Speech Reinforcement Based on Soft Decision under Far-End Noise Environments

Jae-Hun CHOI†, Woo-Sang PARK†, Nonmembers, and Joon-Hyuk CHANG†(a), Member

1. Introduction

In conversations over cellular phones under various noisy background environments, there is a loss of clarity and intelligibility in the far-end speech signal. Recently, several algorithms have been proposed to cure this and they have demonstrated impressive performance, such as a speech enhancement rule [1]–[6]. However, since these speech enhancement techniques try to reduce the effect of the ambient noise from the noisy signal for the far-end listener, the intelligibility of the far-end speech for the near-end listener cannot be directly controlled. Recently, efficient techniques for adjusting the frequency components so that they produce the same SNR for each frequency band have been proposed [7], [8]. However, these methods often amplify background noise components in the far-end signal.

In this letter, we propose an effective speech reinforcement algorithm that amplifies the frequency component of the far-end speech in both near and far-end noisy environments. We employ the soft decision scheme to obtain an estimated clean speech spectrum at the far-end that the target to reinforce. We first derive a speech absence probability (SAP) defined for each frequency component and use it for estimation of clean speech spectra at the far-end. As a result, as shown in Fig. 1, the estimated clean speech spectra at the far-end are then amplified to recover the SNR for each frequency band. Indeed, we devise a robust method to revise the reinforcement spectral gain by a given SAP. The performance of the proposed algorithm is evaluated by a subjective testing, thereby demonstrating it to be better than that of conventional schemes [7].

2. Review of Frequency Dependent SNR Recovery

Recently, Sauert and Vary proposed an effective algorithm for enhancement of intelligibility of far-end speech. Their method, referred to as the SNR recovery scheme amplifies the spectral components while preserving the same SNR for all bands [7]. As we adopt this technique as a baseline, we first review it in brief. Let \( \hat{y}(t, k) \), \( s(t, k) \) and \( n(t, k) \) denote far-end noisy speech signal, clean speech signal, and ambient noise signal, respectively. The discrete Fourier transform (DFT) gives the time-frequency domain representation \( Y(t, k) \) and \( N(t, k) \), where \( t \) and \( k \) are the time and frequency indexes, respectively. Amplifying a far-end speech DFT coefficient with \( G(t, k) \), which is applied to a frequency band, results in [7]:

\[
\hat{Y}(t, k) = G(t, k)Y(t, k).
\]

In [7], it was shown that the SNR of \( \Phi_{\hat{y}y}(t, k) \) and \( \Phi_{NN}(t, k) \) is greater than or equal to a specific target SNR such that [7]:

\[
\frac{\Phi_{\hat{y}y}(t, k)}{\Phi_{NN}(t, k)} \geq \xi
\]

where \( \xi \) (15 dB) is a target SNR. Also, \( \Phi_{YY}(t, k) \), \( \Phi_{\hat{y}y}(t, k) \) and \( \Phi_{NN}(t, k) \) denote short-term power spectral density of far-end speech, amplified far-end speech, and ambient noise signal, respectively. Based on this, a formulation using the gain function is alternatively given by [7]:

---

**SUMMARY** In this letter, we propose a speech reinforcement technique based on soft decision under both the far-end and near-end noise environments. We amplify the estimated clean speech signal at the far-end based on the estimated ambient noise spectrum at the near-end, as opposed to reinforcing the noisy far-end speech signal, so that it can be heard more intelligibly in far-end noisy environments. To obtain an effective reinforcement technique, we adopt the soft decision scheme incorporating a speech absence probability (SAP) in the frequency dependent signal-to-noise ratio (SNR) recovery method where the clean speech spectrum is estimated and the reinforcement gain is inherently derived and modified within the unified framework. Performance of the proposed method is evaluated by a subjective testing under various noisy environments. This is an improvement over previous approaches.

**key words:** soft decision, SAP, speech reinforcement, SNR recovery, masking effect, near-end ambient noise estimation

---

*Manuscript revised March 30, 2009.*

*Manuscript revised March 30, 2009.*

†The authors are with School of Electronic Engineering, Inha University, Incheon, Korea.

a) E-mail: changjh@inha.ac.kr

DOI: 10.1587/transfun.E92.A.2116
\[ G(t, k) \geq \sqrt{\frac{\Phi_{NN}(t, k)}{\Phi_{YY}(t, k)}} \]  
(3)

Since, in our approach, the far-end signal is limited in its amplification, \( G(t, k) \) must also satisfy the following condition [7]:
\[ G(t, k) \geq 1. \]  
(4)

Combining (3) and (4), we then have [7]
\[ G(t, k) = \max \left\{ \sqrt{\frac{\Phi_{NN}(t, k)}{\Phi_{YY}(t, k)}}, 1 \right\}. \]  
(5)

Since \( G(t, k) \) is restricted by maximum gain (\( G_{\text{max}} \equiv 30 \text{ dB} \)), we arrive at [7]:
\[ G(t, k) = \min \left\{ \max \left\{ \sqrt{\frac{\Phi_{NN}(t, k)}{\Phi_{YY}(t, k)}}, 1 \right\}, G_{\text{max}} \right\}. \]  
(6)

For this, \( \Phi_{YY}(t, k) \) and \( \Phi_{NN}(t, k) \) are estimated using a recursive averaging of the periodogram such that [7]
\[ \Phi_{YY}(t, k) = \alpha_Y \Phi_{YY}(t-1, k) + (1 - \alpha_Y)|Y(t,k)|^2 \]
\[ \Phi_{NN}(t, k) = \alpha_N \Phi_{NN}(t-1, k) + (1 - \alpha_N)|N(t,k)|^2 \]  
(7)

where \( \alpha_Y = 0.996 \) and \( \alpha_N = 0.96 \) are the smoothing parameters in our experiments.

3. Proposed Speech Reinforcement Based on Soft Decision

In standard spectral reinforcement rules, the boosting gain applied to each short-time spectral coefficient simply depends on the far-end signal level \(|Y(t,k)|\) measured in the current frame. Indeed, as specified in (6), since \( G(t, k) \) can be expressed as a function of \( \Phi_{YY}(t, k) \) and \( \Phi_{NN}(t, k) \), both speech and noise at the far-end could be amplified when the far-end speech is originally contaminated by noise at the far-end. To avoid this intelligibility of speech degradation phenomenon, we first derive a straightforward way to estimate the clean speech spectra from noisy speech. We then reinforce the estimated clean spectra. This is considered efficient since we can boost only the clean speech component without amplifying noise. To derive the estimated clean speech spectrum at the far-end, we employ the soft decision-based technique proposed in [2]. For this, specifically we first adopt a basic assumption to classify each frequency bin of the far-end noisy speech as either speech or noise [1]:

\[ S(t, k) = Y(t, k) = D(t, k) \]
(8)

where \( S(t, k) \) and \( D(t, k) \) denote the DFT coefficients of clean speech and the added noise at the far-end, respectively.

Assuming that the noisy spectral components are characterized by zero-mean complex Gaussian distributions [4], the probability density functions are conditioned on two hypotheses of \( H_0 \) and \( H_1 \) [1]:
\[ p(Y(t,k)|H_0) = \frac{1}{\pi \lambda_d(t,k)} \exp \left\{ -\frac{\Phi_{YY}(t,k)}{\lambda_d(t,k)} \right\} \]
\[ p(Y(t,k)|H_1) = \frac{1}{\pi \lambda_s(t,k) + \lambda_d(t,k)} \exp \left\{ -\frac{\Phi_{YY}(t,k)}{\lambda_s(t,k) + \lambda_d(t,k)} \right\} \]  
(9)

where \( \lambda_s(t,k) \) and \( \lambda_d(t,k) \) indicate the variances of speech and noise for the \( t \)th spectral component at the \( k \)th frame. Based on the observation at the far-end, the SAP probability of speech absence. Also, the likelihood ratio \( \Lambda(Y(t,k)) \) at the \( k \)th frequency is obtained by [1]
\[ \Lambda(Y(t,k)) = \frac{p(Y(t,k)|H_0)}{p(Y(t,k)|H_1)} \]
\[ = \frac{\frac{1}{1 + \frac{\lambda_H0(Y(t,k))}{\lambda_H1(Y(t,k))}}}{1 + \frac{\lambda_H1(Y(t,k))}{\lambda_H0(Y(t,k))}} \]
\[ = 1 + \frac{\lambda_H0(Y(t,k))}{\lambda_H1(Y(t,k))} \]  
(10)

in which \( p(H_0) = (1 - p(H_1)) \) is the \textit{a priori} probability of speech absence. Also, the likelihood ratio \( \Lambda(Y(t,k)) \) at the \( k \)th frequency is obtained by [1]
\[ \Lambda(Y(t,k)) = \frac{p(Y(t,k)|H_0)}{p(Y(t,k)|H_1)} \]
\[ = \frac{1}{1 + \frac{\lambda_H0(Y(t,k))}{\lambda_H1(Y(t,k))}} \]
\[ = \frac{\lambda_H1(Y(t,k))}{\lambda_H0(Y(t,k))} \]  
(11)

where the predicted SNR \( \gamma(t,k) \) and a posteriori SNR \( \xi(t,k) \) are defined by [1]
\[ \gamma(t,k) = \frac{\Phi_{YY}(t,k)}{\Phi_{DD}(t,k)} \]
\[ \xi(t,k) = \frac{\Phi_{SS}(t,k)}{\Phi_{DD}(t,k)}. \]  
(12)

For the implementation of an on-line algorithm, long-term smoothed noisy speech and speech power estimates are employed as follows [1]:
\[ \hat{\Phi}_{SS}(t+1,k) = \zeta_s \hat{\Phi}_{SS}(t,k) + (1 - \zeta_s)\Phi_{SS}(t,k) \]
\[ \hat{\Phi}_{DD}(t+1,k) = \zeta_s \hat{\Phi}_{DD}(t,k) + (1 - \zeta_s)\Phi_{DD}(t,k) \]  
(13)
in which \( \hat{\Phi}_{DD} \) and \( \Phi_{SS} \) are the estimates of \( \Phi_{DD} \) and \( \Phi_{SS} \), respectively. Also, \( \zeta_s \) (0.99) and \( \zeta_n \) (0.98) are the smoothing parameters with 0 < \( \zeta_s \), \( \zeta_n \) < 1. Using (10) and the statistical assumption for \( S(t) \) and \( D(t) \), the estimates of the speech and noise power spectrum are obtained by the following soft decision scheme [1]:
\[ \hat{\Phi}_{SS}(t,k) = E\left[|S(t,k)|^2|Y(t,k), H_0\|Y(t,k)\) \]
\[ + E\left[|S(t,k)|^2|Y(t,k), H_1\|Y(t,k)\) \]
\[ \hat{\Phi}_{DD}(t,k) = E\left[|D(t,k)|^2|Y(t,k), H_0\|Y(t,k)\) \]
\[ + E\left[|D(t,k)|^2|Y(t,k), H_1\|Y(t,k)\). \]
\[ (14) \]
\[ (15) \]
We derive the proposed reinforcement algorithm based on the SNR recovery method incorporating (6), (7) and (14) such that [7]

\[
G(t, k) = \min \left\{ \frac{\phi_{NN}(t, k)}{\phi_{SS}(t, k)}, G_{\max} \right\}
\]  

(16)

where it can be seen that \(\phi_{YY}(t, k)\) of the conventional method as specified in (6) is substantially replaced by \(\phi_{SS}(t, k)\) in the proposed scheme.

Here, in the case of \(G(t, k)\) specified in (16), we can not obtain a reliable value because it does not take full advantage of the soft decision which prevents us from obtaining robust estimates. Now, we propose another possibility of using the soft decision by modifying the reinforcement rule is to replace \(G(t, k)\) by \(\hat{G}(t, k)\) incorporating \(p(H_0|Y(t, k))\) in (10) as following:

\[
\hat{G}(t, k) = G(t, k)(1 - p(H_0|Y(t, k))) + G_{\min} p(H_0|Y(t, k))
\]  

(17)

where \(G_{\min} = 1\) denotes the minimum value of the reinforcement gain. From this, we can see that \(\hat{G}(t, k)\) replaces \(G(t, k)\) boosting the spectra within the active speech periods on each frequency bin (i.e., \(p(H_0|Y(t, k)) \approx 0\)). On the other hand, in the case of speech absence, \(G(t, k)\) corresponds to 1 (no reinforcement) which could be considered more realistic. Also, in the case of transition periods from speech to silence (\(0 < p(H_0|Y(t, k)) < 1\)), \(\hat{G}(t, k)\) represents the smoothed version of \(G(t, k)\) and \(G_{\min}\) with more statistical reliability. Practically, this is a major contribution of this letter in that the proposed technique has better speech intelligibility compared to the conventional schemes.

Finally, in the proposed method as shown in Fig. 2, the modified reinforcement gain derived from the soft decision specified in (17) is combined with the far-end speech DFT coefficients as given by

\[
\hat{Y}(t, k) = \hat{G}(t, k) Y(t, k).
\]  

(18)

From Fig. 3 showing the representative example, the proposed method seems to perform better than the conventional scheme by taking full advantage of the soft decisions. In comparison, the proposed scheme seems does not boost the noise power degrading speech intelligibility.

4. Experimental Results

Since the purpose of speech reinforcement techniques is to improve intelligibility of far-end speech signals, we performed qualitative tests under various ambient noise conditions. Eight test phrases, spoken by four male and four female speakers whose aged from 20 to 35 from the NTT database [1], were used as experimental data. Each phrase consisted of two different 8-second long meaningful sentences. To add far-end noise, white noise from the NOISEX-92 database was added to the clean speech waveform at SNR 10 dB. In addition, each speech signal at the near-end was degraded by various noise types: white, babble, and vehicular with SNRs of 5, 10, 15, and 20 dB.

Informal subjective tests via a comparison category rating (CCR) were performed in order to assess performance evaluation. Fourteen Korean listeners with normal hearing (eight male and six female) who aged from 20 to 35 participated in the experiment where we simply added the near-end and far-end signal before being played out at the headphone [8]. The CCR test method sheds light on perception quality of the signal of method A over method B. The grades of the seven point scale range are as follows: 3 (much better), 2 (better), 1 (slightly better), 0 (about), −1 (slightly worse), −2 (worse), −3 (much worse).

At first, as given in Table 1, the quality of the signal reinforced (or processed) by the proposed algorithm was compared to that of the unprocessed signal in the aforementioned noisy condition. The results obtained with the proposed method were much better than that of the unprocessed speech when we see the average difference (= 1.54) compared to the average 95% confidence interval (= ±0.17). Secondly, the performance of the proposed approach (in terms of speech quality) was compared with that of the signal reinforced by the previous SNR recovery method as given in Table 2. From the fact that the average difference
Table 1: The CCR results for the proposed reinforced algorithm with respect to the unprocessed under various ambient noise (With 95% confidence interval).

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR (dB)</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>white</td>
<td>5</td>
<td>1.49 ± 0.14</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.46 ± 0.16</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>1.36 ± 0.19</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.33 ± 0.20</td>
</tr>
<tr>
<td>babble</td>
<td>5</td>
<td>1.49 ± 0.14</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.50 ± 0.15</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>1.48 ± 0.19</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.55 ± 0.18</td>
</tr>
<tr>
<td>vehicle</td>
<td>5</td>
<td>1.61 ± 0.14</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.87 ± 0.15</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>1.70 ± 0.19</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.68 ± 0.21</td>
</tr>
</tbody>
</table>

Table 2: The CCR results for the proposed reinforced algorithm with respect to the SNR Recovery algorithm under various background noise (With 95% confidence interval).

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR (dB)</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>white</td>
<td>5</td>
<td>0.77 ± 0.14</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.86 ± 0.15</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>1.07 ± 0.16</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1.04 ± 0.18</td>
</tr>
<tr>
<td>babble</td>
<td>5</td>
<td>0.57 ± 0.14</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.53 ± 0.15</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.52 ± 0.17</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.46 ± 0.17</td>
</tr>
<tr>
<td>vehicle</td>
<td>5</td>
<td>0.11 ± 0.13</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.26 ± 0.13</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.14 ± 0.14</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.18 ± 0.16</td>
</tr>
</tbody>
</table>

(= 0.543) is much bigger than the average 95% confidence interval (= ±0.15), we can see that the proposed algorithm is found to improve the speech quality performance with the soft decision scheme. The performance gains for the white (Average difference = 0.935) and babble (Average difference = 0.52) noise environment is higher than those of the vehicular noise environments. This could be attributable to the fact that the vehicular noise frequency components were concentrated in lower frequency ranges, which could be masked by speech.

5. Conclusion

We have presented a novel approach, based on an estimated near-end background noise spectrum, to enhance clarity and intelligibility of far-end speech under noisy near-end environments. Since estimation of the clean speech spectrum (from the far-end noisy speech) is feasible in a soft decision scheme, amplification of the signal only (that is, without noise amplification) is then done based on the soft decision scheme. The performance of the proposed approach has been shown to be superior to conventional SNR recovery techniques as evidenced by the subjective preference CCR test results.

Acknowledgement

This work was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute for Information Technology Advancement) (IITA-2008-C1090-0902-0010) and This work was partly supported by the IT R&D program of MKE/IITA [2008-F-045-01].

References