Theoretical analysis of musical noise and speech distortion in structure-generalized parametric blind spatial subtraction array

Ryoichi Miyazaki[1], *Hiroshi Saruwatari*[1], *and Kiyohiro Shikano*[1]

[1]Nara Institute of Science and Technology, Nara, Japan

ryoichi-m@is.naist.jp, sawatari@is.naist.jp, shikano@is.naist.jp

Abstract

In this paper, we propose a structure-generalized parametric blind spatial subtraction array (BSSA), and the theoretical analysis of the amounts of musical noise and speech distortion is conducted via higher-order statistics. We theoretically prove a tradeoff between the amounts of musical noise and speech distortion in various BSSA structures. From the analysis and experimental evaluations, we reveal that the structure should be carefully selected according to the application, i.e., a channelwise BSSA structure is recommended for listening but a normal BSSA is more suitable for speech recognition.

Index Terms: speech enhancement, musical noise, higher-order statistics, speech recognition performance

1. Introduction

To achieve high-quality speech enhancement, noise reduction using a microphone array has been widely studied, and recently, speech extraction methods based on independent component analysis (ICA) have been proposed (see, e.g., [1, 2]). We previously proposed a blind spatial subtraction array (BSSA) [3] that consists of accurate noise estimation by ICA and the following speech extraction procedure based on nonlinear noise reduction such as spectral subtraction (SS) [4]. However, BSSA always suffers from artificial distortion, so-called musical noise, owing to nonlinear signal processing. This leads to a serious tradeoff between the noise reduction performance and amount of signal distortion in speech recognition.

In a recent study, two types of BSSA have been proposed (see Fig. 1). One is the traditional BSSA structure that performs SS after delay and sum (DS) (see Fig. 1(a)), and another is that SS is channelwisely performed before DS (chBSSA; see Fig. 1(b)). Also, it has been theoretically clarified that chBSSA is superior to BSSA in the mitigation of the musical noise generation [5]. However, it still remains as an open problem that there is no evaluation of the amount of speech distortion and speech recognition performance in the various BSSA structures.

Therefore, in this paper, we generalize such various types of BSSA as *structure-generalized parametric BSSA*, and we provide a theoretical analysis of the amounts of musical noise generated and speech distortion in several types of methods in the structure-generalized parametric BSSA. From a mathematical analysis based on the higher-order statistics, we prove a tradeoff between the amounts of musical noise and speech distortion in various BSSA structures. From experimental evaluations, we reveal that the structure selection should be carefully performed according to the application, i.e., a chBSSA structure is recommended for listening but normal BSSA is more suitable for speech recognition.

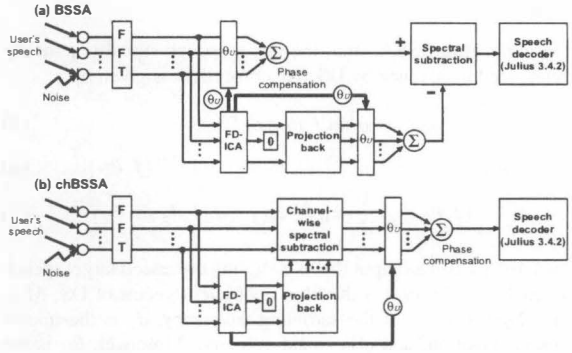

Figure 1: (a) Block diagram of SS after DS (BSSA), and (b) channelwise SS before DS (chBSSA).

2. BSSA and its generalization [3]

2.1. Overview of BSSA

The observed signal vector of the J-channel array in the time-frequency domain, $\boldsymbol{x}(f,\tau) = [x_1(f,\tau),\ldots,x_J(f,\tau)]^{\mathrm{T}}$, is given by

$$\boldsymbol{x}(f,\tau) = \boldsymbol{h}(f)s(f,\tau) + \boldsymbol{n}(f,\tau), \tag{1}$$

where f is the frequency bin, τ is the frame number, $\boldsymbol{h}(f) = [h_1(f),\ldots,h_J(f)]^{\mathrm{T}}$ is a column vector of transfer function from the target signal component to each microphone, $s(f,\tau)$ is a target speech signal component, and $\boldsymbol{n}(f,\tau) = [n_1(f,\tau),\ldots,n_J(f,\tau)]^{\mathrm{T}}$ is a column vector of the additive noise signal.

2.2. ICA-based noise estimation

BSSA includes ICA-based noise estimation. In the ICA part, we perform signal separation using a complex-valued matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$, so that the output signals $\boldsymbol{o}(f,\tau)$ become mutually independent; this procedure can be represented as

$$\boldsymbol{o}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau) = [o_1(f,\tau),\ldots,o_J(f,\tau)]^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) = \mu[\boldsymbol{I} - \langle\boldsymbol{\varphi}(\boldsymbol{o}(f,\tau))\boldsymbol{o}^{\mathrm{H}}(f,\tau)\rangle_\tau]\cdot\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f), \tag{3}$$

where μ is the step-size parameter, $[p]$ is used to express the value of the pth step in iterations, and $\boldsymbol{I}$ is the identity matrix. Besides, $\langle\cdot\rangle_\tau$ denotes a time-averaging operator, and $\boldsymbol{\varphi}(\cdot)$ is an appropriate nonlinear vector function.

Next, the estimated target speech signal is discarded as it is not required because we want to estimate only the noise component. Instead, we construct a *noise-only vector* $\boldsymbol{o}^{(\mathrm{noise})}(f,\tau) = [o_1(f,\tau),\ldots,o_{s-1}(f,\tau),0,o_{s+1}(f,\tau),\ldots,o_J(f,\tau)]^{\mathrm{T}}$ from the output signal obtained by ICA using (2), where $o_s(f,\tau)$ is assumed to be the speech component. Following this, we apply the projection back operation to remove the ambiguity of amplitude and construct the estimated noise signals $\boldsymbol{z}(f,\tau) = [z_1(f,\tau),\ldots,z_J(f,\tau)]^{\mathrm{T}}$ as

$$\boldsymbol{z}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{-1}(f)\boldsymbol{o}^{(\mathrm{noise})}(f,\tau). \tag{4}$$

2.3. Formulation of structure-generalized parametric BSSA

In parametric BSSA, first, the target speech signal is partially enhanced in advance by DS. This procedure is given by

$$y_{\mathrm{DS}} = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f,\theta_U)\boldsymbol{x}(f,\tau), \tag{5}$$

$$\boldsymbol{w}_{\mathrm{DS}}(f,\theta_U) = [w_1^{(\mathrm{DS})}(f,\theta_U),\ldots w_J^{(\mathrm{DS})}(f,\theta_U)]^{\mathrm{T}}, \tag{6}$$

$$w_j^{(\mathrm{DS})}(f,\theta_U) = \frac{1}{J}\exp\left(-2i(f/M)f_s d_j \sin\theta_U/c\right), \tag{7}$$

where y_{DS} is the output that is a slightly enhanced target speech signal, $\boldsymbol{w}_{\mathrm{DS}}(f,\theta_U)$ is the filter coefficient vector of DS, M is the DFT size, f_s is the sampling frequency, d_j is the microphone position, and c is sound velocity. Moreover, θ_U is the estimated direction of arrival of the target speech.

Next, using (4) and (5), we perform generalized SS (GSS) [6] and obtain the target-speech-enhanced signal as

$$y_{\mathrm{BSSA}}(f,\tau) = \begin{cases} \sqrt[2n]{|y_{\mathrm{DS}}(f,\tau)|^{2n} - \beta\overline{|z_{\mathrm{DS}}(f,\tau)|^{2n}}}\, e^{i\arg(y_{\mathrm{DS}}(f,\tau))} \\ (\text{if} \quad |y_{\mathrm{DS}}(f,\tau)|^{2n} - \beta\overline{|z_{\mathrm{DS}}(f,\tau)|^{2n}} > 0), \\ 0 \quad (\text{otherwise}), \end{cases} \tag{8}$$

$$z_{\mathrm{DS}}(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{z}(f,\tau), \tag{9}$$

where $y_{\mathrm{BSSA}}(f,\tau)$ is the final output of parametric BSSA, β is an oversubtraction parameter, n is an exponent parameter, and $\overline{|z_{\mathrm{DS}}(f,\tau)|^{2n}}$ is the smoothed noise component within a certain time frame window.

In parametric chBSSA, we first perform GSS independently in each input channel and derive multiple target-speech-enhanced signals by channelwise GSS. Using (1) and (4), we obtain the target-speech-enhanced signal based on GSS. This procedure can be given by

$$y_j^{(\mathrm{chGSS})}(f,\tau) = \begin{cases} \sqrt[2n]{|x_j(f,\tau)|^{2n} - \beta\overline{|z_j(f,\tau)|^{2n}}}\, e^{i\arg(x_j(f,\tau))} \\ (\text{if} \quad |x_j(f,\tau)|^{2n} - \beta\overline{|z_j(f,\tau)|^{2n}} > 0), \\ 0 \quad (\text{otherwise}), \end{cases} \tag{10}$$

where $y_j^{(\mathrm{chGSS})}(f,\tau)$ is the target-speech-enhanced signal obtained by GSS at a specific channel j. Finally, we obtain the resultant target-speech-enhanced signal by applying DS to $\boldsymbol{y}_{\mathrm{chGSS}} = [y_1^{(\mathrm{chGSS})}(f,\tau),\ldots,y_J^{(\mathrm{chGSS})}(f,\tau)]^{\mathrm{T}}$. This procedure can be given by

$$y_{\mathrm{chBSSA}}(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{y}_{\mathrm{chGSS}}(f,\tau), \tag{11}$$

where $y_{\mathrm{chBSSA}}(f,\tau)$ is the final output of parametric chBSSA.

3. Theoretical analysis of structure-generalized parametric BSSA

3.1. Signal modeling

In this paper, we assume that the input signal x in the power spectral domain is modeled using the gamma distribution as

$$P_{\mathrm{GM}}(x) = \Gamma(\alpha)^{-1}\theta^{-\alpha}\cdot x^{\alpha-1}\exp\left(-x/\theta\right), \tag{12}$$

where $x \geq 0, \alpha > 0$, and $\theta > 0$. Here, α is the shape parameter, θ is the scale parameter, and $\Gamma(\alpha)$ is the gamma function.

3.2. Analysis of amount of musical noise

3.2.1. Metric of musical noise generation: kurtosis ratio

In this study, we apply the *kurtosis ratio* to a *noise-only time-frequency period* of subject signal for the assessment of musical noise [7]. This measure is defined as

$$\text{kurtosis ratio} = \mathrm{kurt}_{\mathrm{proc}}/\mathrm{kurt}_{\mathrm{org}}, \tag{13}$$

where $\mathrm{kurt}_{\mathrm{proc}}$ is the kurtosis of the processed signal and $\mathrm{kurt}_{\mathrm{org}}$ is the kurtosis of the observed signal. Kurtosis is defined as

$$\mathrm{kurt} = \mu_4/\mu_2^2, \tag{14}$$

where μ_m is the mth-order moment, as

$$\mu_m = \int_0^\infty x^m P(x)dx, \tag{15}$$

where $P(x)$ is the probability density function of a power-spectral-domain signal x. Note that μ_m is not a central moment but a raw moment. Thus, (14) is not kurtosis in the mathematically strict definition but a modified version; however, we still refer to (14) as kurtosis in this paper. This measure increases as the amount of generated musical noise increases.

3.2.2. Analysis in the case of parametric BSSA

In this section, we analyze the kurtosis ratio in parametric BSSA. First, using the shape parameter of input noise α_{n}, we express the kurtosis of a gamma distribution, $\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})}$, [7] as

$$\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})} = (\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)/\alpha_{\mathrm{n}}/(\alpha_{\mathrm{n}}+1). \tag{16}$$

The power spectral domain kurtosis after DS is given by [5]

$$\mathrm{kurt}_{\mathrm{DS}}^{(\mathrm{n})} \simeq J^{-0.7}\cdot(\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})} - 6) + 6. \tag{17}$$

Next, we calculate the amount of kurtosis change after parametric BSSA. With the shape parameter after DS, α_{DS} [5], the resultant kurtosis after parametric BSSA is represented as [8]

$$\mathrm{kurt}_{\mathrm{BSSA}}^{(\mathrm{n})} = \mathcal{M}(\alpha_{\mathrm{DS}},\beta,4,n)/\mathcal{M}^2(\alpha_{\mathrm{DS}},\beta,2,n), \tag{18}$$

where $\mathcal{M}(\alpha,\beta,m,n)$ can be expressed as [8]

$$\mathcal{M}(\alpha,\beta,m,n) = \sum_{l=0}^{m/n}\frac{(-\beta)^l\Gamma^l(\alpha+n)\Gamma(m/n+1)}{\Gamma^{l+1}(\alpha)\Gamma(l+1)\Gamma(m/n-l+1)} \Gamma\left(\alpha+m-nl, \left(\beta\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\right)^{\frac{1}{n}}\right), \tag{19}$$

where $\Gamma(\alpha,z)$ is the upper incomplete gamma function as

$$\Gamma(\alpha,z) = \int_z^\infty t^{\alpha-1}\exp(-t)dt. \tag{20}$$

Finally, using (13), (16), and (18), we can determine the resultant kurtosis ratio through parametric BSSA as

$$\text{kurtosis ratio}_{\mathrm{BSSA}}^{(\mathrm{n})} = \mathrm{kurt}_{\mathrm{BSSA}}^{(\mathrm{n})}/\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})}. \tag{21}$$

3.2.3. Analysis in the case of parametric chBSSA

In this section, we analyze the kurtosis ratio in parametric chBSSA. First, we calculate the amount of kurtosis change after channelwise GSS. Using (18) with the shape parameter of input noise α_n, we can express the resultant kurtosis after channelwise GSS as

$$\mathrm{kurt}^{(n)}_{\mathrm{chGSS}} = \mathcal{M}(\alpha_n, \beta, 4, n)/\mathcal{M}^2(\alpha_n, \beta, 2, n). \quad (22)$$

Next, using (17) and (22), we can derive the amount of kurtosis change after parametric chBSSA as

$$\mathrm{kurt}^{(n)}_{\mathrm{chBSSA}} \simeq J^{-0.7} \cdot (\mathrm{kurt}^{(n)}_{\mathrm{chGSS}} - 6) + 6. \quad (23)$$

Finally, we can obtain the resultant kurtosis ratio through parametric chBSSA as

$$\text{kurtosis ratio}^{(n)}_{\mathrm{chBSSA}} = \mathrm{kurt}^{(n)}_{\mathrm{chBSSA}} \,/\, \mathrm{kurt}^{(n)}_{\mathrm{in}}. \quad (24)$$

3.3. Analysis of amount of speech distortion

3.3.1. Analysis in the case of BSSA

In this section, we analyze the amount of speech distortion on the basis of kurtosis ratio in speech components. Hereafter, we define $s(f,\tau)$ and $n(f,\tau)$ as the observed speech and noise component at each microphone. Assuming that speech and noise are disjoint, i.e., there are no overlaps in the time-frequency domain, speech distortion is caused by subtracting the average of noise from the pure speech component. Thus, the distorted speech after BSSA is given by

$$\begin{aligned} |s_{\mathrm{BSSA}}(f,\tau)| &= \sqrt[2n]{|s(f,\tau)|^{2n} - \beta\overline{|z_{\mathrm{DS}}(f,\tau)|^{2n}}} \\ &= \sqrt[2n]{|s(f,\tau)|^{2n} - \beta C_{\mathrm{BSSA}}\overline{|s(f,\tau)|^{2n}}}, \end{aligned} \quad (25)$$

where $s_{\mathrm{BSSA}}(f,\tau)$ is the output speech component in BSSA. Also, calculating the nth moment of gamma distribution, C_{BSSA} is given by

$$\begin{aligned} C_{\mathrm{BSSA}} &= \overline{|z_{\mathrm{DS}}(f,\tau)|^{2n}}/\overline{|s(f,\tau)|^{2n}} \\ &= J^{-n}\overline{|n(f,\tau)|^{2n}}/\overline{|s(f,\tau)|^{2n}} \\ &= J^{-n}\left(\frac{\alpha_s}{\alpha_n}\right)^n \frac{\Gamma(\alpha_n+n)/\Gamma(\alpha_n)}{\Gamma(\alpha_s+n)/\Gamma(\alpha_s)} \left(\frac{\overline{|n(f,\tau)|^2}}{\overline{|s(f,\tau)|^2}}\right)^n, \end{aligned} \quad (26)$$

where α_s is the shape parameter of input speech. Equation (26) indicates that the speech distortion increases when the input SNR, $\overline{|s(f,\tau)|^2}/\overline{|n(f,\tau)|^2}$, and/or the number of microphones, J, decrease. Using (19), and (26) with the input speech shape parameter α_s, we can obtain the speech kurtosis ratio after BSSA as

$$\begin{aligned} &\text{kurtosis ratio}^{(s)}_{\mathrm{BSSA}} \\ &= \frac{\mathcal{M}(\alpha_s, \beta C_{\mathrm{BSSA}}, 4, n)}{\mathcal{M}^2(\alpha_s, \beta C_{\mathrm{BSSA}}, 2, n)} \frac{\alpha_s(\alpha_s+1)}{(\alpha_s+2)(\alpha_s+3)}. \end{aligned} \quad (27)$$

3.3.2. Analysis in the case of chBSSA

In chBSSA, since channelwise GSS is performed before DS, C_{BSSA} is therefore replaced with

$$\begin{aligned} C_{\mathrm{chBSSA}} &= (\overline{|n(f,\tau)|^{2n}}/\overline{|s(f,\tau)|^{2n}}) \\ &= \left(\frac{\alpha_s}{\alpha_n}\right)^n \frac{\Gamma(\alpha_n+n)/\Gamma(\alpha_n)}{\Gamma(\alpha_s+n)/\Gamma(\alpha_s)} \left(\frac{\overline{|n(f,\tau)|^2}}{\overline{|s(f,\tau)|^2}}\right)^n. \end{aligned} \quad (28)$$

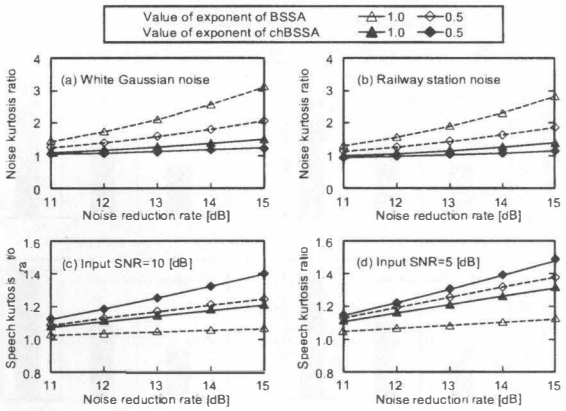

Figure 2: (a) and (b) are theoretical behaviors of noise kurtosis ratio in structure-generalized parametric BSSA. (a) is for white Gaussian noise and (b) is for railway station noise. (c) and (d) are theoretical behaviors of speech kurtosis ratio in structure-generalized parametric BSSA. (c) and (d) are set to 10 and 5 dB.

Equation (28) indicates that the speech distortion increases only when the input SNR decreases, regardless of the number of microphones. Thus, the distortion did not change even if we prepare many microphones, unlike the case of parametric BSSA. Using (19) and (28) with the shape parameter of input speech α_s, we can obtain the speech kurtosis ratio after chBSSA as

$$\begin{aligned} &\text{kurtosis ratio}^{(s)}_{\mathrm{chBSSA}} \\ &= \frac{\mathcal{M}(\alpha_s, \beta C_{\mathrm{chBSSA}}, 4, n)}{\mathcal{M}^2(\alpha_s, \beta C_{\mathrm{chBSSA}}, 2, n)} \frac{\alpha_s(\alpha_s+1)}{(\alpha_s+2)(\alpha_s+3)}. \end{aligned} \quad (29)$$

3.4. Comparison of amount of musical noise and speech distortion under same amount of noise reduction

According to the previous analysis, we can compare the amount of musical noise and speech distortion among parametric BSSA and parametric chBSSA under the same noise reduction rate (NRR) [3] (output SNR - input SNR in dB). Figure 2 shows the theoretical behaviors of the noise kurtosis ratio and speech kurtosis ratio. In Figs. 2(a) and 2(b), the shape parameter of input noise, α_n, is set to 0.95 and 0.83 corresponding to almost white Gaussian noise and railway station noise, respectively. Also, in Figs. 2(c) and 2(d), the shape parameter of input speech, α_s, is set to 0.1, and the input SNR is set to 10 and 5 dB, respectively. Here, we assume an eight-element array with the interelement spacing of 2.15 cm. The NRR is varied from 11 to 15 dB, and the oversubtraction parameter β is adjusted so that the target speech NRR is achieved. In parametric BSSA and parametric chBSSA, the signal exponent parameter $2n$ is set to 1.0 and 0.5.

Figures 2(a) and 2(b) indicate that the noise kurtosis ratio of chBSSA is smaller than that of BSSA, i.e., musical noise generation is lower in parametric chBSSA than in parametric BSSA, and a small amount of musical noise is generated when a lower exponent parameter is used, regardless of the type of noise and NRR. However, Figs. 2(c) and 2(d) show that speech distortion is lower in parametric BSSA than in parametric chBSSA, and a small amount of speech distortion is generated when a higher exponent parameter is used, regardless of the type of noise and NRR. These results theoretically prove a tradeoff between the amount of musical noise and speech distortion in BSSA and chBSSA.

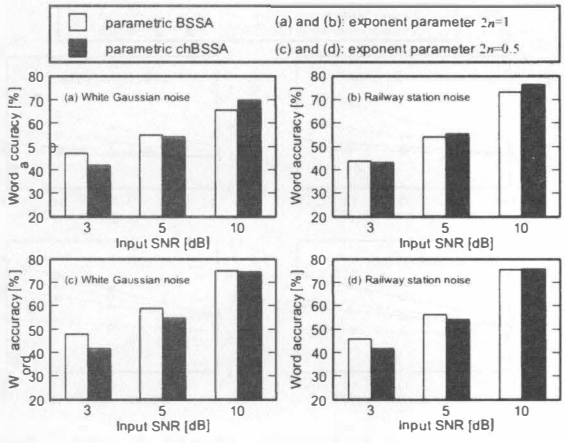

Figure 3: Results of word accuracy in parametric BSSA and parametric chBSSA.

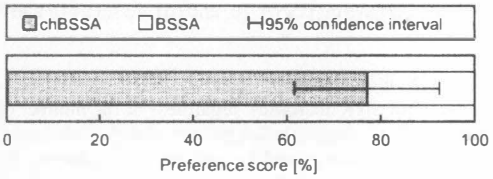

Figure 4: Subjective evaluation results: BSSA vs. chBSSA.

4. Experiment

4.1. Experimental setup

In this study, we conducted a speech recognition experiment. We used an eight-element microphone array with an interelement spacing of 2.15 cm, and the direction of the target speech is set to be normal to the array. The size of the experimental room is $4.2 \times 3.5 \times 3.0$ m^3, and the reverberation time is approximately 200 ms. All the signals used in this experiment are 16-kHz-sampled signals with 16-bit accuracy. The observed signal consists of the target speech signal of 200 speakers (100 males and 100 females) and two types of diffuse noise (white Gaussian noise and railway station noise) emitted from eight surrounding loudspeakers. The input SNR of test data is set to 3, 5, and 10 dB. The FFT size is 1024, and the frame shift length is 256 in BSSA. The speech recognition task is a 20-k-word Japanese newspaper dictation, where we used Julius 3.4.2 [9] as the speech decoder. The acoustic model is a phonetic-tied mixture [9], and we use 260 speakers (150 sentences/speaker) for training the acoustic model. In this experiment, the NRR, target SNR improvement, is set to 10 dB for white Gaussian noise and 5 dB for railway station noise, the exponent parameter $2n$ is set to 1.0 and 0.5, and the oversubtraction parameter β is adjusted so that the target NRR is achieved.

4.2. Evaluation of speech recognition performance and discussion

Figure 3 shows the results of word accuracy in parametric BSSA and parametric chBSSA. This result reveals that a better speech recognition performance can be obtained in parametric BSSA when the input SNR is low (e.g., 3 dB).

This result is of considerable interest because Takahashi et al., [5] mentioned a contradictory result, i.e., chBSSA's sound quality is always superior to that of BSSA. Indeed, we conducted a subjective evaluation. We presented 56 pairs of signals processed by parametric BSSA and parametric chBSSA, selected from sentences used in the speech recognition experiment, in random order to 10 examinees, who selected which signal they preferred. The result is shown in Fig. 4, confirming that chBSSA is preferred in human perception unlike the speech recognition results. This is partially true regarding noise distortion, i.e., the amount of musical noise generation, as shown in Figs. 2(a) and 2(b). Thus, human impression is mostly affected by noise distortion.

However, as proved in Figs. 2(c) and 2(d), the speech distortion in chBSSA is larger than that in BSSA: this leads to the degradation of speech recognition performance. In summary, we should carefully select the structure of signal processing in BSSA, i.e., chBSSA is recommended for listening but BSSA is suitable for speech recognition under low input SNR condition.

5. Conclusions

In this paper, we introduced the structure-generalized parametric BSSA and performed its theoretical analysis via higher-order statistics. Comparing parametric BSSA and parametric chBSSA, we reveal that a channelwise BSSA structure is recommended for listening but a normal BSSA is more suitable for speech recognition.

6. Acknowledgements

This work was supported by the MIC SCOPE, and JST Core Research of Evolutional Science and Technology (CREST), Japan.

7. References

[1] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech and Audio Process.*, vol.13, no.1, pp.120–134, 2005.

[2] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol.14, no.6, pp.2165–2173, 2006.

[3] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol.17, no.4, pp.650–664, 2009.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol.ASSP-27, no.2, pp.113–120, 1979.

[5] Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musical-noise analysis in methods of integrating microphone array and spectral subtraction based on higher-order statistics," *EURASIP Journal on Advances in Signal Process.*, vol.2010, Article ID 431347, 25 pages, 2010.

[6] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech and Audio Process.*, vol.6, no.4, pp.328–337, 1998.

[7] 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," *Proc. IWAENC*, 2008.

[8] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher-order statistics," *IEEE Trans. Audio, Speech and Lang. Process.*, in printing (DOI: 10.1109/TASL.2010.2098871).

[9] A. Lee, T. Kawahara, and K. Shikano, "Julius -An open source real-time large vocabulary recognition engine," *Proc. Eurospeech*, pp.1691–1694, 2001.