an extensive analysis of the DD approach [5]. In [5], it is demonstrated that the a priori SNR is the dominant parameter in the gain function of (6), and the a priori SNR has a significantly reduced variance due to a highly smoothed version of the a posteriori SNR in noise frames. Thus, the variance of the gain depending on the a priori SNR also becomes smoother in noise frames. As a consequence, it is discovered that the smoothness of the spectral gain strongly reduces the musical noise artifacts [5]. We can see that $\hat{\xi}(k, n)$ is obtained from (8) by applying the weighting factor $\alpha$ to an estimate of the instantaneous SNR at the previous frame. Since the weighting factor $\alpha$ is generally chosen very close to 1, the speech spectra estimated in the previous frame are used to estimate the current a priori SNR and the a priori SNR follows the a posteriori SNR with a delay of one frame when the a posteriori SNR $\gamma(k)$ exhibits an abrupt increase. This delay is likely to produce undesired gain distortion and thus generate audible distortion during abrupt transient periods. As a result, when the weighting factor $\alpha$ increases, the musical noise is significantly reduced during noise frames, but the speech signal could be distorted during the speech onset periods.

3. Proposed Adaptive Weighting Factor of the Sigmoid Type Function

In the previous section, we addressed the problem of the tradeoff between musical noise reduction and the speech distortion in transient parts. To cope with this situation [10, 11], we propose the adaptive weighting factor incorporating the sigmoid type function. Specifically, the adaptive value based on the sigmoid type function is applied to the weighting factor of the conventional DD scheme according to the transient of the a posteriori SNR, as given below:

$$\hat{\alpha}(k) = \frac{k \exp\left(-\beta(s(k) - s_0)\right) + \sigma}{1 + \exp\left(-\beta(s(k) - s_0)\right)}$$

(9)

where

$$s(k) = \log(1/|\Delta y(k)|)$$

(10)

Through the speech enhancement experiments, $\hat{\alpha}(k)$ is obtained by using the slope parameter $\beta = 0.4$, the offset $s_0 = -9$, the constant $k = -1.5$ and $\sigma = 0.99$, respectively. As a result, the proposed estimator for the a priori SNR $\xi_p(k, n)$ of $\hat{\xi}(k, n)$ is deduced from (8) and is given as follows:

$$\xi_p(k, n) = \hat{\alpha}(k, n) \frac{[\hat{X}(k, n-1)]^2}{E[D(k)^2]} + (1 - \hat{\alpha}(k, n))P[\gamma(k, n) - 1], \quad 0 \leq \hat{\alpha} < 1.$$
Fig. 1  The sigmoid type function applied to the weighting factor \( \alpha \).

Fig. 2  SNR in short-time frames. A posteriori SNR (solid line), a priori SNR of the DD algorithm (dashed line), a priori SNR of the proposed algorithm (bold line).

Waveform during transient periods. On the otherhand, decreasing \(|\Delta y|\) results in increasing \(\log(1/|\Delta y(k)|)\), and thus the high value of the weighting factor is applied to the previous frame to maintain the advantage of the DD approach, like the highly reduced musical noise. Figure 2 shows the behaviors of the DD algorithm and the proposed sigmoid type function algorithm for estimating the a priori SNR of the noisy speech signal in conjunction with the original input speech samples. The presented scheme efficiently avoids the delay generated by the DD algorithm, and the shape of the proposed a priori SNR estimator resembles the a posteriori SNR during speech onset periods. This implies that the proposed a priori SNR may be more accurate than the conventional a priori SNR and could improve the performance of the noise suppression gain for speech enhancement.

4. Experimental Results

In order to evaluate the performance of the proposed the a priori SNR estimation technique, we conducted extensive subjective quality test experiments under the various noise conditions. Actually, Tables 1, 2 and Fig. 3 show the PESQ scores, the MOS and the speech spectrograms, respectively.

Table 1  The PESQ scores of the conventional DD and proposed DD algorithm under various noise environments.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Method</th>
<th>SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>White</td>
<td>DD</td>
<td>1.915</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>1.977</td>
</tr>
<tr>
<td>Babble</td>
<td>DD</td>
<td>2.136</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>2.167</td>
</tr>
<tr>
<td>Vehicle</td>
<td>DD</td>
<td>3.437</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>3.532</td>
</tr>
</tbody>
</table>

Table 2  The MOS of the conventional DD and proposed DD algorithm under various noise environments.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Method</th>
<th>SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>White</td>
<td>DD</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>1.95</td>
</tr>
<tr>
<td>Babble</td>
<td>DD</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>2.06</td>
</tr>
<tr>
<td>Vehicle</td>
<td>DD</td>
<td>4.17</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>4.19</td>
</tr>
</tbody>
</table>

Fig. 3  Speech spectrograms. (a) Original clean speech. (b) Enhanced speech with the DD algorithm. (c) Enhanced speech with the proposed algorithm.
male speaker and degraded by the various SNR conditions (5, 10, 15 dB). Opinion scores were evaluated by a group of the listeners and then averaged to yield the final MOS results. From the MOS results, it is evident that all noisy conditions, the proposed method yielded higher scores than the conventional DD algorithm. Figure 3 represents the speech spectrograms of enhanced speech signals. The speech spectrograms presented in Fig. 3 use a Hanning window of 256 samples with an overlap of 128 samples and the noisy signals include white noise (SNR=10 dB). Figures (b) and (c) show the spectrograms obtained with the DD algorithm and the proposed algorithm, respectively. In the proposed algorithm, the speech distortion of the enhanced speech is further reduced compared to the DD algorithm during the period of speech.

5. Conclusion

We have presented a novel speech enhancement algorithm in which the sigmoid type function scheme is adopted for the estimation of the a priori SNR. The DD approach provides a significant musical noise reduction but suffers from audible distortions introduced by a delay of one frame. In the proposed algorithm, the weighting factor is applied according to the transient of the a posteriori SNR. This method solves the delay problem about the a priori SNR while maintaining the benefits of a musical noise reduction. Experimental results have shown that the proposed algorithm yields better PESQ scores, higher the MOS, and lower speech distortion compared with the conventional DD algorithm.

Acknowledgement

This work was supported in part by MIC & IITA through IT Leading R&D Support Project.

References