Improved Global Soft Decision Using Smoothed Global Likelihood Ratio for Speech Enhancement

Joon-Hyuk CHANG††*, Member, Dong Seok JEONG†, Nam Soo KIM††, and Sangki KANG††††, Nonmembers

SUMMARY In this letter, we propose an improved global soft decision for noisy speech enhancement. From an investigation of statistical model-based speech enhancement, it is discovered that a global soft decision has a fundamental drawback at the speech tail regions of speech signals. For that reason, we propose a new solution based on a smoothed likelihood ratio for the global soft decision. Performances of the proposed method are evaluated by subjective tests under various environments and show better results compared with our previous work.

key words: global soft decision, SGLR, GSAP, MOS

1. Introduction

Recently, there has been an increasing interest in noisy speech enhancement for speech recognition and transmission since the presence of noise significantly degrades the performance of the systems [1]–[11]. Many approaches have been investigated in order to achieve speech enhancement [12]–[20]. In these techniques, both the spectra of clean speech and added noise are usually characterized by uncorrelated statistical models with several associated parameters. Based on the assumed statistical models, a specific function of the clean speech spectrum is estimated conditioned on the observed noisy speech spectrum. Since the statistical model assumed during the period of speech absence differs from that given when speech is present, the spectral estimation approach should proceed with the considerations on speech absence or presence. In general, speech enhancement algorithms which are based on soft decision gain modification have better performance rather than those employing hard decision where each frame is classified into one of the two (speech absence and presence) cases with the help of a voice activity detection (VAD) algorithm [8], [16], [17], [21]. Recently, a novel spectral enhancement algorithm based on global soft decision was proposed in our previous work [12]. The term global means that the decision is performed globally in a given frame instead of being made independently in each spectral component.

In this letter, we propose an improved global soft decision based on the smoothed likelihood ratio for speech enhancement. We devise a robust method to compute the global speech absence probability (GSAP) incorporating the smoothed version of the global likelihood ratio and apply it not only to spectral gain modification but also to the update of noise spectrum estimate for which a separate VAD algorithm is required in the conventional approaches. The performances of the proposed algorithm are evaluated by the mean opinion score (MOS) tests and they are shown better than those of the method used in [12].

2. Review of Global Soft Decision

Let \( Y(t) = [Y_1(t), Y_2(t), \ldots, Y_M(t)] \) denote the spectrum of the noisy speech signal at \( t \)th frame with \( Y_k(t) \) being the \( k \)th spectral component. Given two hypotheses, \( H_0 \) and \( H_1 \), which respectively indicate speech absence and presence, it is assumed that

\[
H_0 : Y(t) = N(t) \\
H_1 : Y(t) = X(t) + N(t)
\]

where \( N(t) = [N_1(t), N_2(t), \ldots, N_M(t)] \) represents the spectrum of the added noise which is uncorrelated with the clean speech spectrum, \( X(t) = [X_1(t), X_2(t), \ldots, X_M(t)] \). As in a number of other speech enhancement algorithms, we also assume that \( X(t) \) and \( N(t) \) are characterized by separate zero-mean complex Gaussian distributions [16], and consequently it is obtained that

\[
p(Y_k(t)|H_0) = \frac{1}{\pi \lambda_{n,k}} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{n,k}} \right\} \\
p(Y_k(t)|H_1) = \frac{1}{\pi (\lambda_{n,k} + \lambda_{s,k})} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{n,k} + \lambda_{s,k}} \right\}
\]

in which \( \lambda_{s,k}(t) \) and \( \lambda_{n,k}(t) \) are the variances of the clean speech and noise in the \( k \)th frequency bin. Conditioned on the current observation \( Y(t) \), the SAP, \( p(H_0|Y(t)) \) is given by

\[
p(H_0|Y(t)) = \frac{p(Y(t)|H_0)p(H_0)}{p(Y(t)|H_0)p(H_0) + p(Y(t)|H_1)p(H_1)}
\]

where \( p(H_0) \) (= 1 – \( p(H_1) \)) is the a priori probability for speech absence. Since the spectral component in each frequency bin is assumed to be statistically independent, (3) can be rewritten as

\[
p(H_0|Y(t)) = \frac{p(H_0)p(Y(t)|H_0)}{p(H_0)p(Y(t)|H_0)p(H_1)|p(Y(t)|H_1)}
\]
in which $q$ is the ratio defined by

$$q = \frac{p(H_1)}{p(H_0)}$$

and $\Lambda_C$ is a global likelihood ratio (GLR). The GLR is computed by $\Lambda_C(t) = \prod_{k=1}^{K} \Lambda_k(Y_k(t))$. Here, $\Lambda_k(Y_k(t))$ is the likelihood ratio computed in the $k$th frequency bin as follows [8]:

$$\Lambda_k(Y_k(t)) = \frac{p(Y_k(t)|H_1)}{p(Y_k(t)|H_0)}$$

and $\Lambda_k(Y_k(t)) = \frac{1}{1 + \xi_k(t)} \exp \left[ \gamma_k(t) \xi_k(t) \right]$$

where

$$\xi_k(t) = \frac{\lambda_{ak}(t)}{\lambda_{nk}(t)}$$

$$\gamma_k(t) = \frac{|Y_k(t)|^2}{\lambda_{ak}(t)}$$

and $\xi_k(t)$ and $\gamma_k(t)$ are called the a priori SNR and the a posteriori SNR, respectively [13].

### 3. Improved Global Soft Decision Using Smoothed Likelihood Ratio

Some techniques are addressed to prevent the clipping of weak speech tail by modifying the decision on made on the current frame such as the HMM-based hang-over technique [8]. On the other hand, recently, a smoothed LR was introduced to avoid the drawback of the LR for the robust voice activity detection which is known to be a simple but effective method [9]. Specifically, at offset regions, LR becomes too low since the a priori SNR becomes a major parameter in the calculation of the LR when the a posteriori SNR is low. In a similar reason, we consider the smoothed global likelihood ratio (SGLR) to enhance the performance of the GSAP in [12].

Here, we briefly describe the adopted speech enhancement framework given the global soft decision [13]. Actually, the performance of the global soft decision approach adopting (4) depends mostly on the reliable estimation of $|\lambda_{ak}(t)|$ and $|\lambda_{nk}(t)|$. Under a general stationarity assumption of $N(t)$ and $X(t)$, we use the long-term smoothed power spectra of the background noise and clean speech to avoid the drawback of the frame-by-frame implementation as the estimates for $|\lambda_{ak}(t)|$ and $|\lambda_{nk}(t)|$, respectively. Let $\hat{\lambda}_{ak}(t)$ and $\hat{\lambda}_{nk}(t)$ be the estimates for $\lambda_{ak}(t)$ and $\lambda_{nk}(t)$. Then,

$$\hat{\lambda}_{ak}(t+1) = \xi_{ak} \hat{\lambda}_{ak}(t) + (1 - \xi_{ak}) E \left[ |N_k(t)|^2 |Y(t)| \right]$$

$$\hat{\lambda}_{nk}(t+1) = \xi_{nk} \hat{\lambda}_{nk}(t) + (1 - \xi_{nk}) E \left[ |X_k(t)|^2 |Y(t)| \right]$$

Based on (9) and the statistical assumptions made on $X(t)$ and $N(t)$, we get

$$E \left[ |N_k(t)|^2 |Y(t)\right] = E \left[ |N_k(t)|^2 |Y(t), H_0\right] p(H_0|Y(t)) + E \left[ |N_k(t)|^2 |Y(t), H_1\right] p(H_1|Y(t))$$

and

$$E \left[ |X_k(t)|^2 |Y(t)\right] = E \left[ |X_k(t)|^2 |Y(t), H_0\right] p(H_0|Y(t)) + E \left[ |X_k(t)|^2 |Y(t), H_1\right] p(H_1|Y(t))$$

where

$$E \left[ |N_k(t)|^2 |Y(t), H_0\right] = |Y_k(t)|^2$$

$$E \left[ |N_k(t)|^2 |Y(t), H_1\right] = \frac{1}{1 + \xi_k(t)} \hat{\lambda}_{nk}(t) + \frac{1}{1 + \hat{\xi}_k(t)} |Y_k(t)|^2$$

and

$$E \left[ |X_k(t)|^2 |Y(t), H_0\right] = 0$$

$$E \left[ |X_k(t)|^2 |Y(t), H_1\right] = \frac{1}{1 + \hat{\xi}_k(t)} |Y_k(t)|^2$$

with

$$E \left[ |X_k(t)|^2 |Y_k(t)\right] = \frac{\hat{\lambda}_{nk}(t)}{\hat{\lambda}_{nk}(t)} \hat{\xi}_k(t) + (1 + \hat{\xi}_k(t)) |Y_k(t)|^2$$

and $\xi_{ak}(=0.99)$ and $\xi_{nk}(=0.97)$ are the smoothing parameters. It is not difficult to see from (9) that $\hat{\lambda}_{ak}(t)$ and $\hat{\lambda}_{nk}(t)$ do not rely on the current observation, $Y(t)$, which implies that they are some kind of predicted estimates of relevant parameters from the previous frame.

Actually, it is frequently discovered that the GSAP given by (4), (7) and (12) becomes too high at the offset regions due to the low LR since we predict the a priori SNR at the previous frame although it is known that the predicted estimates are more accurate than the a priori SNR in estimating GSAP. For that reason, we require a necessity for the SGLR to emphasize the continuity at speech tails such that

$$\Lambda_C(t) = \kappa \Lambda_C(t-1) + (1 - \kappa) \prod_{k=1}^{M} \Lambda_k(Y_k(t))$$

where $\kappa(=0.9)$ is the long-term smoothing parameter which was an experimentally optimized value. It is noted that smoothing scheme is adopted to the raw GLR since the GLR is used in the computation step of the GSAP, while the smoothed log likelihood ratio is employed to decrease the dynamic range in [9]. It is not difficult to see from Fig. 1 that the GSAP seems to alleviate aforementioned problem, taking into account the relatively high SGLR at the offset regions without serious degradation in the performance at the onset regions.

In addition, we review the notion of the spectral gain in the noise suppression stage. Let $\hat{X}(t)$ =

\[\text{[In [12], it is called the predicted SNR for emphasizing the different derivation step.]}\]
Fig. 1 Examples of the GSAP (a) Noisy speech (b) Clean speech (c) Examples of the GSAP; GSAP [12] (solid line) and GSAP using the SGLR (dotted line).

Table 1 MOS results for the proposed enhancement algorithm (SGLR) and conventional SEGSD technique [12] (with 95% confidence interval).

<table>
<thead>
<tr>
<th>noise</th>
<th>white</th>
<th>babble</th>
<th>vehicular</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR(dB)</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>none</td>
<td>1.24±0.07</td>
<td>1.56±0.08</td>
<td>2.14±0.11</td>
</tr>
<tr>
<td>SEGSD</td>
<td>2.50±0.09</td>
<td>3.20±0.09</td>
<td>3.61±0.10</td>
</tr>
<tr>
<td>SGLR</td>
<td>2.71±0.08</td>
<td>3.32±0.09</td>
<td>3.74±0.10</td>
</tr>
</tbody>
</table>

\[
[\hat{X}_1(t), \hat{X}_2(t), \ldots, \hat{X}_M(t)]
\]
denote the estimated clean speech spectrum at \(t\)th frame. In most of the conventional spectral enhancement techniques, \(\hat{X}(t)\) is obtained by applying a specific gain to each spectral component of the noisy speech signal. Among a number of ways to compute the spectral gains, we choose the noise suppression rule proposed by Ephraim and Malah [16], because of its superiority in reducing musical noise phenomena after enhancement [22].

### 4. Experimental Results

In order to evaluate the performance of the proposed enhancement algorithm, which we denote by SGLR, we conducted subjective quality test experiments under various noisy conditions. Ten test sentences in which five were spoken by a male speaker and the others were generated by a female speaker were used for evaluation. Three types of noise sources, the white, babble and buccaneer noises from the NOISEX-92 database were added to the clean speech waveforms by varying SNR. For a fair comparison, we applied the SGLR scheme to previous our work [12] with which only the GLR of the GSAP computation module was switched by the SGLR. Opinion scores were decided by a group of ten listeners, and then averaged to yield the final MOS results. Table 1 shows the MOS results where for the purpose of comparison. From the MOS results, we can see that in most noisy conditions the proposed method consistently yielded higher scores than the previous our method which is the SEGSD enhancement algorithm [12]. Therefore, it is evident that the incorporation of the proposed SGLR scheme has definitely a positive effect compared to the SEGSD in terms of the subjective quality of enhanced speech.

### 5. Conclusions

We have proposed a novel spectral enhancement algorithm in which the SGLR scheme is adopted. To improve the performance of the estimates of the GSAP at the transition periods, we present the simple but efficient scheme for robust estimation of the GLR. The performance of the proposed approach has been found superior to the previous enhancement technique through a number of MOS evaluation tests.

### Acknowledgement

This work was supported in part by MIC & IITA through IT Leading R&D Support Project. And, this work was partly supported by ETRI SoC Industry Promotion Center.
References


tion 3 for wideband spectrum digital systems,” 1996.