MUSICAL-NOISE-FREE SPEECH ENHANCEMENT WITH LOW SPEECH DISTORTION BY BIASED HARMONIC REGENERATION TECHNIQUE

Masakazu Une[†] *and Ryoichi Miyazaki*[†]

[†] National Institute of Technology, Tokuyama College, Gakuendai, Shunan, Yamaguchi 745–8585, Japan

ABSTRACT

Many noise reduction techniques generate two frequently occurring problems: speech distortion and musical noise. A number of methods have been proposed to solve these problems. One, which addresses the former problem, is harmonic regeneration noise reduction (HRNR). Using restored signals, HRNR regenerates the harmonics that were eliminated. Another, which addresses the latter problem, is a musical-noise-free noise reduction method based on the minimum mean-square error short-time spectral amplitude estimator. The method can suppress noisy speech without generating musical noise. This paper describes a new noise reduction technique we propose that combines these two methods. The technique is shown to be effective in suppressing both speech distortion and musical noise generation.

Index Terms— MMSE-STSA estimator, musical noise, harmonic regeneration, musical-noise-free speech enhancement

1. INTRODUCTION

Many speech communication devices have come into use in recent years. When using them, however, not only human voices but also environmental noises are input at the same time. Noise reduction techniques are necessary to extract the speech component with high accuracy in a noisy environment [1–7]. However, many noise reduction methods generate two severe problems, i.e., speech distortion and musical noise, owing nonlinear signal processing [8–12].

To date many noise reduction methods have been proposed to overcome these problems. The harmonic regeneration noise reduction (HRNR) technique addresses the speech distortion problem [13]. Speech components contain many harmonics that are eliminated by the noise reduction; speech distortion can be argued as due to this elimination of harmonics. The HRNR technique restores the harmonic components that are eliminated by noise suppression. Also, HRNR is applied after noise reduction techniques have been applied and calculates restored signal from the signals suppressed as a consequence of noise reduction. These restored signals have harmonic components based on the suppressed signals.

Part of this work was supported by JSPS KAKENHI Grant Number JP16K21579.

Other methods that suppress noisy speech without generating musical noise have also been proposed [14–16]. These techniques, collectively called *musical-noise-free speech enhancement* methods, are useful in systems such as telecommunication systems, hearing aid systems, and video conference systems [17, 18]. In particular, it is known that a musical-noise-free noise reduction technique based on a minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator (hereafter referred to as a musical-noise-free MMSE-STSA estimator) achieves lower speech distortion than other musical-noise-free noise reduction techniques [15]. By introducing the bias factor into the a priori SNR estimator, the musical-noise-free MMSE-STSA estimator suppresses musical noise generation.

In this paper, we propose a new noise reduction method which applies biased a priori SNR estimation to HRNR. We experimentally demonstrate the tendency of musical noise generation by introducing a bias. The results of the comparative experiments are given in terms of speech distortion and musical noise generation.

2. RELATED WORK

2.1. Signal definition

The observed signal in time-frequency domain $X(p,k)$ consists of the speech signal $S(p,k)$ and the noise signal $N(p,k)$. This is expressed as $X(p,k) = S(p,k) + N(p,k)$, where p is the short-time frame index and k is the frequency bin. The spectral gain $G(p,k) = g(\xi(p,k), \gamma(p,k))$ of each noise reduction method is obtained to estimate the spectrum of the speech signal as $\hat{S}(p,k) = G(p,k)X(p,k)$. The spectral gain is expressed as a function of a priori SNR $\xi(p,k) = \mathrm{E}[|S(p,k)|^2]/\mathrm{E}[|N(p,k)|^2]$ and a posteriori SNR $\gamma(p,k) = |X(p,k)|^2/\mathrm{E}[|N(p,k)|^2]$. Where $\mathrm{E}[\cdot]$ is the expectation operator.

2.2. MMSE-STSA estimator

The MMSE-STSA estimator minimizes the mean-square error between the amplitude spectra of the original and the estimated speech [4]. For this estimator it is necessary to obtain a priori SNR and a posteriori SNR for the spectral

978-1-5386-8151-0/18/$31.00

gain $G_{\mathrm{STSA}}(p,k)$ as

$$G_{\mathrm{STSA}}(p,k) = \frac{\sqrt{\nu(p,k)}}{\gamma(p,k)}\Gamma\left(\frac{3}{2}\right) M\left(-\frac{1}{2};1;-\nu(p,k)\right),$$
$$\nu(p,k) = \frac{\xi(p,k)}{1+\xi(p,k)}\gamma(p,k), \quad (1)$$

where $\Gamma(\cdot)$ and $M(a;b;z)$ are respectively the gamma function and the confluent hypergeometric function.

However, a priori SNR $\xi(p,k)$ contains a spectrum of clean speech $\xi(p,k)$ that is not available in actual environments. Therefore, by using a decision-directed approach [4] we estimate the a priori SNR as

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

where $\hat{N}(p,k)$ is the estimated noise signal and α is the forgetting factor. Generally, the forgetting factor α is set to 0.98 to obtain good sound quality [4]. Also, $\mathrm{Max}\left[a,b\right]$ selects the larger of a and b. Finally, the output signal of the MMSE-STSA estimator is obtained as

$$\hat{S}_{\mathrm{STSA}}(p,k) = G_{\mathrm{STSA}}(p,k)X(p,k)$$
$$= g(\hat{\xi}_{\mathrm{DD}}(p,k),\gamma(p,k))X(p,k). \quad (3)$$

2.3. HRNR

Generally, excessive noise suppression leads to serious speech distortion. The HRNR technique has been proposed as a means to restore the harmonics that were eliminated due to noise reduction [13]. The HRNR process can be mainly divided into two steps. In the first, various noise reduction methods are applied to noisy speech and the speech spectrum is estimated. In the second, the HRNR of a priori SNR $\hat{\xi}_{\mathrm{HRNR}}(p,k)$ is computed as follows:

$$\hat{\xi}_{\mathrm{HRNR}}(p,k) = \frac{\rho|\hat{S}(p,k)|^2 + (1-\rho)|S_{\mathrm{harmo}}(p,k)|^2}{\mathrm{E}[|\hat{N}(p,k)|^2]}, \quad (4)$$

where ρ is used to adjust the mixing level of $|\hat{S}(p,k)|$ and $|S_{\mathrm{harmo}}(p,k)|$. The spectrum of the restored signal $S_{\mathrm{harmo}}(p,k)$ is obtained by

$$S_{\mathrm{harmo}}(p,k) = \mathrm{FT}\left[\mathrm{Max}\left(\mathrm{IFT}[\hat{S}(p,k)],0\right)\right], \quad (5)$$

where $\mathrm{FT}\left[\cdot\right]$ and $\mathrm{IFT}\left[\cdot\right]$ respectively represent the Fourier and the inverse Fourier transforms. Finally, the regenerated harmonics signal $\hat{S}_{\mathrm{HRNR}}(p,k)$ is computed using the HRNR of spectral gain $G_{\mathrm{HRNR}}(p,k)$, which is obtained by the new a priori SNR in Eq. (4) and expressed by

$$\hat{S}_{\mathrm{HRNR}}(p,k) = G_{\mathrm{HRNR}}(p,k)X(p,k)$$
$$= g(\hat{\xi}_{\mathrm{HRNR}}(p,k),\gamma(p,k))X(p.k). \quad (6)$$

2.4. Musical-noise-free MMSE-STSA estimator

The kurtosis ratio (KR) has been proposed [19] as an objective measure of musical noise generation. The KR is defined by $\mathrm{kurt}_{\mathrm{proc}}/\mathrm{kurt}_{\mathrm{org}}$, where $\mathrm{kurt}_{\mathrm{proc}}$ and $\mathrm{kurt}_{\mathrm{org}}$ are the kurtosis of the processed and observed signals, respectively. Musical noise is perceived as the skirt in terms of probability density function in the power spectral domain, and kurtosis represents the shape (or skirt) of the distribution. The increase of kurtosis by the signal processing means the generation of musical noise. Namely, KR evaluates the amount of musical noise. The small KR (> 1) indicates the few generations of musical noise and KR (≤ 1) indicates no generation of that (*musical-noise-free condition*). The musical-noise-free speech enhancement method means that almost no musical noise is generated even with high noise reduction. The musical-noise-free theorem was originally applied to spectral subtraction [3, 14]. In addition, Kanehara et al. have 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 parameter [20]. However, Nakai et al. discovered the existence of a musical-noise-free condition in the MMSE-STSA estimator by introducing biased a priori SNR [15]. The biased a priori SNR $\hat{\xi}_{\mathrm{bias}}$ is computed to provide the bias factor in the term of maximum likelihood estimation in Eq. (2) and given by

$$\hat{\xi}_{\mathrm{bias}} = \alpha\frac{|\hat{S}(p-1,k)|^2}{\mathrm{E}[|\hat{N}(p,k)|^2]} + (1-\alpha)\mathrm{Max}[\gamma(p,k)-1,\varepsilon], \quad (7)$$

where ε is the bias value.

3. PROPOSED METHOD

3.1. Amounts of noise reduction and musical noise generation in HRNR

In this section we describe how we experimentally investigated the relationship between the amounts of noise reduction and musical noise generation obtained with HRNR. The experiment was conducted to determine whether a musical-noise-free condition exists in the HRNR. In this experiment, the internal parameter ρ for HRNR in Eq. (4) was set from 0.0 to 1.0. The target speech signal was generated by adding railway station noise with 0 dB SNR.

Figure 1 shows the behavior of the noise reduction rate (NRR) defined as the difference between output and input SNRs [14] and KR when the internal parameter ρ is increased. From Fig. 1, ρ increases as NRR decreases. It also confirms that a musical-noise-free condition does not exist in HRNR for any internal parameter values in this case. Although we investigated the tendency obtained from the same experiments in other noise condition, it does not almost achieve the musical-noise-free condition.

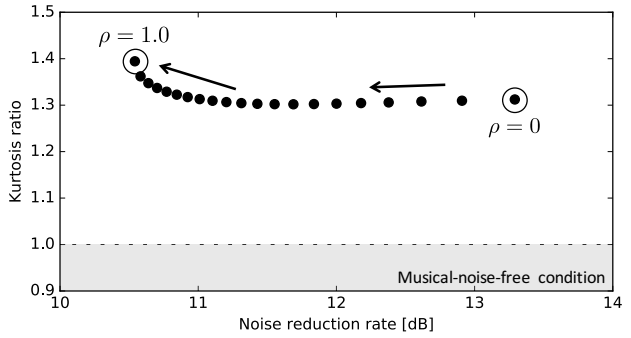

Fig. 1. Relation between NRR and KR in HRNR when internal parameter ρ is increased.

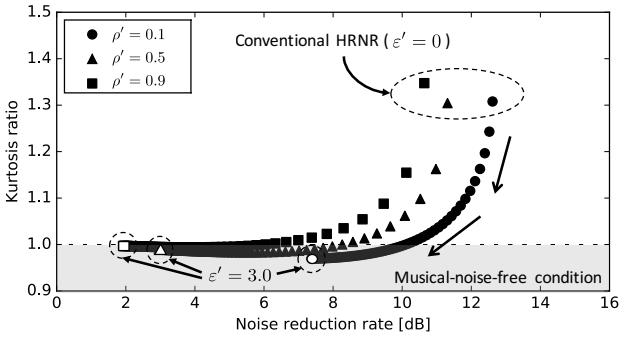

Fig. 2. Relation between NRR and KR in proposed method when internal parameter ε' is increased while internal parameter ρ' remains fixed.

3.2. Biased HRNR

It is known that the amount of musical noise decreases by setting a bias [8]. We therefore introduce biased a priori SNR into HRNR (*biased HRNR*). Eq. (4) is rewritten as

$$\hat{\xi}_{\mathrm{prop}}(p,k) = \rho' \mathrm{Max}\left[\frac{|\hat{S}(p,k)|^2}{\mathrm{E}[|\hat{N}(p,k)|^2]}, \varepsilon'\right] + (1-\rho')\frac{|S_{\mathrm{harmo}}(p,k)|^2}{\mathrm{E}[|\hat{N}(p,k)|^2]}, \quad (8)$$

where $\hat{\xi}_{\mathrm{prop}}(p,k)$ is the biased a priori SNR, ε' is the bias value, and ρ' is the internal parameter ($0 \leq \rho' \leq 1$). The biased HRNR process is the same as the HRNR process. Namely, Eq. (8) corresponds to the second step in HRNR. We finally obtain the estimated speech spectrum using $\hat{\xi}_{\mathrm{prop}}(p,k)$ as

$$\hat{S}_{\mathrm{prop}}(p,k) = g(\hat{\xi}_{\mathrm{prop}}(p,k), \gamma(p,k))X(p,k). \quad (9)$$

Since we confirmed the objective evaluation scores by Eq. (8) are better than those by overall a priori SNR flooring case in preliminary experiment, we adopt Eq. (8) and use it in subsequent comparative experiment.

3.3. Existence of musical-noise-free condition in biased HRNR

We investigate the speech quality obtained with the proposed method to adjust the internal parameter in Eq. (8). In the work, ρ' is fixed to three values of 0.1, 0.5 and 0.9 and the biased value ε' is increased from 0 to 3.0. The other experimental conditions were the same as those given in Subsect. 3.1. Figure 2 represents the relation between NRR and KR in the proposed method. Proposed method can achieve musical-noise-free condition by rising ε'. KR decreases rapidly when ε' begins to increase, whereas KR inflection becomes smaller as ε' approaches 3.0. Also, NRR in the $\rho' = 0.1$ case is larger than in any of the other cases in the musical-noise-free condition. Hence, it is better to set ρ' to a lower value. We also looked into the tendency obtained from the same experiments in other noise condition as in Subsect. 3.1 and confirmed the musical-noise-free conditions which become KR ≤ 1 exist by rising bias ε' in all noise types.

3.4. Effects of harmonic regeneration with biased HRNR

In the previous subsection, we confirmed that the musical-noise-free condition exists in the proposed method. In this subsection we will use spectrograms to show that biased HRNR can restore the eliminated harmonics. For comparison, we applied the MMSE-STSA estimator [4] and the proposed method to noisy speech mixed with white Gaussian noise at 10 dB SNR. To clearly demonstrate the harmonics were restored, we set NRR to 20 dB so that the two output signals would have serious distortion. Figure 3 shows spectrograms of clean speech (a), noisy speech (b), and suppressed speech obtained with the MMSE-STSA estimator (c) and the proposed method (d). From Fig. 3 (c) and (d), we can confirm that the proposed method is able to prevent the generation of musical noise and reconstructs the lost harmonics.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

To ascertain the validity of the proposed method, we performed a comparative experiment with three conventional speech enhancement methods: MMSE-STSA estimator, HRNR, and musical-noise-free MMSE-STSA estimator. The objective scores were KR and cepstral distortion (CD), which indicates the amount of speech distortion [21].

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, street, and white

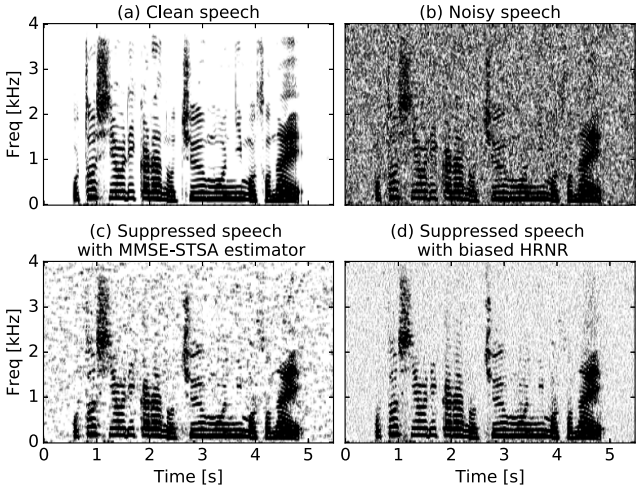

Fig. 3. Spectrograms of clean speech (a), noisy speech (b), and suppressed speech obtained with MMSE-STSA estimator (c) and proposed method (d).

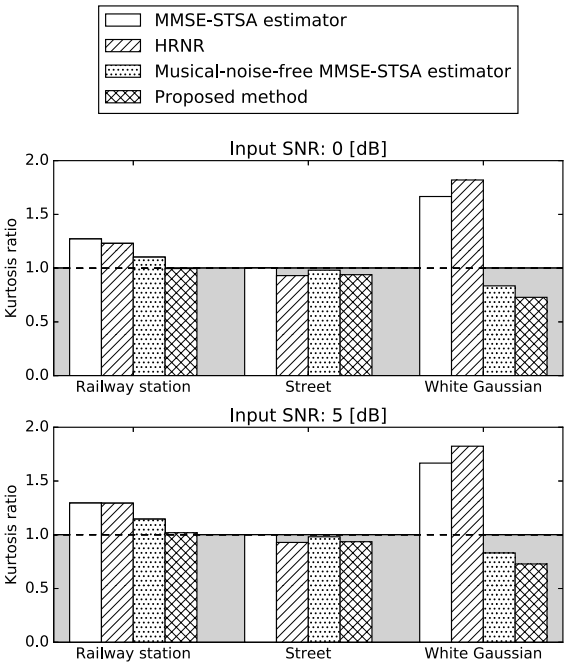

Fig. 4. KR at 0 dB (upper) and 5 dB (lower) input SNRs. Filled areas denote musical-noise-free condition.

Gaussian) at 0 dB and 5 dB input SNRs. Here, the gain functions of HRNR and the proposed method in Eq. (6) and Eq. (9) were set as the MMSE-STSA estimator. To achieve 10 dB NRR in each noise reduction method, we manually controlled the internal parameters ($\rho' = 0.1$ for the proposed method). Note that we made NRR of the three compared methods and proposed method even (i.e., we did not set their parameters so that these KRs are 1 or less); therefore, two of the musical-noise-free approaches we mentioned didn't have to achieve musical-noise-free condition.

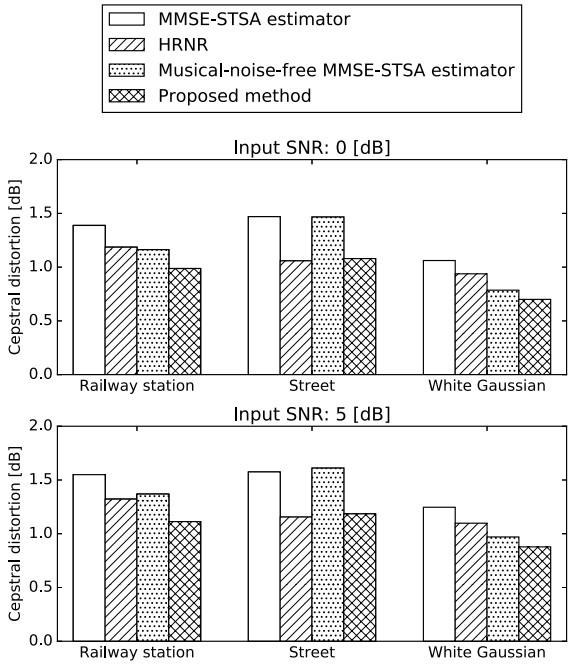

Fig. 5. CD at 0 dB (upper) and 5 dB (lower) input SNRs.

4.2. Results

The objective evaluation results obtained for musical noise generation and speech distortion are shown in Figs. 4 and 5. 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. 4, the musical-noise-free MMSE-STSA estimator and the proposed method achieved the musical-noise-free condition for most noise cases. Although the KR score of the proposed method is slightly more than 1.0 for railway station noise, the method generates little musical noise compared with other methods. This indicates it is effective in terms of musical noise generation. Next, Fig. 5 indicates that the proposed method achieved the lowest CD score in all cases. This shows it achieves high-quality noise reduction comprehensively compared with the other methods under the same NRR conditions.

5. CONCLUSION

In this paper we described a new method we propose that generates almost no musical noise with low speech distortion by applying biased harmonic regeneration noise reduction. Experimental evaluation results confirmed that the musical-noise-free condition could be obtained with the proposed method. A comparative experiment showed that the method is superior to conventional methods in musical noise generation and speech distortion.

6. REFERENCES

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

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

[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] 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.

[5] 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.

[6] 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.

[7] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE SPL*, vol. 14, no. 12, pp. 1036–1039, 2007.

[8] 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.

[9] 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.

[10] 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.

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

[12] J. S. Erkelens, J. Jensen, and R. Heusdens, "Improved speech spectral variance estimation under the generalized gamma distribution," pp. 43–46, 2007.

[13] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 14, no. 6, pp. 2098–2108, 2006.

[14] 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.

[15] 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," *Hands-free Speech Communication and Microphone Arrays*, pp. 122–126, 2014.

[16] H. Saruwatari, "Statistical-model-based speech enhancement with musical-noise-free properties," *Digital Signal Processing*, pp. 1201–1205, 2015.

[17] 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.

[18] 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.

[19] 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.

[20] 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 Echo and Noise Control*, 2012.

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