2012 International Workshop on Nonlinear Circuits,
Communications and Signal Processing
NCSP'12, Honolulu, Hawaii, USA. March 4-6, 2012

Speech enhancement technique in noisy reverberant environment using two microphone arrays

Yuuki Sasaki and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
Phone/FAX:+81-761-51-1391/+81-761-51-1149
Email: {s0910025, akagi}@jaist.ac.jp

Abstract

Recent years, speech enhancement techniques in order to suppress noise and/or reverberation have been introduced into applications like hearing-aid. However, proposed speech enhancement tecniques supporting selective binaural selective hearing cannot work in noisy reverberant environment. This paper aims at constructing speech enhancement supporting binaural selective hearing in noisy reverberant environment. Experiment I verifies whether Two–Stage Binaural Speech Enhancement with Wiener Filter (TS–BASE/WF) can suppress early reflection and late reverberation. Results show that TS–BASE/WF cannot suppress early refractions due to using a wiener filter. Cepstral mean subtraction (CMS) is used as front–end of TS–BASE/WF based on this result. Experiment II is carried out to show whether the modified method is superior to TS–BASE/WF in noisy reverberant environments. This result indicates that the modified method exceeds TS–BASE/WF.

1. Introduction

Speech communication becomes difficult under influence of noise and/or reverberation. Additionally, there are some reports that listening capability of hearing handicapped person declines remarkably in noisy reverberant environment. Therefore, speech enhancement techniques in order to suppress noise and/or reverberation have been introduced into applications like hearing-aid. In speech enhancement techniques proposed until now [1][2][3], some speech enhancement techniques focused on binaural hearing featured of humans.

Frequency domain binaural model (FDBM) [5] based on Lindeman's binaural hearing model [4] was proposed by Usagawa *et al*. This method calculates interaural phase difference and interaural level difference to estimate the direction of the target signal. Then, the received signal is enhanced by FDBM. Two–Stage Binaural Speech Enhancement with Wiener Filter (TS-BASE/WF) [6] was proposed by Li *et al*., to suppress noise with two-step processing; noise estimation stage and noise suppression one. TS–BASE/WF has excellent noise-reduction performance, because TS-BASE/WF has two-step processing.

When these speech enhancement techniques are used indoors, suppression ability of noise and reverberation simultaneously should be required. Room impulse responses (RIR) can be divided into early reflection and late reservation bordering on the time that is dependent on size of the room. Early reflection correlates to the target signal. Late reverberation that is added several reflection sounds have less correlation to the target signal. Moreover, late reverberation diffuses around the room. Almost all of speech enhancement techniques for supporting binaural selective hearing cannot suppress reverberation. However, noise estimation stage of TS–BASE/WF without using cross-spectrum could estimate target signal in reverberant environment. On the one hand, since noise suppression stage of TS–BASE/WF adopt Wiener filter in which it is assumed there is no correlation between target signal and noise. Hence, enhanced signal could be affected by noise suppression stage of TS–BASE/WF.

In this paper, first, performance of TS–BASE/WF in reverberant environment is measured. Those results show that TS–BASE/WF can estimate late reverberation. However, it cannot suppress early reflection because of using a Wiener filter. According to those results, modifided method of TS–BASE/WF will be constructed. Performance of proposed method in noisy reverberant environment is evaluated in this paper.

2. TS–BASE/WF

The block diagram of TS–BASE/WF is shown in Fig. 1. TS–BASE/WF is a speech enhancement technique for supporting binaural selective hearing in noisy environment.

2.1. Noise estimation stage

Noise estimation stage in TS-BASE/WF refers to the equalization-cancellation (EC) model. The EC model was developed by Durlach [7] and further improved by Culling and Sumerfiled [8]. These EC models can explain many psychoacoustic effects.

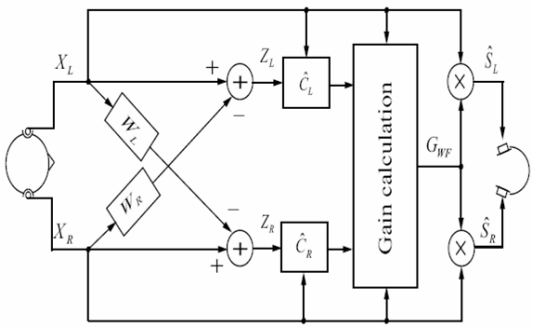

Figure 1: The block diagram of TS–BASE/WF

2.1.1. Equalization process

Equalizing filters are applied to the left and right input signals for equalizing the components of target signal in these input signals. Equalizing filters can be obtained by normalized least mean square (NLMS) algorithm. Based on the assumption that the direction of the target signal is known a priori, two filters are pre-learned in the absence of noise.

2.1.2. cancellation process

In the cancellation process, equalizing filters are fixed and applied to input signals. The target-cancelled signals are derived by subtracting the filter-calibrated inputs at one ear from the input signals at the other ear.

2.2. Noise suppression stage

Wiener filter with synthesized gain functions based on estimated noises is applied to the input signals. Enhanced signals keep direction of each sound source, because of applying common gain function to both channels. Moreover, when gain function is synthesized by Wiener filter, the mean square error between the target-cancelled signals and the input signals is set to minimum. Then enhanced signals are reduced musical noise.

3. Date Base

In experiments, continuous speech sentences uttered by three male speakers and one male speaker were selected from NTT database with sampling rate of 44.1 kHz at 16 bit resolution. The head-related impulse responses (HRIR) measured at the MIT media lab with a sampling rate of 44.1 kHz at 16 bit resolution. Evaluation of all sound sources is 0 degree, angle of the sound source located front of dummy head is 0 degree. The right side of dummy head is +, the left side of dummy head is - .

RIR is synthesized by image method [9]. However, synthesizeed RIR is not fully reflected in direct and reflected sound direction information. Then, direct and reflection sounds were convoluted HRTF corresponding to sound source directions.

The experimental signals were down-sampled to 16 kHz. The analysis frame length for FFT is 512 samples (32 ms), the overlap is 1/2 in TS-BASE/WF by using hanning window. TS-BASE/WF focused on the front of the dummy head.

4. Objective evaluation measures

The segmental signal-to-noise ratio (SEGSNR) and log-spectral distortion (LSD) measures were used for evaluating performances in this paper. SEGSNR and LSD were calculated by those formulas.

SEGSNR

$$= \frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\left(\frac{\sum_{k=0}^{K-1}[s(lK+k)]^2}{\sum_{k=0}^{K-1}[s(lK+k)-\hat{s}(lK+k)]^2}\right), \quad (1)$$

LSD

$$= \frac{10}{L}\sum_{l=0}^{L-1}\left(\frac{1}{K}\sum_{k=0}^{K-1}[\log_{10} AS_d(k,l) - \log_{10} A\hat{S}(k,l)]^2\right). \quad (2)$$

Where s is target signal, $\hat{s}$ is the mean of internal experimental signal or enhanced signal. The target signals were convoluted HRTF corresponding to the sound source direction. L and K indicate the number of frames in the signal and the frame length in samples. Where $AS(k,l) \equiv \max\{|S(k,l)|^2, \delta\}$ is the clipped spectral power, such that the log-spectrum dynamic range is confined to about 50 dB (that is, $\delta = 10^{-50/10}$).

5. Experiment I

The experiment I verifies whether TS-BASE/WF can suppress early reflection and late reverberation.

5.1. How to synthesize experimental sound

Controlling numbers of maximum reflection time and minimum reflection time of reflection sounds, room impulse responses (RIR) are synthesized for Experiment I. The maximum reflection time was varied from 1 to 40 or the minimum reflection time was varied from 1 to 39 while the maximum reflection time was fixing 40. In all RIRs, T_{60} is fixed to 2 s. Experimental sounds were synthesized by convolving the audio signal to the RIR.

5.2. Results

The results are shown in Figs. 2 and 3. Additive reflection show the result of controlling numbers of maximum reflection time, eliminated reflection show the result of controlling

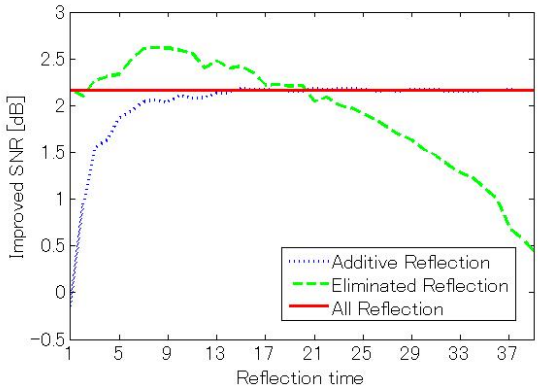

Figure 2: Improved SEGSNR of enhanced signals using by TS-BASE/WF

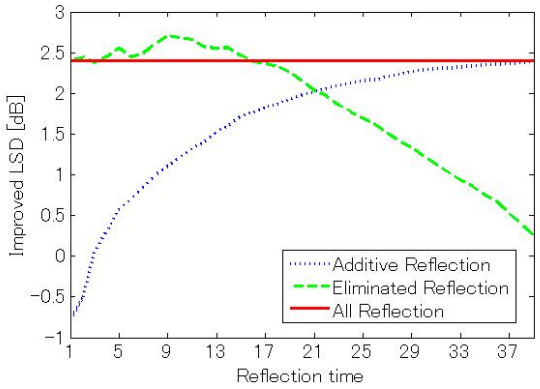

Figure 3: Improved LSD of enhanced signals using by TS-BASE/WF

numbers of minimum reflection time. All refrections show the result of using RIR which is set maximum reflection time to 40. The improved SEGSNR and LSD take maximum values, when low number of reflection sounds were eliminated. Low number of reflection sounds similar to early reflection. Then, Figs 2 and 3 indicate that TS-BASE/WF can suppress late reverberation. However, it cannot suppress early reflection because of using a Wiener filter.

6. Modified method

According to the results of Experiment I, Cepstral Mean Subtraction (CMS) [10] is used as front–end for TS–BASE/WF in order to suppress early reflection. The proposed method is called CMS + TS–BASE/WF.

6.1. CMS

The Amplitude cepstral of the RIR is estimated by normalizing the mean of the left and right input signals in quefrency domain. After that, the amplitude cepstral of the left and right input signal are subtracted by estimated the amplitude cepstral of the RIR for each frame.

6.2. Cepstral mean normalize

The input signal, x is shown by convolution of the RIR and the target signal.

$$x(t) = s(t) * h(t), \tag{3}$$

Where h is the RIR, s is the target signal. The input signal is subjected to Fourier transform. Then, amplitude spectrum is taken the logarithmic.

$$c_x(k,l) = c_s(k,l) + c_h(k,l), \tag{4}$$

Where l is the frame number, k is the quefrency, and c is the amplitude cepstral corresponding to subscript. Next, the amplitude cepstral of the input signal were divided to L frame. And the mean normalize of that was caliculated.

$$\begin{aligned} c_{ave}(k,l) &= \frac{1}{\sum_{l=0}^{L-1}\exp(-\alpha\cdot l)}\sum_{l=0}^{L-1} c_x(k,l)\cdot\exp(-\alpha\cdot l) \\ &= c_{ave:s}(k,l) + c_{ave:h}(k,l) \\ &\approx \hat{c}_h(k,l). \end{aligned} \tag{5}$$

Where subscript ave mean calculating the mean normalize. $\exp(-\alpha\cdot l)$ is forgetting factor for past frames. If the time-invariant RIR, s will be offset. Then, the amplitude cepstral of the RIR can be estimated.

6.3. Subtract on quefrency domain

The amplitude cepstral of the input signal are subtracted by estimated the amplitude cepstral of the RIR in each frame.

$$\hat{c}_s(k,l) = c_x(k,l) - \beta\cdot\hat{c}_h(k,l), \quad 1 > \beta > 0. \tag{6}$$

β is used to compensate for the amplitude cepstral of the RIR in equation (6). After this process, enhanced signal is synthesized using $c_s(k,l)$ and the phase spectrum of the input signal.

7. Experiment II

The Experiment II is carried out to show whether the CMS + TS–BASE/WF is superior to TS–BASE/WF in noisy reverberant environments.

7.1. How to synthesize experimental sound

The angle of the target signal is 0 degree, the angle of the noise is 45 dgree. The specker used noise is different to target signal specker. The target signal and the noise were convoluted each RIRs which include HRTFs. Inaddition, the input signals is sum of the target signal and noise.

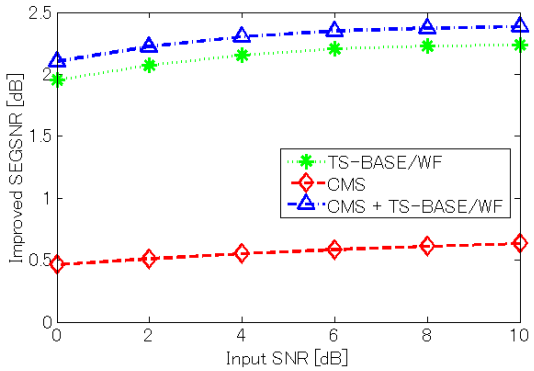

Figure 4: Improved SEGSNR of enhanced signals using by TS-BASE/WF or the modified method

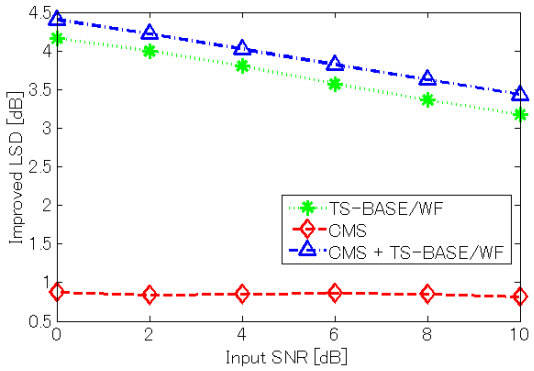

Figure 5: Improved LSD of enhanced signals using by TS-BASE/WF or the modified method

Ratio on all interval is varied from 0 dB to 10 dB in increments of 2 dB. The RIRs, T_{60} is fixed to 2.0 s, number of reflections of the RIR is fixed to 40 times. Experimental sound was synthesized by convolving the audio signal to the RIR.

7.2. Parameter settings for the CMS

The analysis frame length for FFT is 512 samples (32 ms), the overlap is 1/4 by hanning window in CMS. The total number of frames when mean normalizing is calculated is 60. The parameter α is 0.008, the parameter β is 0.14. These parameters are decided by same experiment.

7.3. Results

The results are shown in Figs. 4 and 5. Figures 4 and 5 indicates that the CMS + TS-BASE/WF exceed TS-BASE/WF. On the other hand, improved values of CMS are certain. Then, the perfomance of TS–BASE/WF is improved, because CMS is working well.

8. Conclusion

The results of Experiment I show that TS–BASE/WF not sufficiently suppressed early reflection due to using wiener filter. As the solution, CMS was adopted as the front–end of TS–BASE/WF. The results of Experiment II show that the performance of CMS + TS–BASE/WF is superior to the performance of TS–BASE/WF.

References

[1] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. ASSP, Vol. 27, No. 2, pp. 113–120, 1979.

[2] L. J. Griffiths and C. W. Jim, “An altanative approach to linearly constrained adaptive beamforming,” IEEE Trans. AP, Vol.30, No. 1, pp. 27–34, 1982.

[3] E. A. P. Habets, “Single–channel speech dereverberation based on spectral subtraction” In Proc. ProRISC, pp. 250–254, 2004.

[4] W. Lindemann, “Extension of a binaural cross-correlation model by contralateralinhibition.I. Simulation of lateralization for stationary signals,” J.Acoust. Soc. AM., 80, 1608–1622, 1986.

[5] T. Usagawa, K. Sakai and M. Ebata, “Frequency domain binaural model as the front end of speech recongnition system,” Proc. ICSL98, 1998.

[6] J. Li, S. Sakamoto, M. Akagi, and Y. Suzuki, “A two–stage binaural speech enhancement with wiener filter (TS–BASE/WF) for high–quality speech comunication,” Proc. IEEE WSPAA, New Paltz, New York, 2009.

[7] N. I. Durlach, “Equalization and cancellation theory of binaural masking level differences,” JASA, Vol. 35, no. 8, pp. 1206–1218, 1979.

[8] J. F. Culling, M. L. Hawley and R. Y. Litovsky, “The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources” Jounal of Acoustic Society of America, p1057–1065, 2004.

[9] J. Allen and D. Berkley, “Image method for efficiently simulating small room acoustics,” Journal of Acoustic Society of America, p912–915, 1979.

[10] T. G. Stockham, Jr., “Restoration of old acoustic recordings by means of digital signal processing,” Preprint, 41st Convention, Audio Engineering Society, New York, 1971.