Enhancement of Noisy Speech by Using Improved Global Soft Decision

Vladimir I. Shin, Doh-Suk Kim, Moo Young Kim†, Jeongsu Kim

Human & Computer Interaction Laboratory, Samsung AIT
San 14-1, Nongseo-ri, Kiheung-eup, Yongin-city
Kyonggi-do 449-712, Korea
{vishin,ds,kimjs2}@sait.samsung.co.kr

Abstract

We propose a novel speech enhancement algorithm, termed improved global soft decision (IGSD). IGSD is a unified framework for global soft decision on speech absence/presence, noise spectrum estimation, spectral gain modification based on Ephraim-Malah noise suppression. In IGSD, speech absence probability (SAP) is the most important factor, and we propose an efficient and novel SAP estimation in which the SAP is derived based on the general hypothesis for speech absence/presence. In IGSD, the global SAP based on the global hypothesis for speech absence/presence is used to prevent from the problem caused by insufficient amount of data, but more general hypothesis is utilized in the derivation of global SAP estimation. The performance of IGSD is evaluated both subjectively and objectively, and the quality of speech is improved significantly, compared with conventional GSD speech enhancement algorithm.

1. Introduction

As the mobile communication systems are deployed more and more, the systems suffer from various kinds of background noise. And the quality and intelligibility of speech might be significantly degraded in the presence of background noise, especially when speech signal is subject to subsequent processing. Particularly in speech coding systems, the quality degradation caused by background noise is more severe in the speech processed by speech coding systems than in the input speech. Thus, speech enhancement algorithms have therefore been attracted a great of interest in the past two decades.

Usually speech enhancement problem is addressed in terms of estimation point of view in which the clean speech is estimated under the uncertainty of speech presence in noisy observations. The idea of utilizing the uncertainty of speech presence in the noisy spectral components was applied by many authors to improve the performance of speech enhancement systems [1, 2, 3, 4]. In most of works, the speech absence probability (SAP) is defined "*locally*" for each frequency component, irrespective of other components [5, 6, 7].

However insufficient amount of the data makes the appropriate speech enhancement algorithms statistically unreliable. In order to overcome this problem, global soft decision (GSD) concept was proposed and was shown to perform better than IS-127 standard enhancement method [8]. The term "*global*" was used because the decision is performed globally using the data at all frequency channels in a given time frame [8]. GSD allows to estimated noise power spectrum from noisy speech not only when speech is absent but also when it is present. Also, GSD provides a robust procedure for SAP estimation, spectral gain modification, and noise spectrum estimation in unified framework. In GSD framework, SAP plays a key role and therefore the performance of speech enhancement systems depends on accuracy of SAP estimation. In this paper we presents the improved GSD (IGSD), in which novel SAP estimation is employed based on more precise statistical hypothesis.

This paper is organized as follows. IGSD concept is presented in section 2. The speech enhancement algorithm employing IGSD is described in section 3. Experimental results that compare GSD and IGSD are shown in section 4, followed by the conclusion.

2. Improved Global Soft Decision

Assume that $y(n)$ is clean speech corrupted by additive noise, i.e., $y(n) = s(n) + d(n)$, where $s(n)$ and $d(n)$ denote clean speech and noise, respectively. Then we can consider the statistical model employing two hypotheses, H_0 and H_1, which indicate speech absence and presence, respectively:

$$\begin{aligned} H_0: &\quad Y_k(m) = D_k(m), \\ H_1: &\quad Y_k(m) = S_k(m) + D_k(m), \end{aligned} \tag{1}$$

for $k = 1, 2, \ldots, N$. Here $Y_k(m), S_k(m)$, and $D_k(m)$ denote the short-time Fourier transformations of $y(n), s(n)$, and $d(n)$ for the m-th frame, respectively. Here, k denotes the index of frequency channel and N is the total number of frequency channels. We also assume that $S_k(m)$ and $D_k(m)$ are independent and complex Gaussian random variables with zero mean and variances, $\lambda_{s,k}(m)$ and $\lambda_{d,k}(m)$, respectively:

$$\begin{aligned} p(Y_k(m) \mid H_0) &= \frac{1}{\pi\lambda_{d,k}(m)} \exp\left[-\frac{\mid Y_k(m) \mid^2}{\lambda_{d,k}(m)}\right], \\ p(Y_k(m) \mid H_1) &= \frac{1}{\pi[\lambda_{s,k}(m) + \lambda_{d,k}(m)]} \\ &\quad \cdot \exp\left[-\frac{\mid Y_k(m) \mid^2}{\lambda_{s,k}(m) + \lambda_{d,k}(m)}\right], \end{aligned} \tag{2}$$

for $k = 1, 2, \ldots, N$. In some frequency-domain speech enhancement algorithms, the estimation of $\lambda_{d,k}(m)$ is obtained during the period of speech absence by using a voice activity detector. However, the noise spectrum may also change during speech activity, and this causes the performance degradation. On the other hand, the application of soft decision allows to update $\lambda_{d,k}(m)$ even when speech is present, by using the concept of SAP [8]. In this case, the estimation of SAP becomes critical to the performance of speech enhancement. Thus, the success

This work was partly supported by the Critical Technology Program of Korean Ministry of Science and Technology. † He is now with the Dept. Speech, Music and Hearing, KTH, Sweden.

of failure of the algorithm is mostly dependent on the reliable estimation of SAP.

Kim and Chang [8] proposed the GSD concept to overcome the insufficient amount of the data problem, and the reliable estimate of SAP is expressed as

$$\begin{aligned} p(H_0|Y(m)) &= \frac{p(H_0, Y(m))}{p(Y(m))} \\ &= \frac{1}{1 + q\prod_{k=1}^{N} \Lambda_k(m)}, \end{aligned} \tag{3}$$

where $q = P(H_1)/P(H_0)$, $P(H_0)$ is the *a priori* probability, $P(H_1) = 1 - P(H_0)$, and $Y(m) = [Y_1(m), Y_2(m), \ldots, Y_N(m)]$. The likelihood ratio, $\Lambda_k(m)$, is defined by

$$\Lambda_k(m) = \frac{p(Y_k(m) \mid H_1)}{p(Y_k(m) \mid H_0)}. \tag{4}$$

The denominator $p(Y(m))$ in (3) represents the joint probability of the spectral components, $Y_1(m), \ldots, Y_N(m)$. Based on the statistical independence assumption in their derivation of (3) [8], $p(Y(m))$ are represented as

$$\begin{aligned} p(Y(m)) &= P(H_0)p(Y(m) \mid H_0) + P(H_1)p(Y(m) \mid H_1) \\ &= P(H_0)\prod_{k=1}^{N} p(Y_k(m) \mid H_0) \\ &\quad + P(H_1)\prod_{k=1}^{N} p(Y_k(m) \mid H_1). \end{aligned} \tag{5}$$

Also, the numerator $p(H_0, Y(m))$ in (3) takes the form

$$p(H_0, Y(m)) = P(H_0)\prod_{k=1}^{N} p(Y_k(m) \mid H_0). \tag{6}$$

The derivations in (5) and (6) are based on the hypothesis in (1), which is global in the sense that the model is independent of frequency channel. However, speech can be absent or present depending on the channel in general case. Thus, we propose the improved GSD (IGSD) by taking into account of more general hypothesis, in which $p(Y(m))$ and $p(H_0, Y(m))$ are represented as

$$\begin{aligned} p(Y(m)) &= \prod_{k=1}^{N} [p(Y_k(m))] \\ &= \prod_{k=1}^{N} [P(H_0)p(Y_k(m) \mid H_0) \\ &\quad + P(H_1)p(Y_k(m) \mid H_1)] \end{aligned} \tag{7}$$

and

$$p(H_0, Y(m)) = \prod_{k=1}^{N} [P(H_0)p(Y_k(m) \mid H_0)]. \tag{8}$$

Consequently the accurate and novel representation of SAP can be obtained as

$$\begin{aligned} p(H_0|Y(m)) &= \frac{p(H_0, Y(m))}{p(Y(m))} \\ &= \frac{1}{\prod_{k=1}^{N} [1 + q\Lambda_k(m)]}. \end{aligned} \tag{9}$$

In this IGSD framework, the global SAP is used to prevent from the problem of insufficient amount of data, but the evaluation of SAP is based on the local hypothesis, which is more general than the global one in (1). Thus, IGSD provides robust and more precise estimation of SAP.

3. Speech Enhancement Algorithm Using IGSD

This section provides a brief description of speech enhancement incorporating SAP ((3) or (9)), in which the noise suppression rule proposed by Ephraim and Malah (EMNS) [3] was adopted.

In EMNS, the estimate of clean speech spectrum is obtained as

$$\hat{S}_k(m) = G(\xi_k, \gamma_k)Y_k(m), \tag{10}$$

where $\xi_k(m)$ and $\gamma_k(m)$ are *a priori* and *a posteriori* signal-to-noise ratios (SNR's), defined respectively as

$$\xi_k(m) \equiv \frac{\lambda_{s,k}(m)}{\lambda_{d,k}(m)} \quad \text{and} \quad \gamma_k \equiv \frac{|\, Y_k(m)\,|^2}{\lambda_{d,k}(m)}. \tag{11}$$

Also, spectral gain $G(\cdot, \cdot)$ is given by

$$G(\xi, \gamma) = \frac{\sqrt{\pi}}{2}\sqrt{\frac{\xi}{\gamma(1+\xi)}}M\left[\frac{\gamma\xi}{1+\xi}\right], \tag{12}$$

where

$$M[\theta] = \exp(-\theta/2)\left[(1+\theta)I_0(\theta/2) + \theta I_1(\theta/2)\right] \tag{13}$$

with I_0 and I_1 being the modified Bessel functions of zero and first order, respectively. The estimation of the *a priori* and *a posteriori* SNR's in (11), based on "decision-direct" approach [3], is represented as

$$\hat{\gamma}_k(m) = \frac{|\, Y_k(m)\,|^2}{\hat{\lambda}_{d,k}(m)} \tag{14}$$

and

$$\hat{\xi}_k(m) = \alpha\frac{|\hat{S}_k(m-1)|^2}{\hat{\lambda}_{d,k}(m-1)} + (1-\alpha)F\left[\hat{\gamma}_k(m) - 1\right], \tag{15}$$

where $0 \le \alpha < 1$, and $F[x] = x$ if $x \ge 0$ and $F[x] = 0$ otherwise.

According to [8], the noise spectrum is adaptively estimated by

$$\hat{\lambda}_{d,k}(m+1) = \zeta_d\hat{\lambda}_{d,k}(m) + (1-\zeta_d)\varphi_{d,k}(m), \tag{16}$$

where

$$\varphi_{d,k}(m) = |Y_k(m)|^2 p(H_0 \mid Y(m)) + \left[\left(\frac{\hat{\xi}_k(m)}{1+\hat{\xi}_k(m)}\right) \cdot \hat{\lambda}_{d,k}(m) + \left(\frac{1}{1+\hat{\xi}_k(m)}\right)^2 |Y_k(m)|^2\right] p(H_1 \mid Y(m)).$$

It is worth noting that the noise spectrum, $\lambda_{d,k}(m)$, is updated only when the SAP ((3) or (9)) is larger than a pre-defined threshold, i.e.,

$$p(H_0 \mid Y(m)) \ge p_{th}, \tag{17}$$

in order to prevent from significant speech distortion caused by misadaptation [8]. By using the assumption on the distribution of $S_k(m)$ and $D_k(m)$ (refer to (2)), the likelihood ratio $\Lambda_k(m)$ in (9) can be estimated as

$$\hat{\Lambda}_k(m) = \frac{1}{1+\hat{\xi}_k(m)} \exp\left[\frac{\hat{\gamma}_k(m)\hat{\xi}_k(m)}{1+\hat{\xi}_k(m)}\right]. \quad (18)$$

Speech spectrum, $\hat{\lambda}_{s,k}$, is also estimated similarly as in (16), but it is updated every frame,

$$\hat{\lambda}_{s,k}(m+1) \quad = \zeta_s \hat{\lambda}_{s,k}(m) + (1-\zeta_s)\varphi_{s,k}(m), \quad (19)$$

where

$$\varphi_{s,k}(m) = \left[\left(\frac{1}{1+\hat{\xi}_k(m)}\right)\hat{\lambda}_{s,k}(m) + \left(\frac{\hat{\xi}_k(m)}{1+\hat{\xi}_k(m)}\right)^2 |Y_k(m)|^2\right] p(H_1 \mid Y(m)).$$

4. Experimental Results

This section presents the performance an objective and subjective evaluation of the speech enhancement using GSD and IGSD.

4.1. Objective Evaluation

Considering the difference between clean speech and the output of speech enhancement system as error signal, let us define segmental SNR as

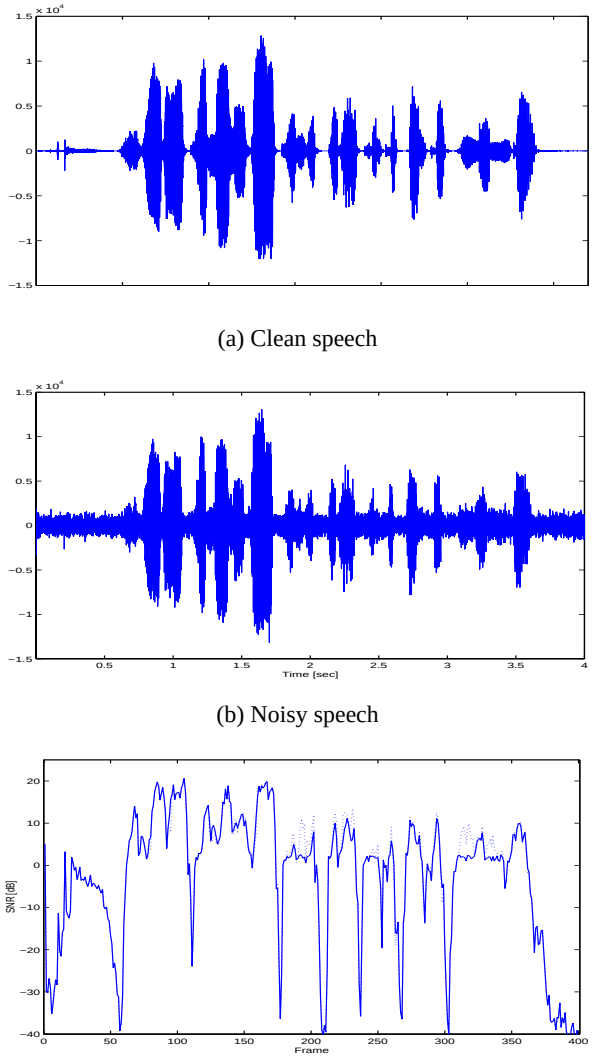
$$\mathrm{SNR}(m) = 10\log \frac{\sum_{i=0}^{L-1} s^2(mL+i)}{\sum_{i=0}^{L-1}\left[s(mL+i)-\hat{s}(mL+i)\right]^2} \quad (20)$$

where L is the number of samples per frame m [9].

Fig. 1 shows an example of segmental SNR's obtained for GSD and IGSD. Upper plot shows clean speech, and the middle one is the speech corrupted by additive car noise of 10 dB SNR. The noisy speech shown in the middle plot was processed by either GSD or IGSD, and the segmental SNR's are depicted in the lowest panel (GSD=solid, IGSD=dotted). In Fig. 1 (c), we can see clearly that IGSD improves the performance compared with GSD. The improvements reaches up to 8 dB, especially in speech regions which have relatively low SNR's.

4.2. Subjective Evaluation

Up until this day, the most reliable way to measure the quality of speech is to perform listening tests. Thus, an absolute category rating test was performed in which the mean opinion score (MOS) was measured to evaluate the performance of the proposed IGSD. In order to make a fair comparison, various parameters for both GSD and IGSD were made the same. The frame interval for both speech enhancement algorithms was set 10 ms, and the total number of frequency channels was set 16. In addition, a priori probability and $q = P(H_0)/P(H_1)$ were determined empirically as $P(H_0) = 0.9411$ and $q = 0.0625$, respectively. Total of three kinds of noises for several SNR's were considered to compare the performance of speech enhancement algorithms in various environmental conditions. Eight Korean



(a) Clean speech

(b) Noisy speech

(c) Segmental SNR

Figure 1: An example of (a) clean speech, (b) noisy speech and (c) segmental SNR's for GSD (solid) and IGSD (dotted).

sentence pairs from four males and four females (sampling rate = 8 kHz), processed by either GSD or IGSD algorithm, were presented monaurally to eight normal-hearing listeners (4 males and 4 females) in a sound proof room, and Table 1 shows the result of the listening test. Here, the quality of speech without being processed by speech enhancement algorithms is also shown in the third column to provide sufficient impression on the quality improvements.

As clearly seen in Table 1, both speech enhancement algorithms enhance the quality for clean speech as well as noisy speech. The maximum improvement in quality is as big as 1.97 MOS (white Gaussian noise of 20 dB SNR). Also, subjective results showed that IGSD is superior to GSD for most of noisy conditions. This is clearly supporting the benefit of improved estimation of speech absence probability.

IGSD performs better than GSD for white Gaussian noise and car noise (as large as 0.24 MOS for 20 dB white Gaussian noise), but the performance of IGSD for speech corrupted by babble noise is comparable to that of GSD. This seems to be

Table 1: MOS results of GSD and IGSD.

Noise	SNR [dB]	None	GSD	IGSD
None	-	4.47	4.73	4.70
White Gaussian	10	1.17	2.17	2.27
	20	1.41	3.14	3.38
Babble	10	2.09	2.73	2.69
	20	3.09	3.47	3.52
Car	10	2.19	2.67	2.78
	15	2.58	3.06	3.16
	20	2.92	3.50	3.61

caused by the disagreement between statistical independence assumptions made in the formulation of speech enhancement algorithms and the characteristics of babble noise.

5. Conclusions

We proposed the IGSD speech enhancement algorithm in this paper. It provides a robust and unified framework in SAP estimation, spectral gain modification, and noise spectrum estimation. In IGSD, the global SAP based on the global hypothesis for speech absence/presence is used to prevent from the problem caused by insufficient amount of data, but more general hypothesis is utilized in derivation of global SAP estimation. By this framework, we can obtain more accurate SAP estimation without having the problem of insufficient data. Objective and subjective experiments demonstrated that the proposed IGSD performs better than GSD.

6. References

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, 1979.

[2] R. McAulay and M. Malpass, "Speech enhancement using a soft decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 137–145, 1980.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109–1121, 1984.

[4] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, pp. 1526–1555, 1992.

[5] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, pp. 789–792, 1999.

[6] J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, pp. 363–366, 1993.

[7] I. Soon, S. Koh, and C. Yeo, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence," *Signal Processing*, vol. 75, pp. 151–159, 1999.

[8] N. Kim and J. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, pp. 108–110, 2000.

[9] R. G. Goldberg and L. Riek, *A Practical handbook of speech coders*. N.W. Corporate Blvd., Boca Raton, Florida 33431: CRC Press, 2000.