平成 30 年度徳山工業高等専門学校特別研究論文集 No.23 (2019)
Graduation Research Reports No.23 (2019), National Institute of Technology, Tokuyama College

バイアス付き倍音復元技術における内部パラメータと音質の関係性の調査

宇根昌和*

Relationship between internal parameters and sound quality in biased harmonic regeneration technique

Masakazu UNE

This paper focuses on two representative problems of single-microphone noise reduction: speech distortion and musical noise. Many noise reduction techniques have been proposed for each problem. For speech distortion, harmonic regeneration noise reduction (HRNR) has been proposed. The HRNR is a method to use a unique signal to regenerate the eliminated harmonics, and improve the estimation accuracy of a priori signal-to-noise ratio (SNR). I have proposed a new noise reduction method which addresses speech distortion and musical noise problems simultaneously by introducing bias into a priori SNR estimator in HRNR (biased HRNR). In this paper, I investigate the behavior of each internal parameter of HRNR and biased HRNR toward speech quality and show the effectiveness of biased HRNR in terms of speech distortion and musical noise. As a result, I found that the bias introduced in the proposed HRNR does not deteriorate the speech quality due to its high-quality noise reduction feature, in consideration of the relation between the estimation accuracy of a priori SNR and speech quality.

Key words : harmonic regeneration, musical noise, musical-noise-free speech enhancement, a priori SNR estimation

1. Introduction

Recent mobile phone has a speech recognition system and recording system including the voice memorandum as well as the verbal communication feature. In order to utilize these systems comfortably, a problem due to the background noise should be resolved. Many noise reduction techniques have been proposed for the problem [1–8]. Multi-channel noise reduction methods such as based on beam-forming [1] and source separation [2] need many microphones. Moreover, the inverse matrix calculation is needed for treating the plural microphone inputs and leads to instability of the applications. Thus, they are not appropriate for the mobile phone in terms of the cost or the scale. On the other hand, single-channel noise reduction [3–8] is low cost, small scale, and small computational complexity. For this reason, single-channel noise reduction is significant to communicate or record a message using the small-sized device.

However, the output speech obtained by single-channel noise reduction exists two critical problems, i.e., speech distortion and musical noise [9–14]. Speech distortion causes the deterioration of the speech articulation and speech recognition accuracy by suppressing the target signal excessively. For the speech distortion problem, harmonic regeneration noise reduction (HRNR) has been proposed [15,16]. The HRNR focuses on the fact that most of the speech distortion are harmonic components and overcome this problem by estimating a priori signal-to-noise ratio (SNR) using a unique signal to restore the harmonics.

As another problem, musical noise makes the output speech unnatural and humans uncomfortable. Then, some methods which suppress noisy signal without generating musical noise have also been proposed [17–19]. These methods (called *musical-noise-free speech enhancement*) are used for systems that human listens to [20,21]. In particular, musical-noise-free speech enhancement based on minimum-mean square error short-time spectral amplitude (MMSE-STSA) estimator (hereafter referred to as a musical-noise-free MMSE-STSA estimator) [18] achieves

* 情報電子工学専攻（指導教員：宮崎亮一）

lower speech distortion than that based on spectral subtraction [17]. In the literature [18], Nakai, et al. introduced bias factor into a classical a priori SNR estimator [5] and succeeded in suppressing the musical noise generation. I have introduced a bias into a priori SNR estimation in HRNR and proposed a new noise reduction technique (biased HRNR) which can achieve lower speech distortion than musical-noise-free MMSE-STSA estimator and no musical noise generation [22].

As I have mentioned above, a priori SNR is estimated by using a unique signal to restore the harmonics in HRNR process or introducing bias in biased HRNR process. The speech quality notably depends on the estimated a priori SNR [23]. HRNR and biased HRNR have the internal parameters [15,22], the internal parameters determine the estimation accuracy of a priori SNR and the speech quality. In [22], the relationship between the internal parameters and the speech quality have demonstrated in extremely limited condition. Revealing the relationship will be the foundation to discover optimal parameters [24,25].

In this paper, I investigate the relationship between the internal parameters and speech quality in HRNR and biased HRNR exhaustively. Also, I consider the relationship between the estimation accuracy of a priori SNR which is determined by the internal parameters, and speech quality.

This paper is organized as follows. In Sec. 2., I show the common noise reduction process and HRNR to overcome the speech distortion problem. In Sec. 3., I explain biased HRNR I have proposed and the a priori SNR estimator. In Sec. 4., I demonstrate the relationship between the internal parameter and the sound quality in the classical a priori SNR estimator, HRNR, and biased HRNR. Finally, in Sec. 5., I present the behavior and the estimation accuracy in each a priori SNR estimator and discuss the relationship between the estimation accuracy and the sound quality.

2. Noise reduction methods and a priori SNR estimation

In this section, I explain the classical a priori SNR estimator and MMSE-STSA estimator proposed by Ephraim and Malah. The MMSE-STSA estimator takes advantage of this a priori SNR estimator. I also explain HRNR which has been proposed to overcome the speech distortion.

2.1 *Classical a priori SNR estimator*

An observed speech in the time domain $x(t)$ is given by

$$x(t) = s(t) + n(t), \tag{1}$$

where $s(t)$ and $n(t)$ are clean and noise speech, respectively. Applying the short-time Fourier transform (STFT) into Eq. (1), k th spectral component ($0 \le k \le K$) of short-time frame p ($0 \le p \le P$) of the observed speech $X(p,k)$ is expressed by

$$X(p,k) = S(p,k) + N(p,k), \tag{2}$$

where $S(p,k)$ and $N(p,k)$ represent the clean and the noise speech spectra, respectively. Hereinafter, I omit the component k and the time frame p unless otherwise stated. Actually, the only X is obtained in a real environment. Generally, the estimate of the clean speech $\hat{S}$ is obtained by multiplying an appropriate spectral gain G by observed speech X as follows:

$$\hat{S} = GX. \tag{3}$$

Spectral gains in common noise reduction techniques such as Wiener filter [4] and MMSE-STSA estimator [5] are expressed as a function of a priori SNR ξ and a posteriori SNR γ:

$$G = g(\xi, \gamma), \tag{4}$$

where $g(\cdot,\cdot)$ is the spectral gain function. The a priori SNR ξ and the a posteriori SNR γ are respectively defined by

$$\xi = \frac{\mathrm{E}\left[|S|^2\right]}{\mathrm{E}\left[|N|^2\right]}, \tag{5}$$

and

$$\gamma = \frac{|X|^2}{\mathrm{E}\left[|N|^2\right]}, \tag{6}$$

where $\mathrm{E}[\cdot]$ is the expectation operator. $\mathrm{E}\left[|N|^2\right]$ is approximated by expected value $\mathrm{E}\left[\left|\hat{N}\right|^2\right]$ of speech absent (noise only) area up to frame T, i.e.,

$$\mathrm{E}\left[|N(p,k)|^2\right] \approx \mathrm{E}\left[\left|\hat{N}\right|^2\right] = \frac{1}{T}\sum_{\tau=0}^{T} |X(\tau,k)|^2 . \tag{7}$$

The a priori SNR ξ is not able to obtain in a real environment, thus, it is estimated using decision-directed (DD) approach [5] as follows:

$$\xi^{\mathrm{DD}}(p,k) = \alpha \frac{\left|\hat{S}(p-1,k)\right|^2}{\mathrm{E}\left[\left|\hat{N}\right|^2\right]} + (1-\alpha)\,\mathrm{Max}\left[\gamma(p,k)-1, 0\right], \tag{8}$$

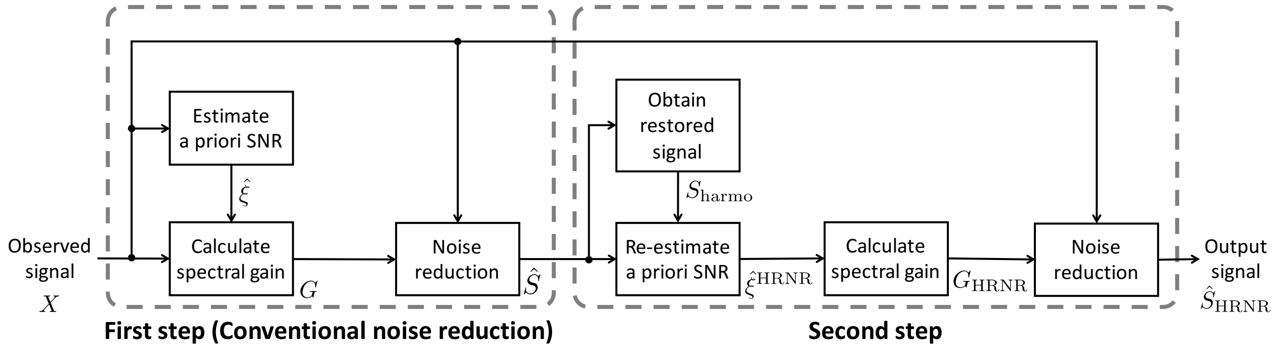

Fig. 1: Brock diagram in HRNR. First step indicates common noise reduction process and output is obtained in second step.

where the internal parameter α is a forgetting factor which controlls the sound quality and is better to set to 0.98 [5], and Max $[a, b]$ returns the larger value.

2.2 *MMSE-STSA estimator*

The MMSE-STSA estimator minimizes the error between the true and estimated speech in amplitude domain [5]. The spectral gain in MMSE-STSA estimator is expressed as a function of a priori SNR ξ and a posteriori SNR γ:

$$g(\xi,\gamma) = \frac{\sqrt{\nu}}{\gamma}\Gamma\left(\frac{3}{2}\right) M\left(-\frac{1}{2};1;-\nu\right), \tag{9}$$

$$\nu = \frac{\xi}{1+\xi}\gamma, \tag{10}$$

where $\Gamma(\cdot)$ and $M(a;b;z)$ are gamma and Kummer functions, respectively. I estimate the a priori SNR from Eq. (8), the final output speech $\hat{S}_{\mathrm{STSA}}$ is calculated from

$$\hat{S}_{\mathrm{STSA}} = G_{\mathrm{STSA}}X = g(\hat{\xi}^{\mathrm{DD}},\gamma)X. \tag{11}$$

2.3 *HRNR*

Approximately 80 % of the pronounced words are voiced sound in a human language. It is known that the power spectrum of the voiced speech decreases as the higher frequency. Due to the small power of the voiced sound, the components in the bandwidth are regarded as noise and suppressed. HRNR focuses on this point and regenerates the higher frequency components (harmonics) mainly suppressed to resolve the speech distortion [15]. Figure 1 shows the block diagram in HRNR. There are two noise reduction steps in HRNR process. The temporal estimation of the speech signal $\hat{S}$ is obtained by the classical noise reduction method such as MMSE-STSA estimator in the first step (mentioned in Sec. 2.2). Next, I apply the temporal estimated signal to the following non-linear function and obtain the restored signal S_{harmo} as follows:

$$S_{\mathrm{harmo}} = \mathcal{F}\left[\mathrm{Max}\left[\mathcal{F}^{-1}\left[\hat{S}\right],0\right]\right], \tag{12}$$

where $\mathcal{F}[\cdot]$ and $\mathcal{F}^{-1}[\cdot]$ respectively indicate the Fourier and the inverse Fourier transforms. The restored signal S_{harmo} which regenerates the pseudo spectrum of the original speech cannot be used directly since contains an unnatural spectrum not in the original one. However, the restored signal has useful information for the harmonic components. The new a priori SNR $\hat{\xi}^{\mathrm{HRNR}}$ is re-estimated by

$$\hat{\xi}^{\mathrm{HRNR}} = \frac{\rho\left|\hat{S}\right|^2 + (1-\rho)\left|S_{\mathrm{harmo}}\right|^2}{\mathrm{E}\left[\left|\hat{N}\right|^2\right]} \tag{13}$$

using the restored signal and the weighting factor ρ ($0 \leq \rho \leq 1$) which corresponds to the internal parameter in HRNR. Moreover, the weighting factor ρ is better to set to the spectral gain which is obtained in the first step [15]. In other words, if the noise reduction method is MMSE-STSA estimator, let ρ be G_{STSA} which is computed by Eq. (9). To distinguish between the case for the spectral gain and one for the constant value, I define the former case as

$$\hat{\xi}^{\mathrm{HRNR}}_{\mathrm{gain}} = \frac{G\left|\hat{S}\right|^2 + (1-G)\left|S_{\mathrm{harmo}}\right|^2}{\mathrm{E}\left[\left|\hat{N}\right|^2\right]}, \tag{14}$$

and the latter case as

$$\hat{\xi}^{\mathrm{HRNR}}_{\mathrm{const}} = \frac{\rho\left|\hat{S}\right|^2 + (1-\rho)\left|S_{\mathrm{harmo}}\right|^2}{\mathrm{E}\left[\left|\hat{N}\right|^2\right]}. \tag{15}$$

Finally, I obtain the new spectral gain and the output in HRNR by using the new a priori SNR such as Eq. (11), i.e.,

$$\hat{S}_{\mathrm{HRNR}} = G_{\mathrm{HRNR}}X = g(\hat{\xi}^{\mathrm{HRNR}},\gamma)X. \tag{16}$$

3. Biased HRNR

It is generally known that introducing the bias into the DD algorithm leads to suppression of the musical noise generation [10]. The biased HRNR is a method to overcome the speech distortion and musical noise generation problems by introducing bias into a priori SNR estimator in HRNR. Three types are given in the way as introducing bias when I change the a priori SNR estimator in HRNR. The first type sets to the bias in the first term (called *1 term*):

$$\hat{\xi}^{\text{1term}} = \rho' \,\text{Max}\left[\frac{\left|\hat{S}\right|^2}{\text{E}\left[\left|\hat{N}\right|^2\right]}, \varepsilon'\right] + (1-\rho')\frac{|S_{\text{harmo}}|^2}{\text{E}\left[\left|\hat{N}\right|^2\right]}, \tag{17}$$

the second type sets in whole the equation:

$$\hat{\xi}^{\text{whole}} = \text{Max}\left[\frac{\rho'\left|\hat{S}\right|^2 + (1-\rho')\,|S_{\text{harmo}}|^2}{\text{E}\left[\left|\hat{N}\right|^2\right]}, \varepsilon'\right], \tag{18}$$

and the third type replaces the first term by biased maximum-likelihood term:

$$\hat{\xi}^{\text{ML}} = \rho' \,\text{Max}\left[\gamma - 1, \varepsilon'\right] + (1-\rho')\frac{|S_{\text{harmo}}|^2}{\text{E}\left[\left|\hat{N}\right|^2\right]}. \tag{19}$$

Here, ρ' $(0 \leq \rho' \leq 1)$ and ε' represent the weighting factor and the bias value, respectively. In this paper, I adopted the 1term version $\hat{\xi}^{\text{1term}}$ as the representative of a priori SNR estimator in the biased HRNR since no distinct differences were found among these estimators in the preliminary study. To distinguish between the case for the spectral gain $\hat{\xi}^{\text{1term}}_{\text{gain}}$ and one for the constant value $\hat{\xi}^{\text{1term}}_{\text{const}}$ in a priori SNR estimator as well as HRNR, namely,

$$\hat{\xi}^{\text{1term}}_{\text{gain}} = G\,\text{Max}\left[\frac{\left|\hat{S}\right|^2}{\text{E}\left[\left|\hat{N}\right|^2\right]}, \varepsilon'\right] + (1-G)\frac{|S_{\text{harmo}}|^2}{\text{E}\left[\left|\hat{N}\right|^2\right]}, \tag{20}$$

$$\hat{\xi}^{\text{1term}}_{\text{const}} = \rho' \,\text{Max}\left[\frac{\left|\hat{S}\right|^2}{\text{E}\left[\left|\hat{N}\right|^2\right]}, \varepsilon'\right] + (1-\rho')\frac{|S_{\text{harmo}}|^2}{\text{E}\left[\left|\hat{N}\right|^2\right]}. \tag{21}$$

4. Behavoir between internal parameter and sound qulity

In Sec. 2., I described a priori SNR estimators in HRNR. I also explained the biased HRNR in Sec. 3.. Each sound quality is decided by adjusting the internal parameters. However, the relationship between the internal parameters and the sound quality is not revealed in detail. In this section, I experimentally investigate the relationships among the noise reduction level, the musical noise generation, and the speech distortion.

4.1 *DD approach and HRNR*

DD approach decides the sound quality by the forgetting factor α [26]. It is supposed that the output in HRNR depends on the forgetting factor α since DD approach is used in the first step in HRNR (see Fig. 1). I examine the sound qualities by adjusting the forgetting factor α in the DD approach and the weighting factor ρ in HRNR.

The observed signals were generated by adding the babble noise (BB) or railway station noise (RS) with 10 dB input SNR. The forgetting factor α in Eq. (8) was set from 0.00 to 0.99. In HRNR, the gain function of MMSE-STSA estimator was adopted in the first noise reduction step and the weighting factor ρ in Eq. (15) was set from 0.0 to 1.0. The forgetting factor α in ξ^{DD} was set 0.5, 0.7 and 0.98. Additionally, I also examined a case that ρ is replaced as the spectral gain G in the first step, i.e., $\hat{\xi}^{\text{HRNR}}_{\text{gain}}$ in Eq. (14) was used as a priori SNR estimator in HRNR. I calculated the noise reduction level, the musical noise generation, and the speech distortion of each output signal with the following objective measurements.

Noise reduction rate (NRR) is a measure of noise reduction level [17] and higher NRR indicates the higher SNR improvement. NRR is computed as the defference between the input and the output signal as follows:

$$\text{NRR} = 10\log_{10}\frac{\text{E}\left[|s_{\text{out}}|^2\right] / \text{E}\left[|n_{\text{out}}|^2\right]}{\text{E}\left[|s_{\text{in}}|^2\right] / \text{E}\left[|n_{\text{in}}|^2\right]}, \tag{22}$$

where s_{in} and s_{out} are the input and the output speech signals, and n_{in} and n_{out} are the input and the output noise signals, respectively.

Next, kurtosis ratio (KR) [27] which quantifies the musical noise generation is defined by

$$\text{KR} = \frac{\text{Kurt}_{\text{proc}}}{\text{Kurt}_{\text{org}}}, \tag{23}$$

where Kurt_{org} and $\text{Kurt}_{\text{proc}}$ are kurtosis of the original and the processed signals, respectively. The small KR (> 1.0) indicates the few musical noise generations, and KR less than 1.0 means no generation of musical noise (refer to *musical-noise-free condition*).

Finally, I introduce cepstral distortion (CD) for measurement of speech distortion [28]. Using the cepstral coefficients of the clean speech C_{ref} and the processed

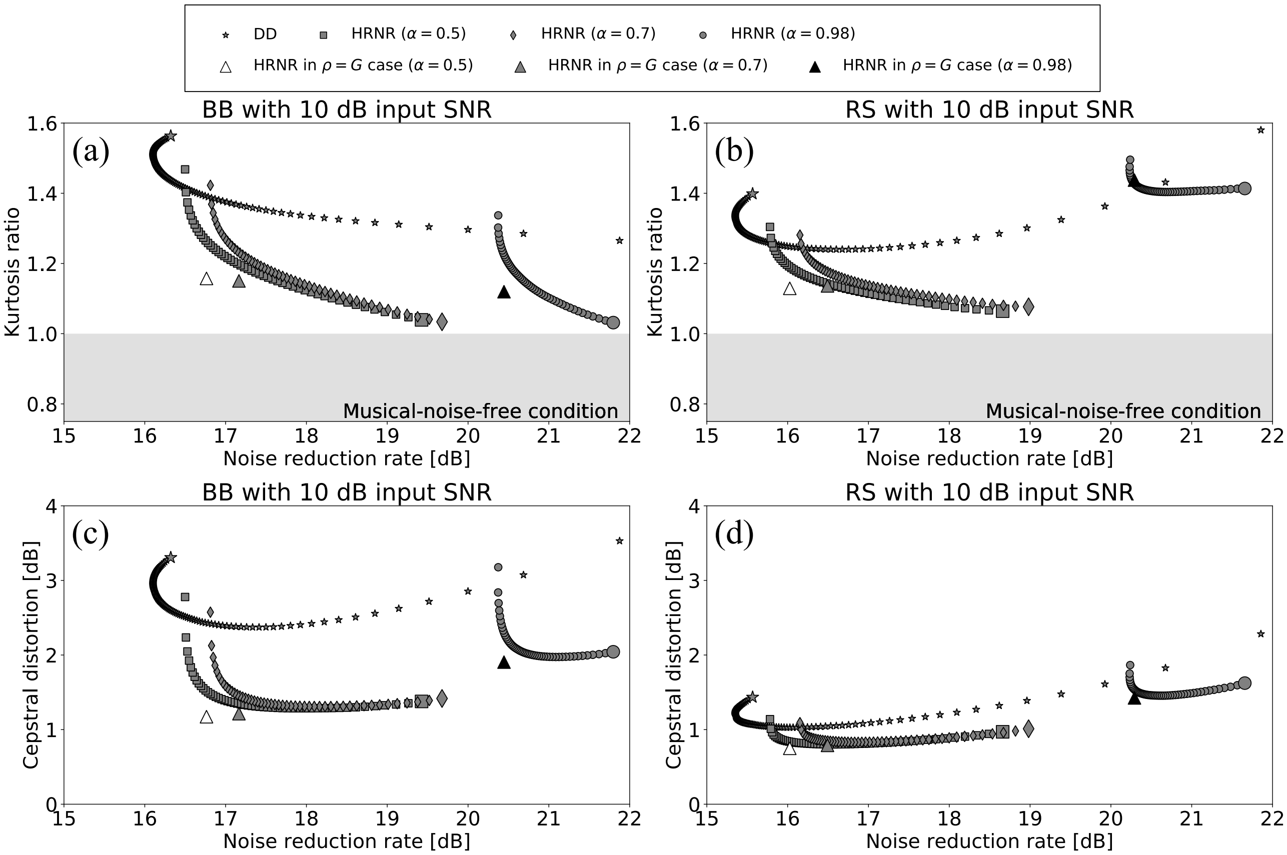

Fig. 2: Results of sound quality by adjusting each internal parameter in DD approach and HRNR. KR versus NRR (top) in (a) BB and (b) RS with 10 dB input SNR. CD versus NRR (bottom) in (c) BB and (d) RS with 10 dB input SNR.

speech C_{out}, CD is calculated by

$$\text{CD} = \frac{20}{P \log 10} \sum_{p}^{P} \sqrt{\sum_{k}^{B} 2(C_{\text{out}}(p,k) - C_{\text{ref}}(p,k))^2}, \quad (24)$$

where B is a dimension of the cepstrum. All signals used in this experiment were sampled as 16 kHz. The hamming window (512 width and 25 % overlap) was applied in the STFT. I provided the speech absent period for computation of the KR. A dimension of the cepstrum B in Eq. (24) was set to 22.

Figures 2 (a) and (b) show the results of KR versus NRR. Some large symbols indicate that each parameter is equal to zero. From Figs. 2 (a) and (b), in terms of the DD approach (the star behavior), NRR increases as increasing the forgetting factor α. In the constant parameter case $\xi_{\text{const}}^{\text{HRNR}}$ (the grey and quadrate, rhomboid or round behaviors), the KR depends on the forgetting factor α in the first step. Namely, the tendencies of the weighting factor ρ are painted in the higher NRR area if the forgetting factor α is large. Moreover, NRR decreases as increasing the weighting factor ρ. On the other hand, in the case for the spectral gain ($\xi_{\text{gain}}^{\text{HRNR}}$), the sound quality is not better than the constant parameter case which parameter is small in terms of NRR and KR. In a comparison between the DD approach and HRNR, I clarify that HRNR is effective in musical noise generation terms.

Next, Figures 2 (c) and (d) show the results of CD versus NRR. Cepstral distortion correspondingly rises when the forgetting factor α increases in the DD approach. The tendencies of the weighting parameter ρ depend on the forgetting factor α as well as the result of KR versus NRR. Cepstral distortion decrease in the part of the low weighting parameter ρ and rapidly increases from a certain point. Since this tendency was found in another condition, I can consider the existence of the optimal value. However, a point of $\xi_{\text{gain}}^{\text{HRNR}}$ marked in the lowest CD of any constant parameters. Hence, I conclude that using spectral gain in a priori SNR estimator in HRNR is the best to keep the speech distortion low.

4.2 *HRNR and biased HRNR*

I described the effectiveness of HRNR in Sec. 4.1. Next, I investigate the relation between the sound quality and the internal parameters of HRNR and biased HRNR. Biased HRNR has two internal parameters (the weighting

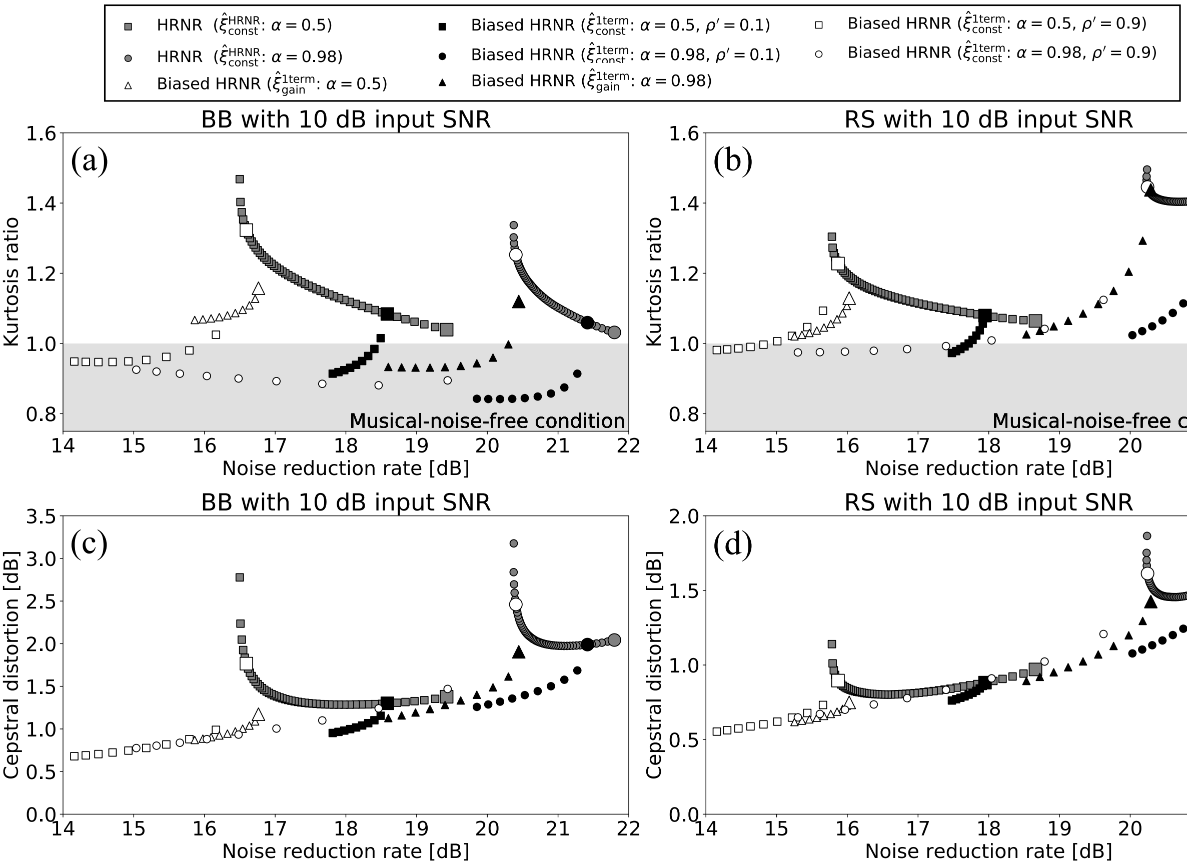

Fig. 3: Results of sound quality by adjusting each internal parameter in HRNR and biased HRNR. KR versus NRR (top) in (a) BB and (b) RS with 10 dB input SNR. CD versus NRR (bottom) in (c) BB and (d) RS with 10 dB input SNR.

factor ρ' and the bias ε'). The weighting factor ρ' was fixed to 0.1 or 0.9. I increased the bias value from 0.0 to 1.0. The value of the forgetting factor α as the first step in biased HRNR was set to 0.7 and 0.98. Other conditions were the same as Sec. 4.1.

I show the behavior of KR by adjusting the internal parameters in Figs. 3 (a) and (b). NRR and KR decrease with the increase of bias value (see the black and white symbols). The biased HRNR suppress the musical noise generation in the same NRR condition. In the constant parameter case $\hat{\xi}^{\mathrm{1term}}_{\mathrm{const}}$ in biased HRNR (the black or white round symbols and quadrate symbols), some symbols reach KR less than 1.0 (musical-noise-free condition) not depending on the weighting parameter ρ'. Furthermore, the weighting parameter ρ' is effective to set to lower value because the higher NRR is better. On the other hand, the tendency of the case with the spectral gain ($\hat{\xi}^{\mathrm{1term}}_{\mathrm{gain}}$) is similar to the one of constant parameter case, but, the points which KR is lower than 1.0 do not exist when noise reduction level is low (i.e., forgetting factor is low). Therefore, the noise reduction level in first step is necessary to large, so that the output in $\hat{\xi}^{\mathrm{1term}}_{\mathrm{gain}}$ can achieve musical-noise-free condition.

Figures 3 (c) and (d) show the tendency by changing each parameter in CD versus NRR. NRR and CD decrease by introducing the bias and biased HRNR prevent the speech distortion compared with HRNR because CD of biased HRNR is lower than in HRNR. Each tendencie in biased HRNR maintains the magnitude correlation and the NRR at $(\rho', \varepsilon') = (0, 0)$ in $\hat{\xi}^{\mathrm{1term}}_{\mathrm{const}}$ is higher than one at $\varepsilon' = 0.0$ in $\hat{\xi}^{\mathrm{1term}}_{\mathrm{gain}}$. In biased HRNR, the constant parameter case $\hat{\xi}^{\mathrm{1term}}_{\mathrm{const}}$ which is small value more effective performance than a case using spectral gain $\hat{\xi}^{\mathrm{1term}}_{\mathrm{gain}}$ in terms of the speech distortion. As a result, biased HRNR is high-quality noise reduction technique comprehensively compared to HRNR and DD approach, and the weighting parameter is effective to use the lower constant value ρ'.

5. Relationship between sound quality and accuracy of a priori SNR

I described the effectiveness of HRNR in Sec. 4.1 and biased HRNR is higher-quality method compared with HRNR in Sec. 4.2. The superiority or inferiority of the sound quality in these methods is only decided by the difference in a priori SNR estimator. From the results in

Sec. 4.1 and Sec. 4.2, I can set up a hypothesis that the high-quality noise reduction method can accurately estimate a priori SNR. Therefore, I investigate the accuracies in each method and discuss the previous hypothesis in this section.

5.1 *Observation of behavior of estimated a priori SNRs*

First, I observe the estimated a priori SNR details in behavior in this section. The observed signal was made by mixing BB with the clean speech at 10 dB input SNR. True a priori SNR was computed by Eq. (5) assuming that clean and noise signals are known. In order to achieve 20 dB NRR in each method, the forgetting factor α in DD approach was set to 0.97, the weighting factor ρ in HRNR was set to 0.04, and the weighting factor ρ' and bias ε' in biased HRNR were set to 0.1 and 0.8, respectively. Other experimental conditions were the same as Sec. 4.1.

Figure 4 shows the result of the behavior of the estimated a priori SNRs. Figures 4 (a) and (b) represent each the estimated a priori SNR in a frame and a frequency bin, respectively. First, I compare true a priori SNR and one by the DD approach. The behavior in DD approach tracks the true one in the lower position (i.e., underestimation). In particular, the underestimation is outstanding at high-frequency (around 4 kHz) in Fig. 4 (b). The underestimation leads to the speech distortion and the harmonic distortion in high frequency part. Additionally, the transition of the behavior by DD approach is slow. This phenomenon is caused by smoothing down using the previous frame at the estimation, the delay is occurred in an area from speech absent to speech presence and produces the underestimation. Therefore, I confirm the a priori SNR estimator of the DD approach triggers the underestimation leading to the speech distortion.

Next, I compare the true a priori SNR and one by HRNR. From Figs. 4 (a) and (b), HRNR estimates a priori SNR excessively (i.e., overestimation). Although the overestimation is caused by a restored signal S_{harmo} which regenerates the pseudo spectrum of the harmonics, the signal S_{harmo} has the harmonics based on the suppressed spectrum $\hat{S}$. Namely, the protrusions in DD approach and HRNR mostly synchronize.

Finally, I consider the effect of bias in biased HRNR. Biased HRNR overestimates a priori SNR when true one is small (see the 200 th frame in Fig. 4 (b)). However, the overestimation by bias does not lead to the deterioration of the speech quality since biased HRNR is the best performance in the methods I mentioned from the experiments in Sec. 4.1 and Sec. 4.2.

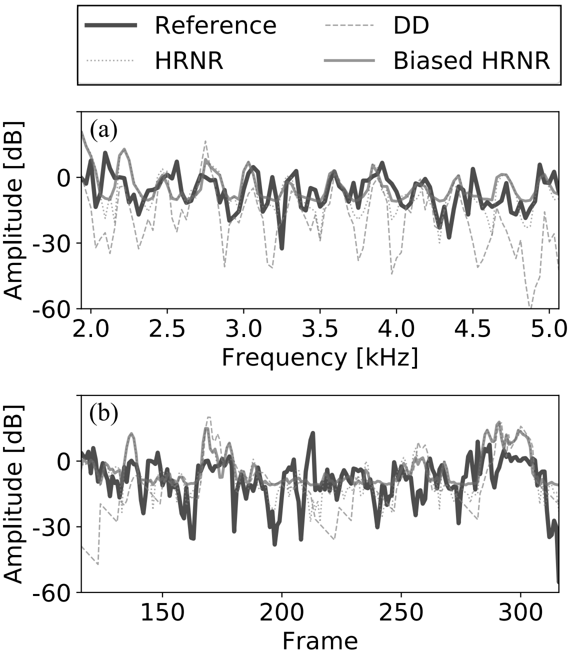

Fig. 4: Comparison of true a priori SNR and estimated ones by each method. (a) Behavior in a frame and (b) behavior in a frequency bin.

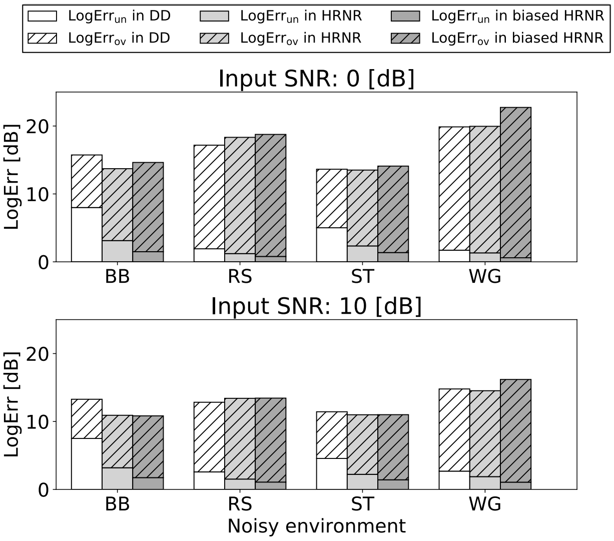

Fig. 5: Results of log-error in 0 dB input SNR (upper) and 10 dB input SNR (lower)

5.2 *Objective evaluation of amounts of overestimation and underestimation*

To support the conjecture regarding overestimation and underestimation I stated in Sec. 5.1, I investigate the observation result objectively. The observed signals were mixed

BB, RS, street noise (ST), or white Gaussian noise (WG) with the clean speech at 0 and 10 dB input SNRs. Log-error (LogErr) was used as the measurement of the amount of the estimation error. LogErr is calculated as the sum of the amounts of underestimation $\text{LogErr}_{\text{un}}$ and overestimation $\text{LogErr}_{\text{ov}}$, i.e.,

$$\text{LogErr} = \text{LogErr}_{\text{un}} + \text{LogErr}_{\text{ov}}, \tag{25}$$

$$\text{LogErr}_{\text{ov}} = \frac{10}{PK}\sum_{p,k}\left|\text{Min}\left[\log_{10}\frac{\xi(p,k)}{\hat{\xi}(p,k)}, 0\right]\right|, \tag{26}$$

$$\text{LogErr}_{\text{un}} = \frac{10}{PK}\sum_{p,k}\text{Max}\left[\log_{10}\frac{\xi(p,k)}{\hat{\xi}(p,k)}, 0\right], \tag{27}$$

where Min $[a, b]$ returns the smaller value. I obtained the a priori SNRs by DD approach, HRNR, biased HRNR, and computed each LogErr with respect to true one. Each parameter was set to achieve the same NRR.

Figure 5 shows the results of LogErr in various noisy environments. Many large underestimations occur in DD approach, and this outcome corresponds to the result in Fig. 4. Next, HRNR suppresses the amount of underestimation compared to DD approach; however, HRNR overestimates a priori SNR more largely. This result also matches the result in Fig. 4. Finally, biased HRNR increases the total error compared to other methods. HRNR and biased HRNR keep the amount of underestimation low and speech distortion is prevented. Biased HRNR reduces the amount of underestimation relative to HRNR, especially. I consider the effect is caused by bias. The result that the amount of the overestimation in biased HRNR is large agrees with the result in Sec. 5.1. However, Sec. 4.2 and Sec. 5.1 have mentioned biased HRNR is comprehensively high-quality noise reduction method; therefore, the overestimation by bias does not degrade the speech quality. Rather, introducing bias leads to a favorable result.

5.3 *Discussion*

The fact I mentioned in Sec. 5.2 gives some tasks. As an obvious case, if a true a priori SNR is used in Eq. (11), I can obtain the significantly high-quality output. A priori SNR close to a true one is better, intuitively. However, I have indicated introducing bias improves the speech quality even if the estimation accuracy declines. Introducing bias is the representative instantiation as one of the factors.

Many a priori SNR estimators which outperform the DD approach have been proposed [29–33] and consider compensation for frame delay, adaptivizing the forgetting factor, or its optimization. These methods mainly focus on the speech distortion problem; hence, it is important to analyze the effects including the musical noise problem in each method. Elucidation of the special factors including bias, not the estimation accuracy, can be used in the selection of an optimal loss function for the deep neural network.

6. Conclusion

In this paper, I investigated the relationship between the internal parameters and the sound qualities of HRNR and biased HRNR exhaustively. In addition, I compared these estimation accuracies of a priori SNR and a true a priori SNR.

In Sec. 4.1, HRNR is effective compared to the DD approach. The internal parameter is better to set to the small constant value in terms of the musical noise problem. In contrast, speech distortion is low in case that the internal parameter is set spectral gain. Next, in Sec. 4.2, I showed biased HRNR is higher-quality noise reduction technique than HRNR. Additionally, I confirmed that introducing bias reduces the speech distortion and musical noise generation in biased HRNR and the weighting parameter is more effective to set to small constant value.

Finally, in Sec. 5.1 and Sec. 5.2, I confirmed a priori SNR estimator in HRNR prevent the underestimation and suppress the speech distortion. Biased HRNR also prevents the underestimation by the bias and the amount of the overestimation increases. However, I concluded the bias does not deteriorate the speech quality since biased HRNR is high-quality noise reduction method.

References

[1] R. Z. J. L. Flanagan, J. D. Johnston, and G. W. Elko, “Computer-steered microphone arrays for sound transduction in large rooms,” *Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, 1985.

[2] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, and T. Nishikawa, “Blind source separation combining independent component analysis and beamforming,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[3] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] N. Wiener, “Extrapolation, interpolation and smoothing of stationary time series with engineering applications,” *Cambridge, MA: MIT Press*, 1949.

[5] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech and*

Signal Processing, vol. 27, no. 6, pp. 1109–1121, 1984.

[6] K. Yamashita, S. Ogata, and T. Shimamura, “Spectral subtraction iterated with weighting factors,” *Proceedings of IEEE Speech Coding Workshop*, pp. 138–140, 2002.

[7] C. H. You, S. N. Koh, and S. Rahardja, “β-order mmse spectral amplitude estimation for speech enhancement,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 13, no. 5, pp. 475–486, 2005.

[8] J. Benesty and Y. Huang, “A single-channel noise reduction MVDR filter,” *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 273–276, 2011.

[9] P. C. Loizou, “Speech enhancement theory and practice,” *CRC Press, Taylor & Francis Group, FL*, 2007.

[10] O. Cappe, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[11] Z. Goh, K. C. Tan, and B. Tan, “Postprocessing method for suppressing musical noise generated by spectral subtraction,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 287–292, 1998.

[12] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, “Instantaneous a priori SNR estimation by cepstral excitation manipulation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, 2017.

[13] J. S. Erkelens, J. Jensen, and R. Heusdens, “A data-driven approach to optimizing spectral speech enhancement methods for various error criteria,” *Speech Communication*, vol. 49, no. 7, pp. 530–541, 2007.

[14] J. S. Erkelens, J. Jensen, and R. Heusdens, “Improved speech spectral variance estimation under the generalized gamma distribution,” *IEEE BENELUX/DSP Valley Signal Processsing Symposium*, pp. 43–46, 2007.

[15] C. Plapous, C. Marro, and P. Scalart, “Improved signal-to-noise ratio estimation for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, 2006.

[16] M. Une and R. Miyazaki, “Evaluation of sound quality and speech recognition performance using harmonic regeneration for various noise reduction techniques,” *2017 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing*, pp. 377–380, 2017.

[17] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, “Musical-noise-free speech enhancement based on optimized iterative spectral subtraction,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 7, pp. 2080–2094, 2012.

[18] S. Nakai, H. Saruwatari, R. Miyazaki, S. Nakamura, and K. Kondo, “Theoretical analysis of biased mmse short-time spectral amplitude estimator and its extension to musical-noise-free speech enhancement,” *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pp. 122–126, 2014.

[19] H. Saruwatari, “Statistical-model-based speech enhancement with musical-noise-free properties,” *Proceedings of International Conference on Digital Signal Processing*, pp. 1201–1205, 2015.

[20] R. Mukai, S. Araki, H. Sawada, and S. Makino, “Removal of residual cross-talk components in blind source separation using time-delayed spectral subtraction,” *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1789–1792, 2002.

[21] Y. Takahashi, T. Takatani, H. Saruwatari, and K. Shikano, “Blind spatial subtraction array with independent component analysis for hands-free speech recognition,” *Proceedings of International Workshop on Acoustic Echo and Noise Control*, 2006.

[22] M. Une and R. Miyazaki, “Musical-noise-free speech enhancement with low speech distortion by biased harmonic regeneration technique,” *Proceedings of International Workshop on Acoustic Signal Enhancement*, pp. 31–35, 2018.

[23] C. Breithaupt, T. Gerkmann, and R. Martin, “A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing,” *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 4897–4900, 2008.

[24] S. Kubo and R. Miyazaki, “Estimation of spectral subtraction parameter-set for maximizing speech recognition performance,” *5th IEEE Global Conference on Consumer Electronics*, pp. 567–568, 2016.

[25] S. Kubo and R. Miyazaki, “Estimation of beta-order MMSE-STSA parameter set for maximizing speech recognition performance with multiple regression analysis,” *2018 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing*, pp. 180–183, 2018.

[26] S. Kanehara, H. Saruwatari, R. Miyazaki, K. Shikano, and K. Kondo, “Theoretical analysis of musical noise generation in noise reduction methods with decision directed a priori SNR estimator,” *Proceedings of International Workshop on Acoustic Signal Enhancement*, 2012.

[27] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K Kondo, “Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics,” *Proceedings of International Workshop on Acoustic Echo and Noise Control*, 2008.

[28] L. Rabiner and B. Juang, “Fundamentals of speech recognition,” *Upper Saddle River, NJ: Prentice-Hall*, 1993.

[29] M. K. Hansan, S. Salahuddin, and M. R. Khan, “A modified a priori SNR for speech enhancement using spectral subtraction rules,” *IEEE Signal Processing Letters*, vol. 11, no. 4, pp. 450–453, 2004.

[30] S. Suhadi, C. Last, and T. Fingscheidt, “A data-driven approach to a priori SNR estimation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 186–195, 2011.

[31] P. Yun-Sik and J. H. Chang, “A novel approach to a robust a priori SNR estimator in speech enhancement,” *IEICE Transactions on Acoustics Speech and Signal Processing*, vol. 90, no. 8, pp. 2182–2185, 2007.

[32] P. C. Yong, S. Nordholm, and H. H. Dam, “Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement,” *Speech Communication*, vol. 55, no. 2, pp. 358–376, 2013.

[33] L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, “Improved a priori SNR estimation in speech enhancement,” *23rd Asia-Pacific Conference on Communications*, pp. 1–5, 2017.