2017 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP'17)
Guam, USA, February 28th to March 3rd, 2017

Evaluation of Sound Quality and Speech Recognition Performance using Harmonic Regeneration for Various Noise Reduction Techniques

Masakazu Une[1] and Ryoichi Miyazaki[1]

[1] Department of Computer Science and Electronic Engineering, National Institute of Technology, Tokuyama College
Gakuendai, Shunan-shi, Yamaguchi, 745–8585 Japan
Phone:+81-834-29-6303 E-mail: i12une@tokuyama.ac.jp

1. Introduction

Many applications of speech communication systems, such as telecommunication systems, hearing aid systems, and video conference systems have been used. Although, these systems have the problem of speech quality deterioration under noisy conditions. Thus, it is necessary to obtain high quality speech signal by noise reduction. However, the very annoying distortion, so called musical noise arises owing to nonlinear signal processing. Also, the excessive noise reduction generates especially the speech distortion that leads to the degradation of the speech quality and the speech recognition performance.

To solve this problem, harmonic regeneration noise reduction (HRNR) has been proposed [1]. HRNR can regenerate the lost harmonic and obtain the good quality speech. Nevertheless, HRNR mentions only the validity for Wiener Filtering (WF) [2]. In this paper, we evaluate the sound quality and the speech recognition performance using HRNR technique for three noise reduction techniques (WF, Spectral Subtraction (SS) [3], Minimum Mean-Square Error Short-Time Spectral Amplitude (MMSE-STSA) estimator [4]).

2. Noise Reduction Methods

2.1 Definition of Signals

The observed signal $x(t)$ is composed of the speech signal $s(t)$ and the noise signal $n(t)$ as $x(t) = s(t)+n(t)$. Then, we apply the short-time Fourier transform to the observed signal to obtain the time-frequency signal $X(p, k)$, 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. Next, we estimate the spectrum of the speech signal $\hat{S}(p, k)$ as

$$\hat{S}(p, k) = G(p, k)X(p, k) \tag{1}$$

where $G(p, k)$ is the spectral gain of each noise reduction method.

2.2 WF

WF is a classical noise reduction technique [2] and its spectral gain $G_{\mathrm{WF}}(p, k)$ is defined as

$$G_{\mathrm{WF}}(p, k) = \frac{\xi(p, k)}{1 + \xi(p, k)} \tag{2}$$

Here, $\xi(p, k)$ is a priori SNR, defined as $\xi(p, k) = \mathrm{E}[|S(p, k)|^2]/\mathrm{E}[|N(p, k)|^2]$, where $\mathrm{E}[\cdot]$ is the expectation operator. However, since we cannot estimate $|S(p, k)|^2$ in advance, we calculate the a priori SNR via the decision-directed approach as [4]

$$\begin{aligned}\hat{\xi}(p, k) =& \alpha \frac{|G(p-1, k-1)X(p-1)|^2}{\mathrm{E}[|\hat{N}(p, k)|^2]} \\ &+ (1-\alpha)\mathrm{Max}[\gamma(p, k) - 1, 0]\end{aligned} \tag{3}$$

where $\hat{N}(p, k)$ is the estimated noise signal, α is the forgetting factor. Generally, the forgetting factor α is set to 0.98 to obtain the better sound quality [4]. Also, $\gamma(p, k)$ is a posteriori SNR defined as $\gamma(p, k) = |X(p, k)|^2/\mathrm{E}[|\hat{N}(p, k)|^2]$.

2.3 SS

SS is a usually used noise reduction technique that has small amount of calculation and high noise suppression performance [3]. This method subtracts the estimated noise signal from the observed signal in power spectral domain. The spectral gain $G_{\mathrm{SS}}(p, k)$ is expressed as

$$G_{\mathrm{SS}}(p, k) = \begin{cases} \sqrt{1 - \frac{\beta \mathrm{E}[|\hat{N}(p,k)|^2]}{|X(p,k)|^2}} & (|X(p, k)|^2 > \beta \mathrm{E}[|\hat{N}(p, k)|^2]) \\ 0 & (\mathrm{otherwise}) \end{cases} \tag{4}$$

where β is the oversubtraction parameter that adjusts the amount of noise suppression.

2.4 MMSE-STSA Estimator

The MMSE-STSA estimator minimizes the mean-square error between the amplitude spectra of the original speech and the estimated speech signals. The spectral gain $G_{\mathrm{STSA}}(p,k)$ is expressed as

$$G_{\mathrm{STSA}}(p,k) = \frac{\sqrt{\nu(p,k)}}{\gamma(p,k)}\Gamma\left(\frac{3}{2}\right) M\left(-\frac{1}{2};1;-\nu(p,k)\right) \tag{5}$$

where $\Gamma(h)$ and $M(a;b;z)$ are the gamma function and the confluent hypergeometric function, respectively. $\nu(p,k)$ is expressed as follows:

$$\nu(p,k) = \frac{\xi(p,k)}{1+\xi(p,k)}\gamma(p,k) \tag{6}$$

where $\xi(p,k)$ is estimated using the decision-directed approach in eq. (3) as WF.

3. Harmonic Regeneration

The excessive noise reduction generates harmonic distortion and degrades the speech quality. The HRNR technique is proposed to restore the missing harmonics due to noise reduction [1]. The block diagram of HRNR technique is shown in Fig. 1.

HRNR spectral gain $G_{\mathrm{HRNR}}(p,k)$ consists of the refined a priori SNR $\hat{\xi}_{\mathrm{HRNR}}(p,k)$ and the a posteriori SNR $\gamma(p,k)$ as

$$G_{\mathrm{HRNR}}(p,k) = \upsilon(\hat{\xi}_{\mathrm{HRNR}}(p,k), \gamma(p,k)) \tag{7}$$

where function υ can be chosen from various spectral gain functions (WF, MMSE-STSA estimator, etc.). Also, the refined a priori SNR $\hat{\xi}_{\mathrm{HRNR}}(p,k)$ is computed by

$$\hat{\xi}_{\mathrm{HRNR}}(p,k) = \frac{\rho(p,k)|\hat{S}(p,k)|^2 + (1-\rho(p,k))|S_{\mathrm{harmo}}(p,k)|^2}{\mathrm{E}[|\hat{N}(p,k)|^2]} \tag{8}$$

where $\rho(p,k)$ is used to adjust the mixing level of $|\hat{S}(p,k)|$ and $|S_{\mathrm{harmo}}(p,k)|$. It is better to set $\rho(p,k)$ equal to the spectral gain in noise reduction part ($G_{\mathrm{WF}}(p,k)$, $G_{\mathrm{SS}}(p,k)$, $G_{\mathrm{STSA}}(p,k)$, etc.) [1]. The restored signal spectrum $S_{\mathrm{harmo}}(p,k)$ is obtained by

$$S_{\mathrm{harmo}}(p,k) = \mathrm{FT}\left[NL\left(\mathrm{IFT}\left[\hat{S}(p,k)\right]\right)\right] \tag{9}$$

where $NL(\cdot)$ is non-linear function (e.g., absolute value, minimum, or maximum relative to a threshold, etc.), $\mathrm{FT}[\cdot]$ and $\mathrm{IFT}[\cdot]$ represent the Fourier and the inverse Fourier transforms, respectively. Finally, the regenerated harmonics signal computed as

$$\hat{S}_{\mathrm{HRNR}}(p,k) = G_{\mathrm{HRNR}}(p,k)X(p,k) \tag{10}$$

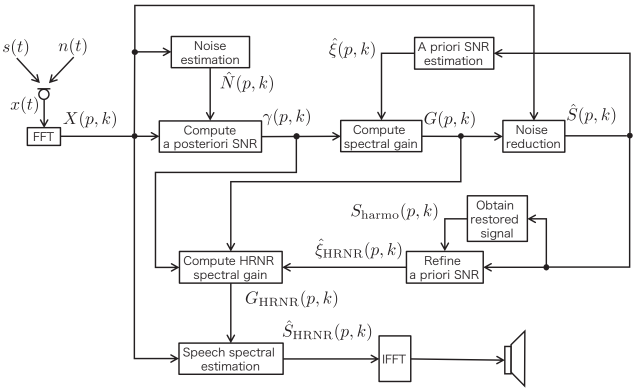

Figure 1: Block diagram for HRNR

4. Objective Evaluation Experiment

4.1 Overview

In ref. [1], it was reported only the validity for WF in terms of speech distortion. Then, we evaluate the second quality on the basis of the amount of musical noise generation and SNR improvement for various noise reduction techniques. In this section, we use following three objective evaluations (Noise Reduction Rate: NRR [5] for SNR improvement, Kurtosis Ratio: KR [5] for the amount musical noise generation, and Cepstral Distortion: CD [6] for speech distortion) and investigate the sound quality for HRNR.

4.2 Definition of Objective Evaluations

4.2.1 NRR

We introduce NRR to evaluate the SNR improvement as

$$\mathrm{NRR} = 10\log_{10}\frac{\mathrm{E}[s_{\mathrm{out}}^2]/\mathrm{E}[n_{\mathrm{out}}^2]}{\mathrm{E}[s_{\mathrm{in}}^2]/\mathrm{E}[n_{\mathrm{in}}^2]} \tag{11}$$

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

4.2.2 KR

KR is an objective measurement for musical noise generation. KR is defined by

$$\mathrm{KR} = \mathrm{kurt}_{\mathrm{proc}}/\mathrm{kurt}_{\mathrm{org}} \tag{12}$$

where $\mathrm{kurt}_{\mathrm{proc}}$ is the kurtosis of the processed signal and $\mathrm{kurt}_{\mathrm{org}}$ is the kurtosis of the observed signal. KR decreases as the amount of generated musical noise decreases.

4.2.3 CD

CD is an objective measurement for speech distortion defined by

$$\mathrm{CD} = \frac{20}{T\log 10}\sum_{p=1}^{T}\sqrt{\sum_{k=1}^{B}2(C_{\mathrm{out}}(p,k)-C_{\mathrm{ref}}(p,k))^2} \tag{13}$$

where T is the frame length, B is the number dimensions of cepstrum, $C_{\mathrm{out}}(p,k)$ and $C_{\mathrm{ref}}(p,k)$ are cepstral coefficients of after processing and the clean speech, respectively.

4.3 Experimental Condition

We evaluated the effect of HRNR using three objective measurements (NRR, KR, CD). We used 10 speakers (five males and five females) from the JNAS database [7] for the target signals. Also, we mixed the speech signals with three types of noise (railway station noise, street noise and white Gaussian noise). The input SNR of the test data is set to 10 dB or 15 dB. All the signals used in this experiment were sampled at 16 kHz with 16 bit accuracy. The size of the Fourier transform is 512, and frame shift length was is 128.

In this experiment, the signals suppressed by three classical noise reduction techniques (WF, SS, and MMSE-STSA estimator) and restored by HRNR technique. We set the gain function υ in eq. (7) and parameter $\rho(p,k)$ in eq. (8) to the spectral gain of each noise reduction method (e.g., υ is set to $\xi_{\mathrm{HRNR}}/(1+\xi_{\mathrm{HRNR}})$ and $\rho = G_{\mathrm{WF}}$ in WF case). Also, we define this case as WF + HRNR, as well as SS+HRNR, MMSE-STSA estimator + HRNR.

The chosen nonlinear function $NL(\cdot)$ in eq. (9) is the half-wave function. We manually controlled the forgetting factor α for WF and MMSE-STSA estimator and the oversubtraction parameter β for SS to achieve 10-dB NRR. We set the dimension of cepstrum B in eq. (13) to 22.

4.4 Experimental Results

Figure 2 shows the spectrograms that the noisy signal is mixed the clean speech signal with white Gaussian noise in 10-dB input SNR. From Fig. 2 (c) we can confirm that the degraded harmonic component is regenerated by HRNR as compared to Fig. 2 (b).

Next, Figs. 3 and 4 indicate the result of each objective evaluation for NRR, KR and CD, respectively. Fig. 3 (a) and Fig. 4 (a) show that the variation of NRR by applying HRNR is little. Regarding Fig. 3 (b) and Fig. 4 (b), it is shown that KR decrease in all noisy environment when the spectral gain is set to SS and the enhanced signal is restored by HRNR. Consequently, the amount of musical noise generation decreases. In addition, from Fig. 3 (c) and Fig. 4 (c), HRNR reduces CD in all noisy environment. It follows that when we apply HRNR, we can obtain an enhanced speech signal with less speech distortion in SS and MMSE-STSA estimator.

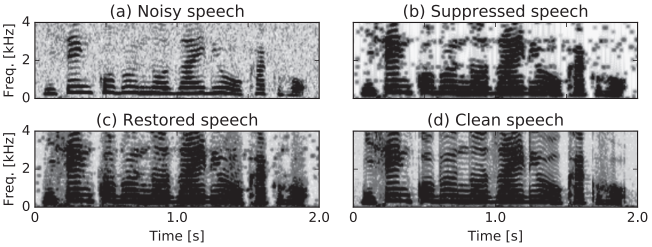

Figure 2: Spectrograms of (a): Noisy speech mixed with white Gaussian noise at 10-dB SNR, (b): Suppressed speech by SS, (c): Suppressed speech signal by HRNR, and (d): Clean speech signal

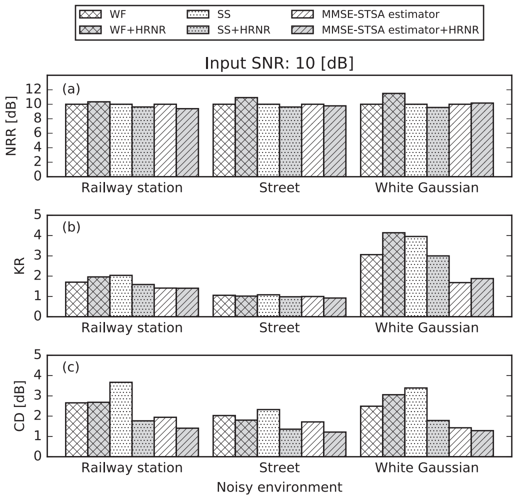

Figure 3: Objective evaluation at 10-dB input SNR. (a): NRR (b): KR, and (c): CD

5. Speech Recognition Experiment

5.1 Experimental Condition

In this section, we evaluated the speech recognition performance for HRNR. We used 200 speakers (100 males and 100 females) for the target speech signals. Also, we compared the speech recognition performance of HRNR with those of WF, SS, MMSE-STSA estimator, HRNR, clean speech signals, and unprocessed signals. Other experimental conditions are the same in section 4.3.

5.2 Speech Recognition Result

Figures 5 and 6 show the result of the speech recognition performance. From Figs. 5 and 6, we can confirm that the

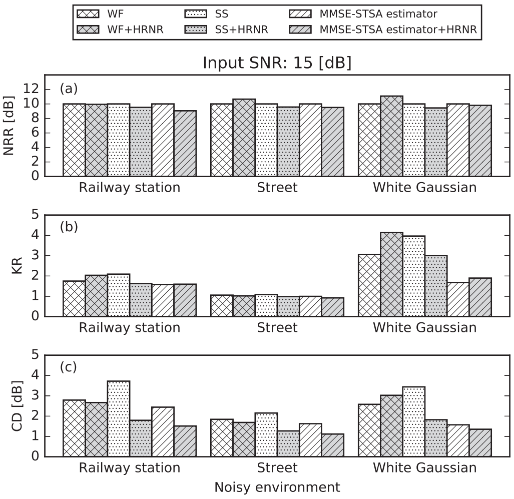

Figure 4: Objective evaluation at 15-dB input SNR. (a): NRR (b): KR, and (c): CD

speech recognition performance of the signals suppressed by SS is improved a lot by HRNR technique compared with the signals suppressed by WF or MMSE-STSA estimator case. Consequently, HRNR technique is effectual using SS.

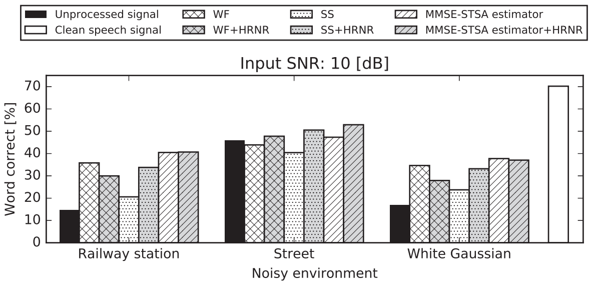

Figure 5: Speech recognition performance in three noisy environments at 10-dB input SNR

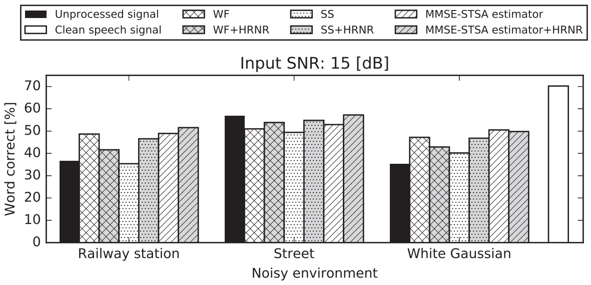

Figure 6: Speech recognition performance in three noisy environments at 15-dB input SNR

6. Conclusion

In this paper, we evaluated the sound quality of the signal processed by HRNR for various noise reduction techniques. From the objective evaluation, the variation of NRR by applying HRNR is little, and the scores of KR and CD are improved by applying HRNR in almost all the cases. Also, from the speech recognition experiment, we can confirm that applying HRNR leads to the improvement of the speech recognition performance in almost all the cases.. We concluded that SS is particularly effective for HRNR in various noisy environment.

Acknowledgment

This work was supported by Grant-in-Aid for Young Scientists (B) 16K21579.

References

[1] 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.

[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] R. Miyazaki, *et al.*, "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.

[6] L. Rabiner and B. Juang, "Fundamentals of Speech Recognition," *Upper Saddle River, NJ: Prentice-Hall*, 1993.

[7] K. Ito, *et al.*, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of Acoustical Society of Japan*, vol.20, pp.196–206, 1999.