Applied Acoustics 168 (2020) 107410

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Musical-noise-free noise reduction by using biased harmonic regeneration and considering relationship between a priori SNR and sound quality

Masakazu Une [a], Ryoichi Miyazaki [b,*]

[a] University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, Japan
[b] National Institute of Technology, Tokuyama College, Gakuendai, Shunan, Yamaguchi, Japan

ARTICLE INFO

Article history:
Received 13 September 2019
Received in revised form 30 March 2020
Accepted 29 April 2020

Keywords:
MMSE-STSA estimator
Musical noise
Harmonic regeneration
Musical-noise-free noise reduction
A priori SNR estimation

ABSTRACT

This paper focuses on two representative single-microphone noise reduction problems: speech distortion and musical noise. Many noise reduction methods have been proposed for each problem. Harmonic regeneration noise reduction (HRNR) was introduced for the improvements of speech distortion and the a priori signal-to-noise ratio (SNR) estimator. HRNR involves using a unique signal to regenerate harmonics, which had been eliminated. Musical-noise-free noise reduction based on the minimum-mean square error short-time spectral amplitude estimator (musical-noise-free MMSE-STSA estimator) has also been proposed. This method can suppress a noisy signal without generating musical noise by introducing a bias into the classical a priori SNR estimator. We propose a noise reduction method for addressing these problems simultaneously by improving the classical a priori SNR estimator. We investigated the behavior of the internal parameters for the proposed and conventional methods with regard to speech quality and show the effectiveness of the proposed method in terms of speech distortion and musical noise. We consider and discuss the relationship between the estimation accuracy of an a priori SNR and speech quality. Specifically, we consider the factors to improve speech quality in terms of biasing.

1. Introduction

Recent mobile phones have speech-recognition and recording systems that include voice memorandum as well as a verbal communication function. To effectively use these systems, significant problems due to background noise should be resolved. Many noise reduction methods have been proposed for solving such problems [1–8]. Multi-channel noise reduction methods, such as those based on beam-forming [1] and source separation [2], require many microphones; thus, they are not appropriate in terms of cost and scale. Inverse matrix calculation is also required for treating the spatial characteristics of each microphone input, which leads to system instability. Single-channel microphone noise reduction methods [3–8], however, are low cost, small in scale, and have low computational complexity. For this reason, single-channel microphone noise reduction is essential to communicate or record a message using a small device. However, the output speech with these methods results in two critical problems, i.e., speech distortion and musical noise [9–14]. Speech distortion makes listening difficult and reduces speech-recognition accuracy by excessively suppressing the target signal. Minimum-mean square error short-time spectral amplitude (MMSE-STSA) with harmonic regeneration noise reduction (HRNR) was proposed for the speech-distortion problem [15–18]. Since a kind of speech distortion can be discussed as harmonics, which are removed by excessive noise suppression, HRNR aims to restore these components for high-quality noise reduction and prior SNR estimation [15,16,19,20]. MMSE-STSA estimator with HRNR focuses on the fact that most speech distortions are harmonic component distortions, which are overcome by estimating the a priori signal-to-noise ratio (SNR) using a unique signal to restore the harmonics. The name "HRNR" originally refers to the noise reduction method, but Plapous et al. mainly focus on a priori SNR estimation by a harmonic [15,16]. For this reason, we treat the name "HRNR" as an a priori SNR estimator in this paper.

Other methods for suppressing noisy signals without generating musical noise have also been proposed [21–23]. These methods, collectively called *musical-noise-free noise reduction* theorem, are used for systems that humans use to listen [24,25]. In particular, the musical-noise-free noise reduction method based on the

* Corresponding author.
E-mail addresses: s1920628@s.tsukuba.ac.jp (M. Une), miyazaki@tokuyama.ac.jp (R. Miyazaki).

https://doi.org/10.1016/j.apacoust.2020.107410
0003-682X/

MMSE-STSA estimator (this method is called the musical-noise-free MMSE-STSA estimator) results in lower speech distortion than spectral subtraction based on that [21]. Nakai et al. introduced a bias factor into the classical a priori SNR estimator and succeeded in suppressing musical noise generation. Many noise reduction methods, including HRNR, focus on a priori SNR estimation [26–30]. These methods are based on the most classical a priori SNR estimator proposed by Ephraim and Malah [5], which takes into account compensation for frame delay and adapts to the forgetting factor or its optimization so that the estimated a priori SNR is close to the true one. On the contrary, the musical-noise-free MMSE-STSA estimator places more importance on the perception of musical noise than accurate a priori SNR estimation.

We propose a noise reduction method that applies the biasing concept of the musical-noise-free MMSE-STSA estimator to HRNR (we call this a priori SNR estimator *biased HRNR* and our method *MMSE-STSA estimator with biased HRNR*). Biased HRNR suppresses speech distortion futher than HRNR and generates no musical noise, the same as the musical-noise-free MMSE-STSA estimator. We first experimentally investigated the tendency of musical noise generation and speech distortion by adjusting the internal parameters of HRNR. Our MMSE-STSA estimator with biased HRNR can meet the condition in which musical noise is not generated. We then investigated the behavior of the estimated a priori SNR by each estimator and described the contribution of biasing. Finally, we compared our MMSE-STSA estimator with biased HRNR with conventional methods and found that our method suppresses speech distortion without generating musical noise.

This paper is organized as follows. In Section 2, we discuss the relationship among standard noise reduction processes and conventional a priori SNR estimators including HRNR. In Section 3, we introduce biased HRNR as a new a priori SNR estimator to decrease of the amounts of speech distortion and musical noise. We also discuss the relationship among the internal parameters, sound quality, and the estimation accuracy of a priori SNR for each a priori SNR estimator. Next, we discuss the objective evaluations we conducted to show the effectiveness of our MMSE-STSA estimator with biased HRNR compared to conventional methods in Section 4. Finally, we conclude this paper in Section 5.

2. Related works

In this section, we explain the classical a priori SNR estimator and the MMSE-STSA estimator proposed by Ephraim and Malah. We also explain the MMSE-STSA estimator with HRNR and musical-noise-free MMSE-STSA estimator, which were proposed to overcome speech distortion and musical noise generation, respectively.

2.1. Classical noise reduction and 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 signals, respectively. Applying the short-time Fourier transform (STFT) into Eq. (1), the k th spectral component ($0 \leqslant k \leqslant K$) of short-time frame p ($0 \leqslant p \leqslant P$) 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 noise speech spectra, respectively. Hereafter, we omit components p and k unless otherwise stated. Generally, the estimate of the clean speech $\widehat{S}$ is obtained by multiplying an appropriate spectral gain G by observed speech X as follows:

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

Spectral gains with common noise reduction methods, such as Wiener filter [4] and MMSE-STSA estimator [5], are expressed as a function of an 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 ξ and 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, and $\mathrm{E}\left[|N|^2\right]$ is approximated by the expected $\mathrm{E}\left[\left|\widehat{N}\right|^2\right]$ of the speech-absent (noise only) area up to frame T, i.e.,

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

The ξ cannot be obtained in a real environment; thus, it is estimated using the decision-directed (DD) approach [5] as follows:

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

where internal parameter α is a forgetting factor that controls the sound quality and is best set to 0.98 [5], and $\mathrm{Max}[a,b]$ returns a larger value. We specify the components p and k because Eq. (8) needs the current frame p and the previous frame $p-1$. Since the frame delay generally deteriorates the quality of the output, many a priori SNR estimators for solving this problem have been proposed [26–30].

It is well known that the MMSE-STSA estimator is a classical noise reduction method and it minimizes the error between the true and estimated speech in the amplitude-spectrum domain [5]. The spectral gain of the MMSE-STSA estimator is expressed as a function of ξ and γ:

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

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

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

$$\widehat{S}^{\mathrm{DD}} = G^{\mathrm{DD}}X = g^{\mathrm{STSA}}\left(\hat{\xi}_{\alpha}^{\mathrm{DD}},\gamma\right)X. \tag{11}$$

2.2. MMSE-STSA estimator with HRNR

Approximately 80% of pronounced words are voiced in human languages. It is well known that the power spectrum of a voiced sound decreases as the frequency increases. Due to the small power of a voiced sound, the voiced sound's components are regarded as noise and suppressed, especially in high bandwidth. HRNR focuses on this point and regenerates the higher-frequency components (harmonics) mainly suppressed to resolve speech dis-

tortion [16]. Fig. 1 shows a block diagram of the MMSE-STSA estimator with HRNR; there are two noise reduction steps. Originally, HRNR means the noise reduction process, but we define HRNR as a priori SNR estimator to avoid confusion in this paper. First, we obtained the temporal estimated signal by applying the classical noise reduction as the first step in Fig. 1. Next, the temporal estimated signal is applied to the following non-linear function, and the restored signal S^{harmo} is obtained in the second step in Fig. 1 as follows:

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

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

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

using S^{harmo} and the weighting factor ρ $(0 \leqslant \rho \leqslant 1)$, which corresponds to the internal parameters of HRNR. Moreover, ρ is best set to the spectral gain, which is obtained in the first step [16]. In other words, if we use the MMSE-STSA estimator as the noise reduction method in the first step, let ρ be G, which is computed using Eq. (11). Also, the form which replaces ρ with spectral gain G is expressed as

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

Finally, we obtain the new spectral gain and output from MMSE-STSA estimator with HRNR by using the new a priori SNR $\hat{\xi}_{\rho}^{\text{HRNR}}$ such as Eq. (11), i.e.,

$$\widehat{S}^{\text{HRNR}} = G^{\text{HRNR}}X = g^{\text{STSA}}\left(\hat{\xi}_{\rho}^{\text{HRNR}}, \gamma\right)X. \tag{15}$$

2.3. Musical-noise-free speech enhancement based on biased DD approach

The kurtosis ratio (KR) was proposed [31] as an objective measure for musical noise generation. The KR is defined by $\text{kurt}_{\text{proc}}/\text{kurt}_{\text{org}}$, where $\text{kurt}_{\text{proc}}$ and kurt_{org} are the kurtosis of the processed and observed signals, respectively. Musical noise is perceived as the tail in terms of the probability density function in the power-spectral domain, and kurtosis represents the tail of the distribution. The increase in kurtosis by nonlinear signal processing generates musical noise. Namely, the KR is used to evaluate the amount of musical noise. A large KR (> 1.0) indicates more generation of musical noise and a small KR (< 1.0) indicates no generation of musical noise (*musical-noise-free condition*). The musical-noise-free noise reduction method generates almost no musical noise even with high noise reduction. Miyazaki et al. established the musical-noise-free theorem for spectral subtraction and Wiener filtering as single-channel microphone noise reduction [21]. They also extended the theorem to multi-channel for dealing with nonstationary noise [32]. Although these methods can suppress noise without musical noise generation, the amount of speech distortion is large. Also, Kanehara et al. revealed the theoretical relationship between the amounts of noise reduction and musical noise generation in the MMSE-STSA estimator and concluded that no musical-noise-free condition exists regardless of the value of the internal parameters [33–35]. On the other hand, Nakai et al. discovered the existence of the musical-noise-free condition in the MMSE-STSA estimator for the biased a priori SNR $\hat{\xi}_{\alpha,\varepsilon}^{\text{B-DD}}$ [22], which suppressed speech distortion further than the conventional musical-noise-free methods. The biased a priori SNR $\hat{\xi}_{\alpha,\varepsilon}^{\text{B-DD}}$ is computed to provide the bias factor in the term of maximum likelihood estimation in Eq. (8) and is given by

$$\hat{\xi}_{\alpha,\varepsilon}^{\text{B-DD}}(p,k) = \alpha\frac{|\widehat{S}(p-1,k)|^2}{\text{E}\left[|\widehat{N}(p,k)|^2\right]} + (1-\alpha)\text{Max}[\gamma(p,k)-1, \varepsilon], \tag{16}$$

where ε is the bias value. We specify p and k because Eq. (16) needs the current frame p and previous frame $p-1$. The same as in Eqs. (11) and (15), we obtain the spectral gain and output of the musical-noise-free MMSE-STSA estimator by using the new a priori SNR $\hat{\xi}_{\alpha,\varepsilon}^{\text{B-DD}}$, i.e.,

$$\widehat{S}^{\text{B-DD}} = G^{\text{B-DD}}X = g^{\text{STSA}}\left(\hat{\xi}_{\alpha,\varepsilon}^{\text{B-DD}}, \gamma\right)X. \tag{17}$$

3. Proposed method based on biased HRNR

3.1. Overview

It is generally known that introducing a bias into the DD approach suppresses musical noise generation [10]. Our MMSE-STSA estimator with biased HRNR is used to overcome the speech distortion and musical noise generation problems by introducing a bias into HRNR. The process of the proposed method is illustrated in Fig. 2. In this paper, we express the bias as the following equation:

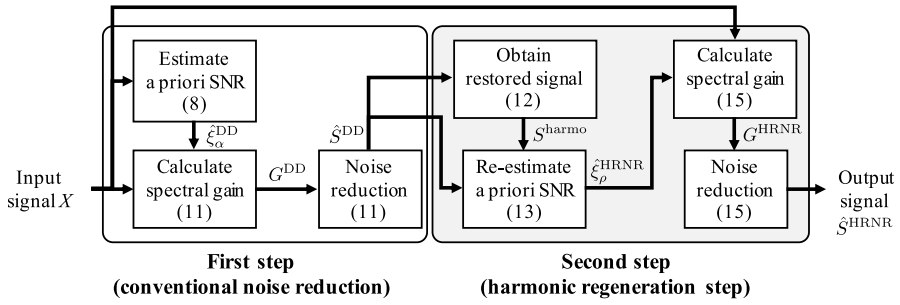

Fig. 1. Block diagram of MMSE-STSA estimator with HRNR. First step indicates common noise reduction process, and second step outputs speech with new a priori SNR estimated by restored signal S^{harmo}.

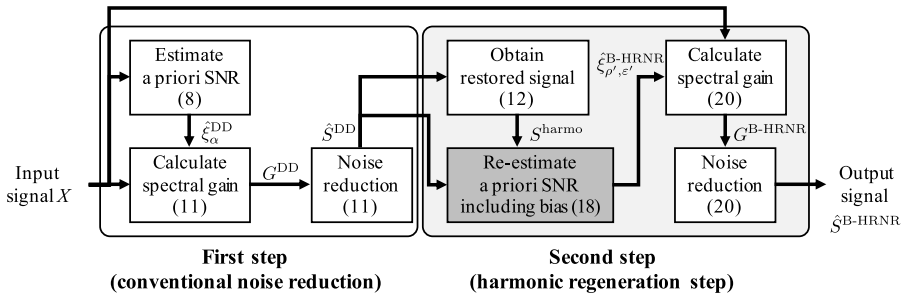

Fig. 2. Block diagram of proposed method. Basic procedure is same as MMSE-STSA estimator with HRNR in Fig. 1. In second step, we re-estimate a priori SNR including bias.

$$\hat{\xi}^{\mathrm{B-HRNR}}_{\rho',\varepsilon'} = \rho' \mathrm{Max}\left[\frac{\left|\widehat{S}\right|^2}{\mathrm{E}\left[\left|\widehat{N}\right|^2\right]}, \varepsilon'\right] + (1-\rho')\frac{\left|S^{\mathrm{harmo}}\right|^2}{\mathrm{E}\left[\left|\widehat{N}\right|^2\right]}, \tag{18}$$

Here, ρ' $(0 \leqslant \rho' \leqslant 1)$ and ε' represent the weighting factor and bias value, respectively. Also, the form which replaces ρ' with spectral gain G is expressed as

$$\hat{\xi}^{\mathrm{B-HRNR}}_{G,\varepsilon'} = G\mathrm{Max}\left[\frac{\left|\widehat{S}\right|^2}{\mathrm{E}\left[\left|\widehat{N}\right|^2\right]}, \varepsilon'\right] + (1-G)\frac{\left|S^{\mathrm{harmo}}\right|^2}{\mathrm{E}\left[\left|\widehat{N}\right|^2\right]}. \tag{19}$$

Finally, the output of the proposed method is obtained as follows:

$$\widehat{S}^{\mathrm{B-HRNR}} = G^{\mathrm{B-HRNR}}X = g^{\mathrm{STSA}}\left(\hat{\xi}^{\mathrm{B-HRNR}}_{\rho',\varepsilon'}, \gamma\right)X. \tag{20}$$

3.2. Discussion on existence of musical-noise-free condition

In Section 2.2, we described the HRNR. We also explained biased HRNR in Section 3.1. The sound qualities obtained using HRNR and biased HRNR are determined by adjusting the internal parameters. However, the detailed relationship between the internal parameters and sound quality has not been revealed. Moreover, the existence of the musical-noise-free condition has not been clarified. The aim of this section is to confirm the existence of this for the MMSE-STSA estimator with HRNR and our MMSE-STSA estimator with biased HRNR by adjusting their internal parameters. We summarize the internal parameters of each method to clarify what we show in Table 1.

Table 1
Summary of relationship among noise reduction methods, a priori SNR estimators, biasing, and spectral gains.

Noise reduction ethod	A priori SNR estimator	Biasing	Spectral gain
MMSE-STSA estimator	$\hat{\xi}^{\mathrm{DD}}_{\alpha}$	NOT Included	$G^{\mathrm{DD}} = g^{\mathrm{STSA}}\left(\hat{\xi}^{\mathrm{DD}}_{\alpha}, \gamma\right)$
MMSE-STSA estimator with HRNR	$\hat{\xi}^{\mathrm{HRNR}}_{\rho}$	NOT Included	$G^{\mathrm{HRNR}} = g^{\mathrm{STSA}}\left(\hat{\xi}^{\mathrm{HRNR}}_{\rho}, \gamma\right)$
Musical-noise-free MMSE-STSA estimator	$\hat{\xi}^{\mathrm{B-DD}}_{\alpha,\varepsilon}$	Included	$G^{\mathrm{B-DD}} = g^{\mathrm{STSA}}\left(\hat{\xi}^{\mathrm{B-DD}}_{\alpha,\varepsilon}, \gamma\right)$
MMSE-STSA estimator with biased HRNR	$\hat{\xi}^{\mathrm{B-HRNR}}_{\rho',\varepsilon'}$	Included	$G^{\mathrm{B-HRNR}} = g^{\mathrm{STSA}}\left(\hat{\xi}^{\mathrm{B-HRNR}}_{\rho',\varepsilon'}, \gamma\right)$

3.2.1. DD approach and HRNR

The DD approach is used to determine sound quality by using the forgetting factor α [33]. It is assumed that the output of the MMSE-STSA estimator with HRNR depends on α since the DD approach is used in the first step of this noise reduction method (see Fig. 1). To determine the relationship between the internal parameters and sound quality, we conducted an experiment by adjusting α in the DD approach and weighting factor ρ in HRNR.

The observed signals were generated by adding babble noise or railway station noise to the target speech from the JNAS database [36] with a 10-dB input SNR. For DD, the α in Eq. (8) was set from 0.00 to 0.99. For HRNR, the α in Eq. (8) was set to 0.5, 0.7, and 0.98 as the first step of Fig. 1, and the ρ in Eq. (13) was set from 0.0 to 1.0 as the second step. We also examined a case in which ρ was replaced with the spectral gain G obtained in the first step, i.e., $\hat{\xi}^{\mathrm{HRNR}}_{G}$ in Eq. (14) was used. We introduced two objective measurements to evaluate the noise reduction level and speech distortion of each output signal in addition to the KR as a measurement of musical noise generation, as mentioned in Section 2.3.

The noise reduction rate (NRR) is a measure of noise reduction level [21], and a higher NRR indicates significant SNR improvement. The NRR is computed as the difference between the input and output signals as follows:

$$\mathrm{NRR} = 10\log_{10}\frac{\mathrm{E}\left[|s_{\mathrm{out}}|^2\right]/\mathrm{E}\left[|n_{\mathrm{out}}|^2\right]}{\mathrm{E}\left[|s_{\mathrm{in}}|^2\right]/\mathrm{E}\left[|n_{\mathrm{in}}|^2\right]}, \tag{21}$$

where s_{in} and s_{out} are the input and output speech signals, and n_{in} and n_{out} are the input and output noise signals, respectively. Cepstral distortion (CD) is used for measuring speech distortion [37]. By using the cepstral coefficients of clean speech C_{ref} and processed speech C_{out}, CD is calculated by

$$\mathrm{CD} = \frac{20}{P\log 10}\sum_{p=0}^{P}\sqrt{\sum_{b=0}^{B}2(C_{\mathrm{out}}(p,b) - C_{\mathrm{ref}}(p,b))^2}, \tag{22}$$

where b is an index of the cepstral coefficient $(0 \leqslant b \leqslant B)$ and B is a dimension of the cepstrum. All signals used in this experiment were sampled at 16 kHz. The Hamming window (512 width and 25 % overlap) was applied in the STFT. We provided the speech-absent period for computing the KR. The B in Eq. (22) was set to 22.

Figs. 3 (a) and (b) show the results of the KR versus NRR. The larger symbols indicate that each parameter is set to zero, i.e., large ☆ shows $\hat{\xi}^{\mathrm{DD}}_{\alpha=0}$ and large ■, ◆, and • represent $\hat{\xi}^{\mathrm{HRNR}}_{\rho=0}$ when α is set to 0.5, 0.7, and 0.98, respectively. The gray areas mean that the KR is less than 1.0 (musical-noise-free condition). In terms of the DD approach (star symbols), NRR increased as α increased. For the constant parameter $\hat{\xi}^{\mathrm{HRNR}}_{\rho}$ (grey and quadrate, rhomboid or round behaviors), the KR depended on α in the first step. Namely, the tendencies of ρ were plotted in the higher NRR area when α was large.

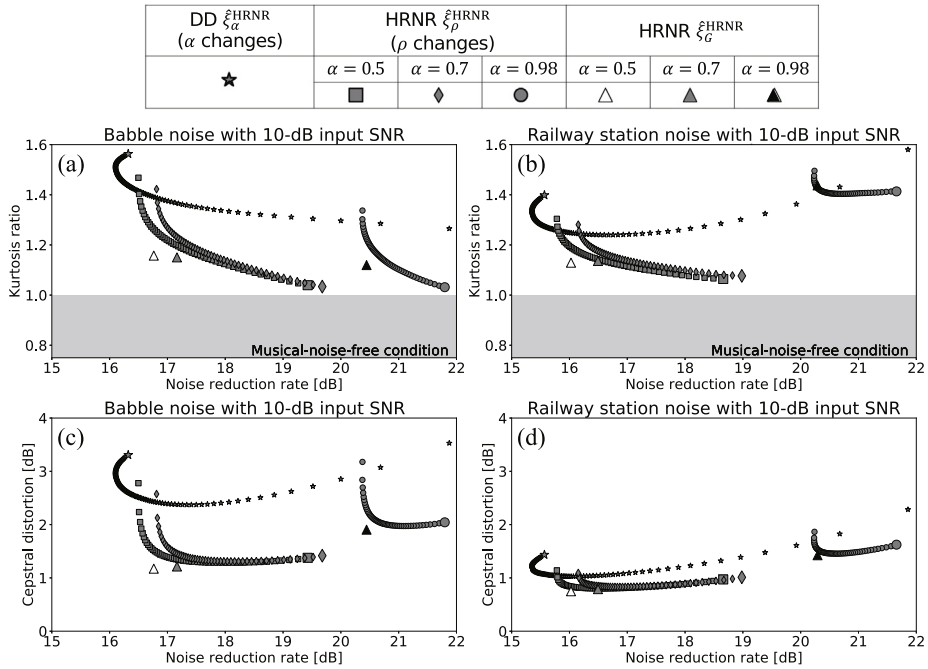

Fig. 3. Results of sound quality by adjusting each internal parameter of DD approach and HRNR. (a) and (b) indicate relationship between NRR and KR, (c) and (d) indicate relationship between NRR and CD. Larger symbols indicate that each parameter is set to zero, i.e., large ☆ shows $\hat{\xi}^{\mathrm{DD}}_{\alpha=0}$ and large ■, ◆, and ● represent $\hat{\xi}^{\mathrm{HRNR}}_{\rho=0}$ when α is set to 0.5, 0.7, and 0.98, respectively. Note that $\hat{\xi}^{\mathrm{HRNR}}_{G}$ has no adjusting parameters.

Moreover, NRR decreased when ρ increased. For $\hat{\xi}^{\mathrm{HRNR}}_{G}$, however, the sound quality was not better than that for $\hat{\xi}^{\mathrm{HRNR}}_{\rho}$in which the parameter is small in terms of the NRR and KR. However, no plots are in the gray areas in this figure. Although HRNR is more effective in terms of musical noise generation than the DD approach, it is limited in meeting the musical-noise-free condition.

Figs. 3 (c) and (d) show the results of CD versus NRR. CD correspondingly increased when α increased with the DD approach. The tendencies of ρ depended on α as with the results of the KR versus NRR. CD decreased in the part of the low ρ and rapidly increased from a certain point. Since this tendency was found under another condition, we can consider the existence of the optimal value. However, a point of $\hat{\xi}^{\mathrm{HRNR}}_{G}$ is marked in the lowest CD of any constant parameter. Hence, we conclude that using spectral gain in the a priori SNR estimator of HRNR is best to keep speech distortion low.

3.2.2. HRNR and biased HRNR

We described the effectiveness of HRNR in Section 3.2.1. Next, we discuss the relation between the sound quality and internal parameters from HRNR and biased HRNR. Biased HRNR has two internal parameters (ρ' and ε'). The ρ' was fixed to 0.1 or 0.9 and we increased ε' from 0.0 to 1.0. For HRNR and biased HRNR, the α in the first step was set to 0.5 and 0.98. Other conditions were the same as those mentioned in Section 3.2.1.

Figs. 4 (a) and (b) show the behavior of the KR by adjusting the internal parameters. Larger symbols indicate that each parameter is set to zero, i.e., larger ■ and ● are $\hat{\xi}^{\mathrm{HRNR}}_{\rho=0}$ with $\alpha = 0.5$ and $\alpha = 0.98$, larger ■ and □ indicate $\hat{\xi}^{\mathrm{B-HRNR}}_{\rho'=0.1,\varepsilon'=0}$ and $\hat{\xi}^{\mathrm{B-HRNR}}_{\rho'=0.9,\varepsilon'=0}$ with $\alpha = 0.5$, larger ● and ○ represent $\hat{\xi}^{\mathrm{B-HRNR}}_{\rho'=0.1,\varepsilon'=0}$ and $\hat{\xi}^{\mathrm{B-HRNR}}_{\rho'=0.9,\varepsilon'=0}$ with $\alpha = 0.9$, and △ and ▲ indicate the scores of $\hat{\xi}^{\mathrm{B-HRNR}}_{G,\varepsilon'=0}$ with $\alpha = 0.5$ and $\alpha = 0.98$, respectively. The gray areas mean that the KR is less than 1.0 (musical-noise-free condition). The NRR and KR decrease with the increase in ε' (see black and white symbols). Biased HRNR suppresses musical noise generation under the same NRR condition. In the constant parameter case $\hat{\xi}^{\mathrm{B-HRNR}}_{\rho',\varepsilon'}$ of biased HRNR (black or white round and quadrate symbols), some symbols reach a KR less than 1.0 not depending on ρ'. Furthermore, a lower ρ' is effective because a higher NRR is better. On the other hand, the KR of $\hat{\xi}^{\mathrm{B-HRNR}}_{G,\varepsilon'}$ with $\alpha = 0.5$ (△) does not meet the musical-noise-free condition (KR < 1.0 area). Therefore, in order for the output of $\hat{\xi}^{\mathrm{B-HRNR}}_{G,\varepsilon'}$ to meet the musical-noise-free condition, it is necessary to increase the noise reduction level in the first step.

Figs. 4 (c) and (d) show the tendency by changing each parameter in CD versus NRR. The NRR and CD decrease by introducing ε' and biased HRNR suppresses speech distortion compared with HRNR. The $\hat{\xi}^{\mathrm{B-HRNR}}_{\rho',\varepsilon'}$ achieves the lower CD than $\hat{\xi}^{\mathrm{B-HRNR}}_{G,\varepsilon'}$ under the same NRR condition (e.g., in Fig. 4 (c) at NRR = 20, the black circle is plotted under the black triangle symbol). As a result, biased HRNR is a high-quality noise reduction method compared to HRNR and the DD approach, and a lower constant ρ' is effective.

3.3. Discussion on introducing bias

We described the behavior of the internal parameters for each a priori SNR estimator in the previous section. The results indicate that biasing contributes to decreasing KR and enables biased HRNR to meet the musical-noise-free condition. In this section, we focus on the effectiveness of biasing and discuss its contribution in terms of the estimation accuracy of a priori SNR. First, we conducted a quantitative evaluation of estimation accuracy. The observed signals were mixed babble noise, railway station noise, street noise, and white Gaussian noise with clean speech at 0- and 10-dB input SNRs. The log-error (LogErr) was used for measuring the amount of estimation error. LogErr is calculated as the sum of the amounts of underestimation $\mathrm{LogErr}_{\mathrm{un}}$ and overestimation $\mathrm{LogErr}_{\mathrm{ov}}$, i.e.,

$$\mathrm{LogErr} = \mathrm{LogErr}_{\mathrm{un}} + \mathrm{LogErr}_{\mathrm{ov}}, \tag{23}$$

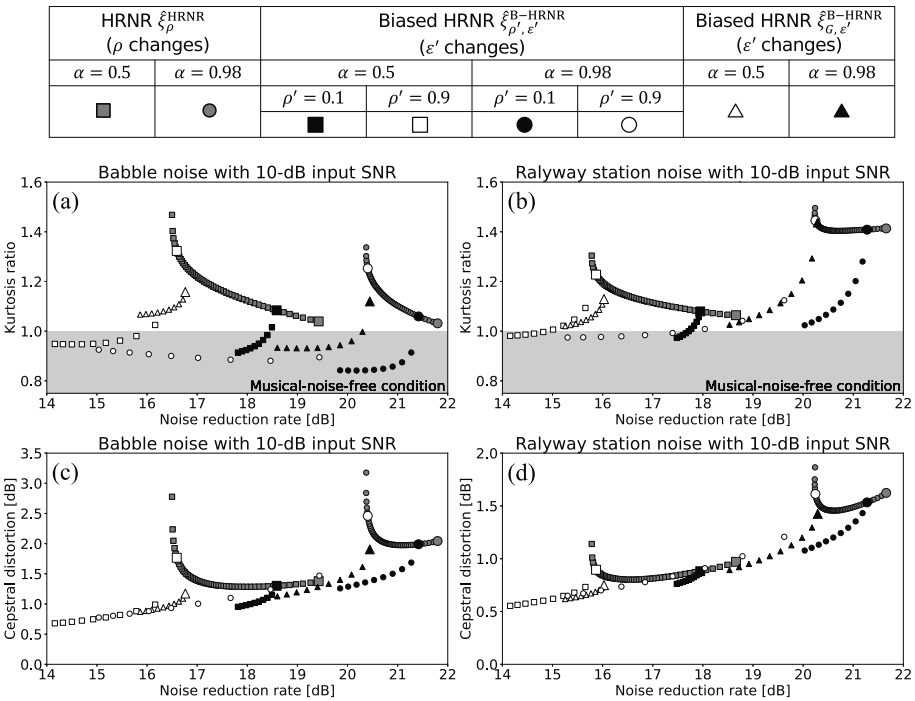

Fig. 4. Results of sound quality by adjusting each internal parameter of HRNR and biased HRNR. (a) and (b) indicate relationship between NRR and KR, (c) and (d) indicate relationship between NRR and CD. Larger symbols indicate that each parameter is set to zero, i.e., larger ■ and ● are $\hat{\xi}_{\rho=0}^{\mathrm{HRNR}}$ with $\alpha = 0.5$ and $\alpha = 0.98$, larger ■ and □ indicate $\hat{\xi}_{\rho'=0.1,\varepsilon'=0}^{\mathrm{B-HRNR}}$ and $\hat{\xi}_{\rho'=0.9,\varepsilon'=0}^{\mathrm{B-HRNR}}$ with $\alpha = 0.5$, larger ● and ○ represent $\hat{\xi}_{\rho'=0.1,\varepsilon'=0}^{\mathrm{B-HRNR}}$ and $\hat{\xi}_{\rho'=0.9,\varepsilon'=0}^{\mathrm{B-HRNR}}$ with $\alpha = 0.9$, and △ and ▲ indicate the scores of $\hat{\xi}_{G,\varepsilon'=0}^{\mathrm{B-HRNR}}$ with $\alpha = 0.5$ and $\alpha = 0.98$, respectively.

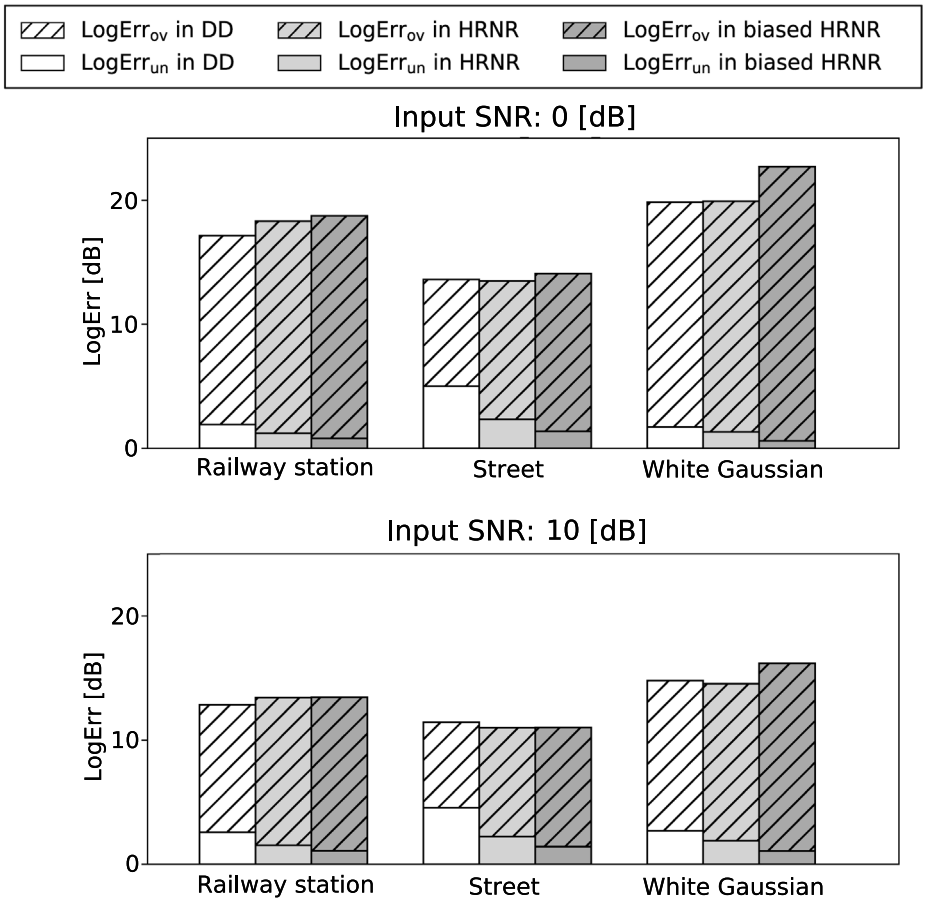

Fig. 5. Results of log-error in 0-dB input SNR (upper) and 10-dB input SNR (lower).

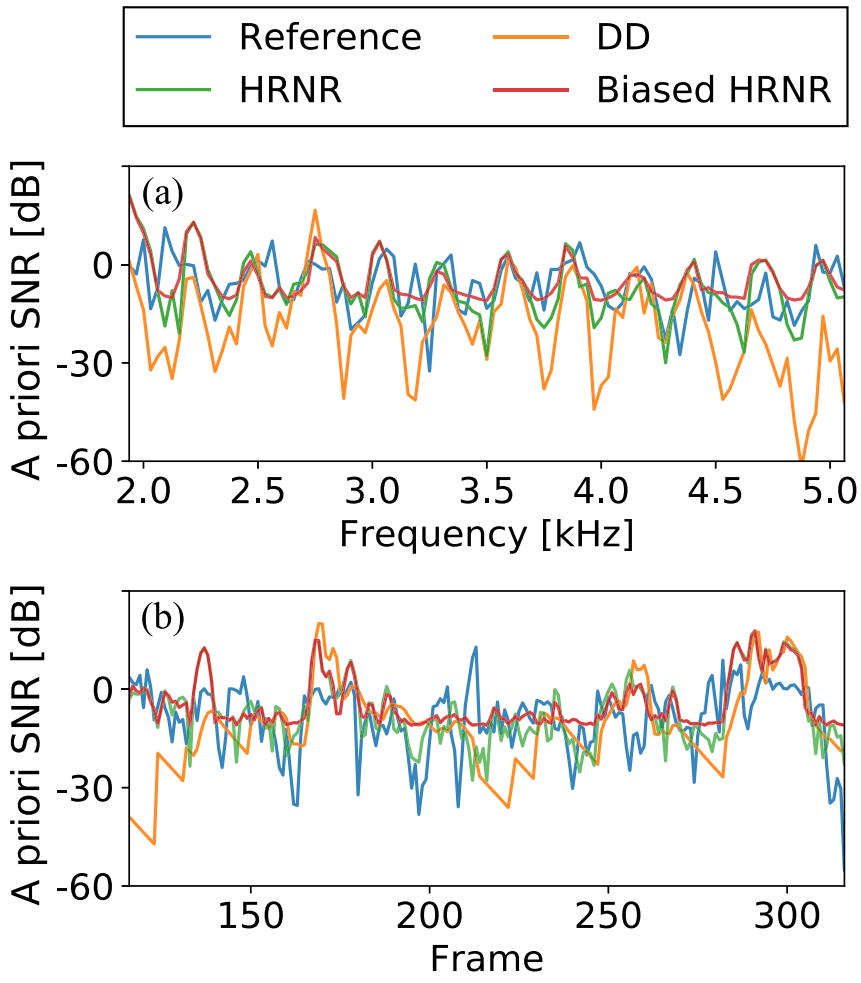

Fig. 6. Comparison of true and estimated a priori SNRs under speech presence interval in (a) frequency bin and (b) frame.

$$\mathrm{LogErr}_{\mathrm{un}} = \frac{10}{PK}\Sigma_{p,k}\mathrm{Max}\left[\log_{10}\frac{\xi}{\hat{\xi}},0\right], \tag{24}$$

$$\mathrm{LogErr}_{\mathrm{un}} = \frac{10}{PK}\Sigma_{p,k}\mathrm{Max}\left[\log_{10}\frac{\xi}{\hat{\xi}},0\right], \tag{25}$$

where $\mathrm{Min}[a,b]$ returns a smaller value, and the true a priori SNR ξ was computed using Eq. (5) assuming that clean and noise signals were known. We obtained a priori SNRs estimated using the DD approach, HRNR, and biased HRNR, and computed each LogErr with respect to the true one. Each parameter was set to achieve the same NRR.

Fig. 5 shows the results of LogErr in various noisy environments. Many large underestimations occurred with the DD approach. HRNR suppressed the more underestimation compared to the DD approach; however, HRNR overestimated the a priori SNR more. Although biased HRNR increased the total error compared to the other methods, it kept the amount of underestimation low. We argue that this is due to introducing bias.

Next, we observed the actual behavior of a priori SNR to confirm the effect of biasing. The observed signal was made by mixing babble noise with the clean speech at a 10-dB input SNR. To achieve 20-dB NRR for each method, the α of the DD approach was set to 0.97, ρ of HRNR was set to 0.04, and ρ' and ε' of biased HRNR were set to 0.1 and 0.8, respectively. Other experimental conditions were the same as those mentioned in Section 3.2.1.

Fig. 6 shows the results of the estimated a priori SNRs of these methods under speech presence interval. Figs. 6 (a) and (b) represent each estimated a priori SNR in a frame and frequency bin, respectively. First, we compared the true a priori SNR and that with the DD approach. The behavior with the DD approach tracks the true a priori SNR in the lower position (i.e., underestimation). In particular, the underestimation is outstanding at high frequency (around 4 kHz), as shown in Fig. 6 (a). The underestimation leads to speech distortion and harmonic distortion in the high-frequency part. Additionally, the transition in the behavior with the DD approach is slow. This phenomenon is caused by smoothing down using the previous frame at the estimation, and the delay occurs in an area from speech absent to speech presence and produces the underestimation. Therefore, we confirm that the a priori SNR estimator of the DD approach triggers underestimation, leading to speech distortion. Next, we compared HRNR with the DD approach. From Figs. 6 (a) and (b), HRNR estimated the a priori SNR highly compared with DD approach. A restored signal S^{harmo} regenerates the pseudo spectrum of the harmonics. Namely, the protrusions with the DD approach and HRNR mostly synchronize and we expect the improvement of speech distortion from to this effect. In biased HRNR, bias smoothens the a priori SNR (see the period from the 200 th frame to 240 th frame in Fig. 6 (b)). This smoothing contributes to reducing musical noise generation which is perceived by isolated components in the processed power spectrogram. Therefore, biased HRNR can suppress speech distortion and musical noise generation.

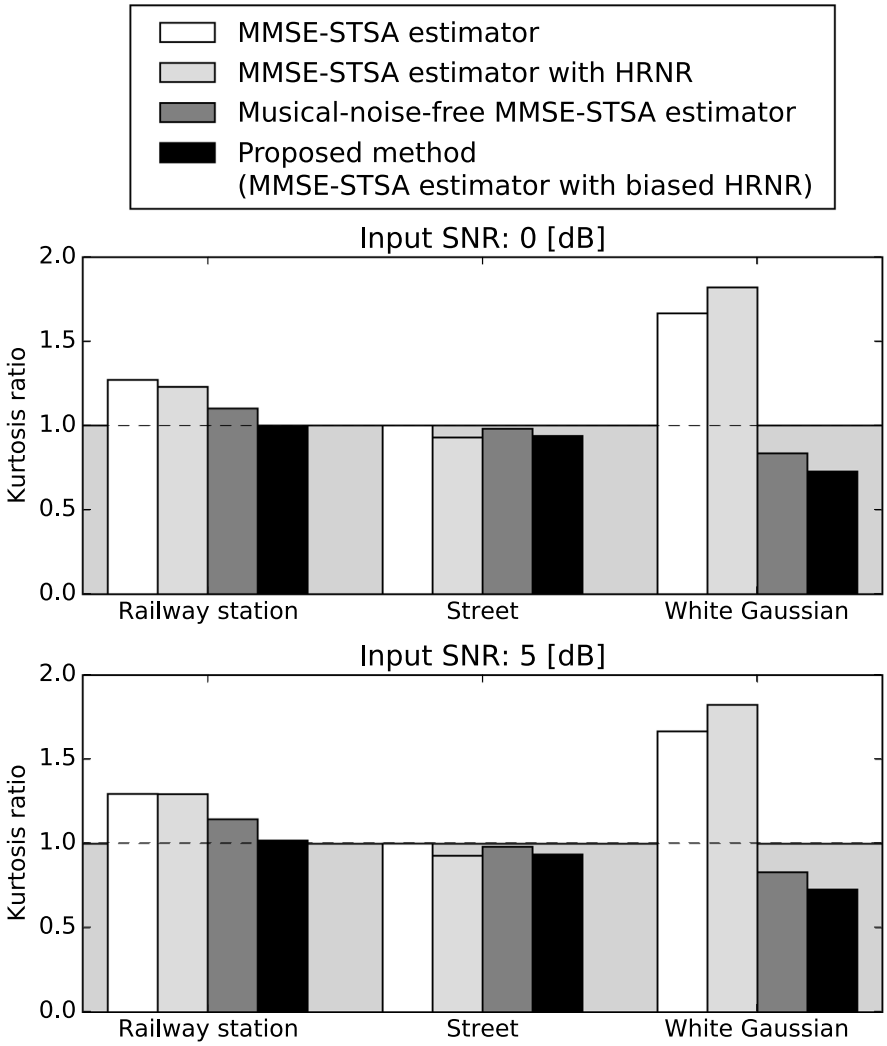

Fig. 7. KR obtained at 0-dB (upper) and 5-dB (lower) input SNRs. Filled areas denote musical-noise-free condition.

Naturally, if a correct a priori SNR is used in Eq. (11), we can obtain significantly high-quality output signals. An a priori SNR close to a true one is intuitively better. However, we indicated that introducing bias improves speech quality despite the increase in estimation error. The biasing flattens out the power spectrum in a noise-dominant area after processing and suppresses musical noise generation. Therefore, we argue the biasing contributes to suppressing the amounts of speech distortion and musical noise generation.

4. Experimental evaluation

4.1. Experimental conditions

To validate the effectiveness of biased HRNR, we conducted a comparative experiment involving three conventional noise reduction methods: MMSE-STSA estimator, MMSE-STSA estimator with HRNR, and musical-noise-free MMSE-STSA estimator. The objective scores were KR and CD.

We used ten sentences (five for male speech and five for female speech) as the target speech signals, which were mixed with three types of noise (railway station noise, street noise, and white Gaussian noise) at 0- and 5-dB input SNRs. The gain function of the MMSE-STSA estimator with HRNR in Eq. (15) was set as the MMSE-STSA estimator. It is best to set ρ' to a constant value for biased HRNR, as mentioned in Section 3.2.2. We used Eq. (18) as the a priori SNR formula in this experiment. To achieve 10-dB NRR for each noise reduction method, we manually controlled the internal parameters of each method. Other conditions, e.g., sampling rate, window size, and shift length were the same as those mentioned in Section 3.2.1. Note that we made the NRRs of the three conventional methods and our MMSE-STSA estimator with biased HRNR even (i.e., we did not set their parameters in order for these KRs to be 1 or less); therefore, two of the musical-noise-free methods we mentioned do not necessarily meet the musical-noise-free condition.

4.2. Results

The objective evaluation results obtained for musical noise generation and speech distortion are shown in Figs. 7 and 8, respectively. In both figures, the input SNR in the upper part was 0 dB and that in the lower part was 5 dB. First, from Fig. 7, the musical-noise-free MMSE-STSA estimator and MMSE-STSA estimator with biased HRNR met the musical-noise-free condition for most noise cases. Although the KR of our MMSE-STSA estimator with biased HRNR was slightly more than 1.0 for railway station noise, it generated little musical noise compared with the other methods. This indicates that it is effective in terms of musical noise generation. Next, Fig. 8 indicates that our MMSE-STSA estimator

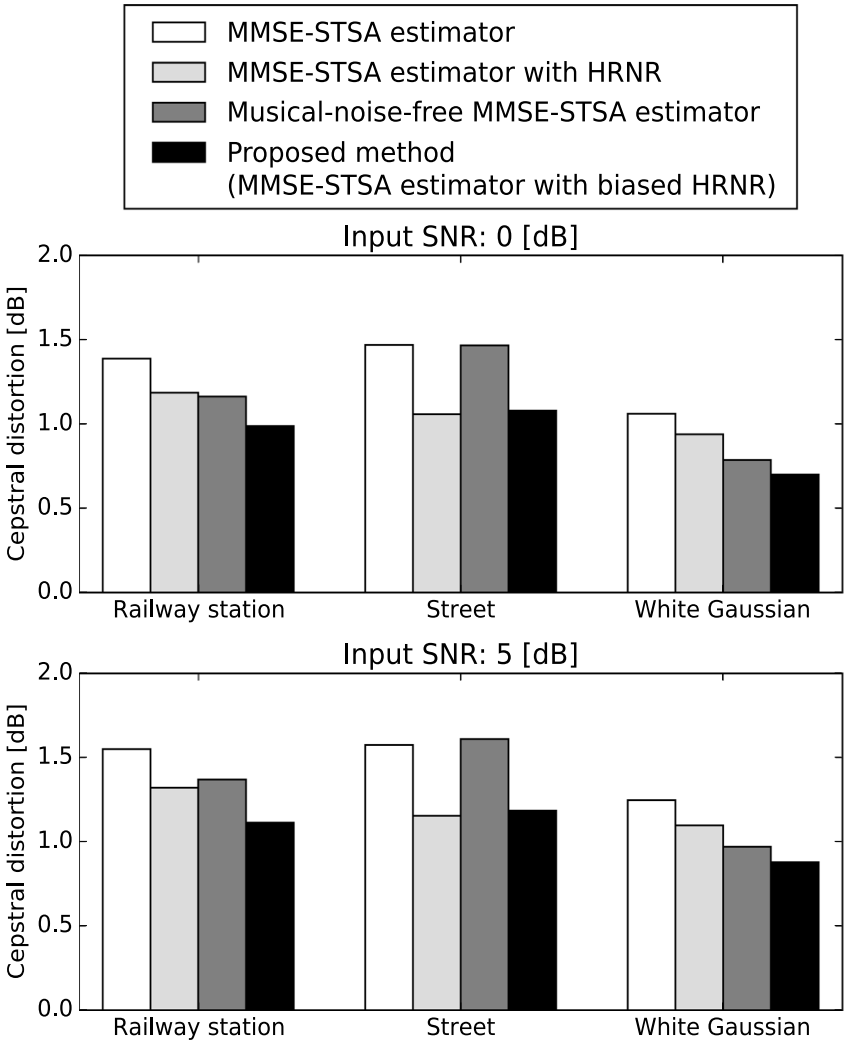

Fig. 8. CD at 0-dB (upper) and 5-dB (lower) input SNRs.

with biased HRNR achieved the lowest CD in all cases. Therefore, we can confirm that MMSE-STSA estimator with biased HRNR is a better high-quality noise reduction method compared with the other methods under the same NRR conditions.

5. Conclusion

We proposed a noise reduction method for generating no musical noise with low speech distortion by applying biased HRNR. The contribution of this paper is suppressing lower speech distortion under the musical-noise-free condition with the introduction of bias. In Section 3, we confirmed that HRNR is effective compared to the DD approach. The internal parameters are best to set small and constant in terms of the musical-noise-generation problem. We also mentioned that HRNR is limited in meeting the musical-noise-free condition. In contrast, speech distortion is low when the internal parameters are set to the spectral gain obtained in the first step of HRNR. Next, we showed that the proposed method is a higher-quality noise reduction method than HRNR, and meets the musical-noise-free condition in all cases. We also confirmed that introducing a bias suppresses musical noise generation, and speech distortion with biased HRNR and weighting parameter is more effective as a small constant value. Finally, we conducted a comparative experiment (Section 4), and the results indicate that our MMSE-STSA estimator with biased HRNR is superior to conventional methods in terms of both suppressing musical noise generation and speech distortion.

CRediT authorship contribution statement

Masakazu Une: Investigation, Methodology, Writing - original draft. **Ryoichi Miyazaki:** Writing - original draft, Writing - review & editing.

Acknowledgments

This project was partially supported by Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Young Scientists (B).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apacoust.2020.107410.

References

[1] Flanagan RZJL, Johnston JD, Elko GW. Computer-steered microphone arrays for sound transduction in large rooms. J Acoust Soc Am 1985;78(5):1508–18.

[2] Saruwatari H, Kurita S, Takeda K, Itakura F, Nishikawa T. Blind source separation combining independent component analysis and beamforming. EURASIP J Appl Signal Process 2003;2003(11):1135–46.
[3] Boll SF. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust, Speech Signal Process 1979;27(2):113–20.
[4] Wiener N. Extrapolation, interpolation and smoothing of stationary time series with engineering applications. Cambridge, MA: MIT Press; 1949.
[5] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans Acoust, Speech Signal Process 1984;27(6):1109–21.
[6] Yamashita K, Ogata S, Shimamura T. Spectral subtraction iterated with weighting factors. Proceedings of IEEE Speech Coding Workshop. p. 138–40.
[7] You CH, Koh SN, Rahardja S. β-order MMSE spectral amplitude estimation for speech enhancement. IEEE Trans Acoust, Speech Signal Process 2005;13(5):475–86.
[8] Benesty J, Huang Y. A single-channel noise reduction MVDR filter. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing. p. 273–6.
[9] Loizou PC. Speech enhancement theory and practice. FL: CRC Press, Taylor & Francis Group; 2007.
[10] Cappe O. Elimination of the musical noise phenomenon with the Ephraim andMalah noise suppressor. IEEE Trans Speech Audio Process 1994;2(2):345–9.
[11] Goh Z, Tan KC, Tan B. Postprocessing method for suppressing musical noise280generated by spectral subtraction. IEEE Trans Speech Audio Process 1998;6(3):287–92.
[12] Elshamy S, Madhu N, Tirry W, Fingscheidt T. Instantaneous a priori snr estimation by cepstral excitation manipulation. IEEE/ACM Trans Audio, Speech, Language Process 2017;25(8):1592–605.
[13] Erkelens JS, Jensen J, Heusdens R. A data-driven approach to optimizing spectral speech enhancement methods for various error criteria. Speech Comm 2007;49(7):530–41.
[14] Erkelens JS, Jensen J, Heusdens R. Improved speech spectral variance estimation under the generalized gamma distribution. IEEE BENELUX/DSP Valley Signal Process. Symp., Antwerp, Belgium. p. 43–6.
[15] Plapous C, Marro C, Scalart P. Speech enhancement using harmonic regeneration. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, vol. 1. p. 157–60.
[16] Plapous C, Marro C, Scalart P. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans Audio, Speech, Language Process 2006;14(6):2098–108.
[17] Une M, Miyazaki R. Evaluation of sound quality and speech recognition performance using harmonic regeneration for various noise reduction techniques, in:2017 RISP International Workshop on Nonlinear Circuits. Commun Signal Process 2017:377–80.
[18] Une M, Miyazaki R. Musical-noise-free speech enhancement with low speech distortion by biased harmonic regeneration technique. In: Proceedings of International Workshop on Acoustic Signal Enhancement. p. 31–5.
[19] Vihari S, Murthy AS, Soni P, Naik DC. Comparison of speech enhancement305algorithms. Procedia Computer Sci 2016;89:666–76.
[20] Hu Y. A simulation study of harmonics regeneration in noise reduction for electric and acoustic stimulation. J Acoust Soc Am 2010;127:3145–53.
[21] Miyazaki R, Saruwatari H, Inoue T, Takahashi Y, Shikano K, Kondo K. Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. IEEE Trans Audio Speech Lang Process 2012;20(7):2080–94.
[22] Nakai S, Saruwatari H, Miyazaki R, Nakamura S, Kondo K. 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 2014:122–6.
[23] Saruwatari H. Statistical-model-based speech enhancement with musical-noise-free properties. In: Proceedings of International Conference on Digital Signal Processing. p. 1201–5.
[24] Mukai R, Araki S, Sawada H, Makino S. 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. p. 1789–92.
[25] Takahashi Y, Takatani T, Saruwatari H, Shikano K. Blind spatial subtraction array with independent component analysis for hands-free speech recognition. In: Proceedings of International Workshop on Acoustic Echo and Noise Control.
[26] Hansan MK, Salahuddin S, Khan MR. A modified a priori SNR for speech enhancement using spectral subtraction rules. IEEE Signal Process Lett 2004;33011(4):450–3.
[27] Suhadi S, Last C, Fingscheidt T. A data-driven approach to a priori SNR estimation. IEEE Trans Audio Speech Lang Process 2011;19(1):186–95.
[28] Yun-Sik P, Chang JH. A novel approach to a robust a priori SNR estimator in speech enhancement. IEICE Trans Acoust Speech Signal Process 2007;90(8):2182–5.
[29] Yong PC, Nordholm S, Dam HH. Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement. Speech Commun 2013;55(2):358–76.
[30] Nahma L, Yong PC, Dam HH, Nordholm S. Improved a priori SNR estimation in speech enhancement. In: 23rd Asia-Pacific Conference on Communications. 1–5.24..
[31] Uemura Y, Takahashi Y, Saruwatari H, Shikano K, Kondo K. Automaticoptimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics. In: Proceedings of International Workshop on AcousticEcho and Noise Control.
[32] Miyazaki R, Saruwatari H, Nakamura S, Shikano K, Kondo K, Blanchette J, Bouchard M. Musical-noise-free blind speech extraction integrating microphone array and iterative spectral subtraction. Signal Process 2014;102:226–39.
[33] Kanehara S, Saruwatari H, Miyazaki R, Shikano K, Kondo K. Theoretical analysis of musical noise generation in noise reduction methods with decision-directed a priori SNR estimator. In: Proceedings of International Workshop on Acoustic Signal Enhancement.
[34] Kanehara S, Saruwatari H, Miyazaki R, Shikano K, Kondo K. Comparative study on various noise reduction methods with decision-directed a priori SNRestimator via higher-order statistics. In: Proceedings of APSIPA Annual Summit and Conference 2012.
[35] Saruwatari H, Kanehara S, Miyazaki R, Shikano K, Kondo K. Musical noise analysis for bayesian minimum mean-square error speech amplitude estimators360based on higher-order statistics. In: Proceedings of INTERSPEECH2013. p. 441–5.
[36] Ito K, Yamamoto M, Takeda K, Takezawa T, Matsuoka T, Kobayashi T, Shikano K, Itahashi S. JNAS Japanese speech corpus for large vocabulary continuous speech recognition research. J Acoust Soc Japan 1999;20:196–206.
[37] Rabiner L, Juang B. Fundamentals of speech recognition. Upper Saddle River, NJ: Prentice-Hall; 1993.