Optimization Scheme of Joint Noise Suppression and Dereverberation Based on Higher-Order Statistics

Fine Dwinita Aprilyanti*, Hiroshi Saruwatari*, Kiyohiro Shikano* and Tomoya Takatani†
*Nara Institute of Science and Technology, Nara, Japan
E-mail: dwinita-a@is.naist.jp Tel: +81-743-72-5287
†Toyota Motor Corporation, Aichi, Japan
E-mail: tomoya_takatani@mail.toyota.co.jp Tel: +81-565-98-6462

***Abstract*— In this paper, we apply the higher-order statistics parameter to automatically improve the performance of blind speech enhancement. Recently, a method to suppress both diffuse background noise and late reverberation part of speech has been proposed combining blind signal extraction and Wiener filtering. However, this method requires a good strategy for choosing the set of its parameters in order to achieve the optimum result and to control the amount of musical noise, which is a common problem in non-linear signal processing. We present an optimization scheme to control the value of Wiener filter coefficients used in this method, which depends on the amount of musical noise generated, measured by higher-order statistics. The noise reduction rate and cepstral distortion are also evaluated to confirm the effectiveness of this scheme.**

I. INTRODUCTION

A hands-free speech recognition system provides the natural and efficient human-machine interaction. On the other hand, since this system often uses a microphone array to pick up the target speech signal at a distance, its performance is limited by the adverse effect of background noise and reverberation. Therefore, a method that can suppress the interference sound is required to improve the speech quality.

Many microphone array techniques have been studied to solve this problem. The method proposed in [1] has shown that frequency-domain independent component analysis (FD-ICA) can estimate the diffuse background noise component better than the target speech one, thus combining FD-ICA as noise estimator and spectral subtraction (SS) as nonlinear postprocessing has been proved to be effective in improving the target sound quality. However, this method does not suppress the effect of reverberation.

It is known that there are two parts of reverberation: the early and late reverberation. While the early reverberation is considered harmless to speech intelligibility, the late part can deteriorate the sound quality, depending on the length and strength of this reverberation [2]. Some of the authors have proposed a method that jointly suppresses diffuse background noise and the late reverberation part of the speech signal [3] (hereafter this method is referred to as *joint method*). This method is based on frequency domain blind signal extraction (FD-BSE) [4] combined with two stages of multichannel Wiener filtering (WF) as nonlinear postfilters. This method uses *a priori* knowledge of the room reverberation time (T_{60}) to synthesize the late reverberation part. Other researchers have published similar approach using different BSS algorithm [5], but although the method does not required estimation of T_{60}, it does not consider the effect of background noise. There is also related research using linear prediction (LP) analysis combined with SS for suppressing late reverberation and WF to compensate background noise [6]. This method requires pre-recorded handclaps to estimate the late reverberation component and noise level.

Another problem that arises in the nonlinear signal processing technique is artificial distortion called *musical noise*. Some of the authors have reported that the amount of musical noise generated is strongly correlated with the difference between the higher-order statistics of the power spectra before and after nonlinear signal processing [7]. The theoretical analysis of the amount of musical noise generated in various types of WF has also been presented in [8].

The joint method proposed in [3] has been proved to be effective to improve the performance of automatic speech recognition system in term of word accuracy. However, some parameter values are still chosen manually and the generation of musical noise has not been considered. Therefore, in this paper, motivated by our previous works [7, 8] on musical noise assessment, an optimization scheme is presented, which can automatically choose the best set of parameters of the joint method. Also, this assessment of musical noise generated can be performed blindly, and consequently our proposed system can be driven in the whole blind fashion. We also conduct objective evaluations to confirm the effectiveness of this scheme.

This paper is organized as follows. In Section II, we describe the related works on the joint method. In Section III, we present the overview of the proposed scheme including the assessment algorithm of musical noise generated via the higher-order statistics. We show the experimental results and evaluations in Section IV. Finally, conclusions are provided in Section V.

II. RELATED WORKS

A. FD-BSE-Based Joint Noise Suppression and Dereverberation [3]

In this subsection, we explain our previously proposed joint method for noise reduction and dereverberation based on FD-BSE and WF postprocessing. The architecture of this method

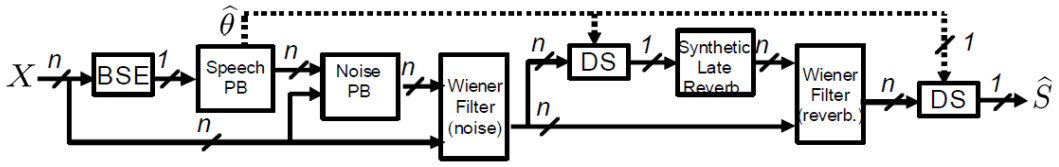

Fig. 1 Block diagram of joint blind noise suppression and dereverberation method.

is shown in Fig. 1. In this system, the microphone array with m sensors is utilized to acquire the multichannel observations. The observed signal of m components, $x(t)$, is the superposition of the speech contribution $x_S(t)$ and noise contribution $n(t)$, as given by

$$x(t) = x_S(t) + n(t), \quad (1)$$
$$x_S(t) = \big(h_E(\tau) + h_L(\tau)\big) * s(t), \quad (2)$$

where $s(t)$, $h_E(\tau)$ and $h_L(\tau)$ are the clean speech source, early and late parts of the room impulse response, respectively. Most hidden-Markov-model (HMM)-based speech recognizers are capable to handle the effect of $h_E(\tau)$ up to certain time delay τ_d, while the effect of $h_L(\tau)$ must be handled by array signal processing.

This method uses the new FD-BSE proposed in [4] that utilizes the sparsity of the modulus of the target speech signal to estimate the diffuse background noise component. In frequency domain, the estimate $Y(f,t)$ is obtained by applying extracting vector to the observed signal

$$Y(f,t) = W(f)X(f,t). \quad (3)$$

The vector $W(f)$ is updated using gradient decent method to minimize the cost function $J\big(W(f)\big)$ that based on signal modulus sparsity, given by

$$J\big(W(f)\big) = \tfrac{1}{2} E\{|Y(f,t)|\}^2, \quad (4)$$
$$E\{|Y(f,t)|^2\} = 1\ . \quad (5)$$

The estimated noise $N(f,t)$ is obtained by applying projection back to the separated output vector $Y^{(\text{noise})}(f,t)$ that contains only noise components (speech component is substituted with zero)

$$\widehat{N}(f,t) = W^{-1}(f)Y^{(noise)}(f,t). \quad (6)$$

After FD-BSE, a set of multichannel WF is applied to suppress the estimated noise components, as given by

$$\widehat{X_S}(f,t) = G|X(f,t)|e^{jarg(X(f,t))}, \quad (7)$$
$$G = \frac{|X(f,t)|^2}{|X(f,t)|^2 + \beta_N|\widehat{N}(f,t)|^2}\ , \quad (8)$$

where β_N is a parameter for controlling the strength of noise suppression.

The late reverberation speech part is synthesized afterwards from the output of WF under assumption that the noise suppression was efficient. A synthetic late reverberation filter is estimated using the information of T_{60} of the room, and the late reverberation part is suppressed in the same manner with noise suppression stage, as represented by

$$\hat{S}(f,t) = G\big|\widehat{X_S}(f,t)\big|e^{jarg(\widehat{X_S}(f,t))}, \quad (9)$$
$$G = \frac{\big|\widehat{X_S}(f,t)\big|^2}{\big|\widehat{X_S}(f,t)\big|^2 + \beta_R\big|\widehat{X_L}(f,t)\big|^2}\ , \quad (10)$$

where β_R is a parameter for controlling the strength of dereverberation.

After two stages of WF, the delay-and-sum beamformer (DS) is applied to the signal, using information of the estimated DOA $\hat{\theta}$ from the speech projection back. Therefore, there are at least three parameters that should be estimated: WF coefficient for noise suppression stage (β_N), WF coefficient for dereverberation stage (β_R), and T_{60}. But the main problem in this method is that there are so many combinations of the parameters and we cannot know, in advance, which set of the parameters will lead to the best result.

III. Proposed Optimization Scheme

A. Problem and Motivation

In the conventional method described in Sect. II, it is of great interest to realize a good optimization scheme for the internal parameters, namely, β_N, β_R and T_{60}. The correct setting of these parameters will give optimum performance. However, we have to manually optimize and there is no efficient method to control the parameters automatically. Therefore, in this section, we propose a new optimization scheme for the parameters. In particular, we introduce higher-order statistics in the system.

Motivated by our previous study on relation between the higher-order statistics of power spectra and musical noise perception, we newly utilize *kurtosis,* the 4th-order-moment based statistics, for the assessment of musical noise generated after nonlinear postprocessing in WF parts. The overview of the proposed assessment is as follows.

1. Detect the speech-pause time period (noise-only time period) by comparing the temporal noise power estimated by FD-BSE in (6) and that of observed signal.
2. In the detected speech-pause period, we can estimate the amount of generation of musical noise based on the kurtosis ratio.

In the following subsections, we describe the details of the algorithm.

B. Speech-Pause Detection Based on FD-BSE

The authors of [1] have proved that under the diffuse background noise condition, an ICA method described by (3) performs better as noise estimator than as target speech estimator. Assuming the estimation of noise is efficient, we can detect the speech activity within observed signal using the short-time power of noise estimation, given by

$$\sigma_N{}^{2(i)} = \frac{1}{M}\sum_{t\in T_i}\sum_{f\in F}\widehat{N}(f,t)^2\,, \tag{11}$$

where $\sigma_N{}^{2(i)}$ is the power of noise estimation in the *i*th time frame where signal can be considered as stationary (about 25 ms). According to (1), signal power will become higher if there is speech part included. Thus, by setting the maximum noise estimation power as threshold, we can divide the observed signal into two categories: time frames with power more than threshold will be regarded as speech-active frames, and the rest as speech-pause frames.

Note that the estimation of noise may include some late reverberation, which we do not need in assessment of musical noise based on higher-order statistics. Therefore, we only select noise frames that precede the first speech-active frame.

C. Musical Noise Analysis via Higher Order Statistics

The amount of musical noise is highly correlated to the number of isolated power spectral components and their level of isolation. These isolated components are called *tonal components*. Since such tonal components have relatively high power, they are strongly related to the weight of the tail of their probability density function (pdf). Therefore, quantifying the tail of the pdf makes it possible to measure the number of tonal components. Kurtosis has been introduced in our studies to evaluate the tail of the pdf [9, 10], successfully showing the effectiveness of using the higher-order statistics.

The assessment of musical noise is done by applying the *kurtosis ratio* to noise-only part (speech-pause time period detected using (11)) of the observed signal [7]. This measure is defined as

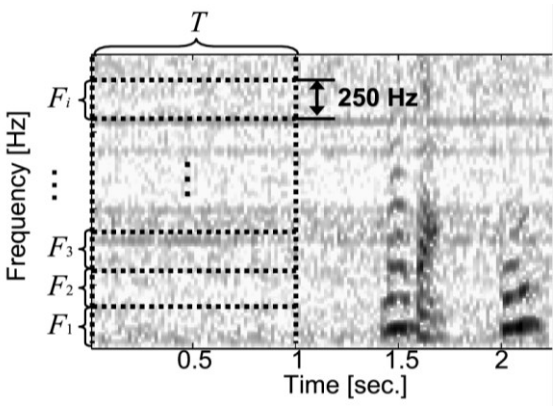

Fig. 2 Time-frequency domain noise-only signal for calculation of subband kurtosis.

$$\textit{kurtosis ratio} = \mathrm{kurt}_{\mathrm{proc}}/\mathrm{kurt}_{\mathrm{org}}\,, \tag{12}$$

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

$$\mathrm{kurt} = \mu_4/\mu_2^2, \tag{13}$$

where μ_n is the *n*th-order moment, given by

$$\mu_n = \int_0^\infty x^n P(x)dx\,, \tag{14}$$

where $P(x)$ is the probability density function of a signal x.

In this paper, we measure the amount of musical noise generated by calculating the frequency subband-wise kurtosis [8] as given by

$$\mathrm{kurt}^{(i)} = \frac{(1/M)\sum_{f\in F_i}\sum_{t\in T}(|X(f,t)|^2)^4}{\{(1/M)\sum_{f\in F_i}\sum_{t\in T}(|X(f,t)|^2)^2\}^2} \tag{15}$$

where $\mathrm{kurt}^{(i)}$ is the *i*th subband kurtosis of a signal x. F_i and T represent subband time-frequency grid indexes to be

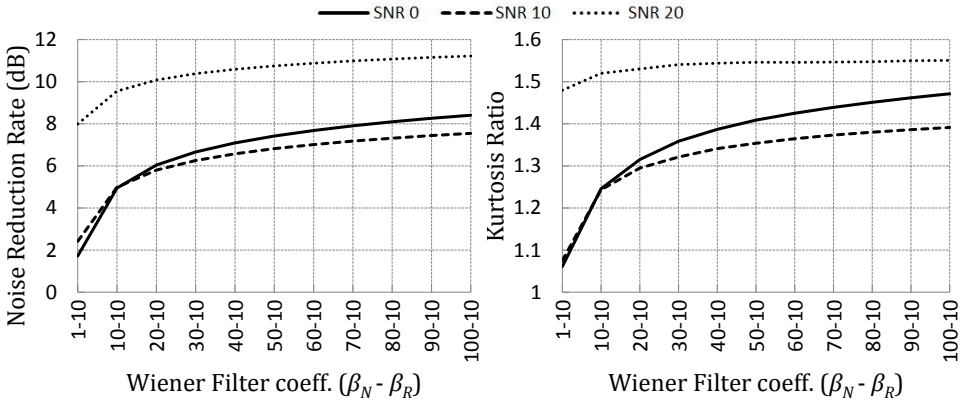

Fig. 3 Noise Reduction Rate and kurtosis ratio behavior of joint method.

evaluated, which have been detected in (11), while M is the total number of grids in each subband. We use 250-Hz-width F_i and T of 5 s, which are taken from noise-only time-frequency region preceding a speech utterance. See Fig. 2 for the example of the subband procedure.

D. Objective Evaluations of Joint Method

The previous work only evaluates the word accuracy rate of speech recognition system. Also, the combination of parameters value is set manually by simulation of several combinations and the ones who give optimum result is chosen. In this subsection, we study the effect of choice of parameters to the sound quality of the input signal.

We use kurtosis ratio, noise reduction rate (NRR) and cepstral distortion (CD) as objective evaluation measures. NRR is defined as the difference of the SNR before and after processing. The SNR of a signal is represented by

$$\mathrm{SNR} = 10\log_{10}\frac{\mathrm{E}[S(t)]^2}{\mathrm{E}[N(t)]^2}, \qquad (16)$$

where $S(t)$ and $N(t)$ are the speech and noise component of signals, respectively. Since high SNR indicates good signal quality, high NRR result is preferable.

CD is a measure of how the target signal is distorted after processing, define as

$$CD = (20/\ln 10)\sqrt{2\sum_{k=1}^{D}(\widehat{C_k} - C_k)^2}, \qquad (17)$$

where C_k is the kth cepstrum coefficient. To compensate the early reverberation speech part that was left unprocessed in this method, the signal convoluted with early room impulse response is used as reference signal.

Fig. 3 shows the behavior of kurtosis ratio for joint method. As one can expect, large value of β_N will suppress large amount of noise, thus increase the NRR. However, the kurtosis ratio value is also increased, which means more amount of musical noise generated. Note that for signal with high level of SNR (in this case is 20 dB), small value of WF coefficients already results in high NRR and kurtosis ratio. This is logical since the amount of interference signals to be suppressed are very low. Since both of NRR and kurtosis ratio increase with the increasing WF coefficient, we can maximize the NRR automatically by setting the largest possible WF coefficient under a certain kurtosis ratio constraint. Note that after some point the increase of NRR and kurtosis ratio become slower and insignificant. By assuming that this indicates the noise suppression has been optimum, we can also limit the processing when the difference between the updated value and the previous value of kurtosis ratio has reached a certain threshold.

The behavior of CD and the effect of T_{60} value choice is shown in Fig. 4. We can see that the CD value decreases as the NRR increases. However, at some points, it starts to increase. This happens because the late reverberation distorts the cepstrum of the original signal, causing a signal with strong reverberation to have high CD. WF suppresses the late

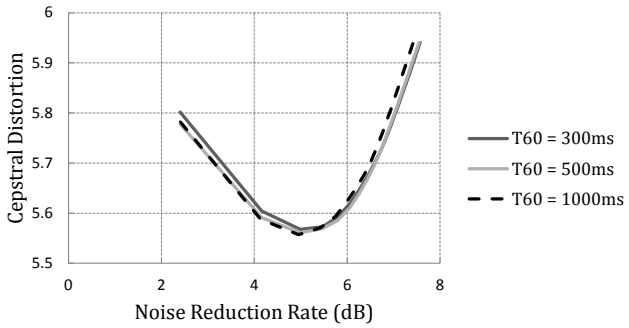

Fig. 4 Cepstral distortion behavior and effect of choice of T_{60} to input signal with SNR of 10 dB

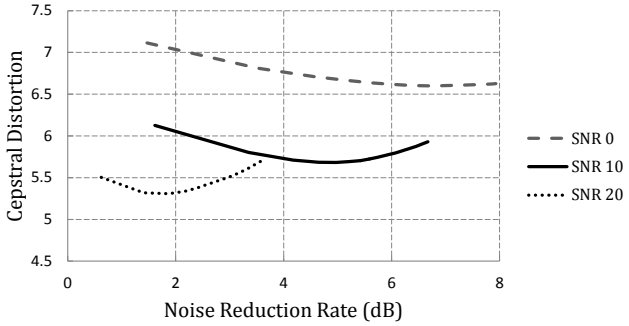

Fig. 5 Cepstral distortion behavior of noise suppression stage.

reverberation part, thus results in lower CD. On the other hand, if the WF coefficient is too strong, not only the late reverberation part but also the original speech part will be suppressed. Furthermore, it is also shown that the value of T_{60} does not significantly affect the result. This is possible because the effect of mismatched T_{60} can be compensated by the choice of the WF coefficient parameter.

While theoretically the signal processing is done in two stages, in practice the WF in noise suppression stage also acts as dereverberation filter, as can be seen in Fig. 5. This is due to the mixing model of signals used in FD-BSE which includes some amount of late reverberation in noise [3]. On the other hand, the WF part in the dereverberation stage also partly acts as a noise suppression filter, indicated by increased NRR corresponding to the increasing β_R. According to these characteristics, we choose to optimize the WF coefficient separately.

E. Optimization of WF Coefficient

Based on the objective evaluation results from the joint method, we can conclude that setting appropriate value of β_N and β_R is more important than T_{60}. Therefore, we focus on optimization of these parameters and set T_{60} to be fixed value. We use the amount of musical noise generated, corresponds to kurtosis ratio as constraints. The procedure is described as follow.

Step 0: First, set initial β_N and β_R to a low value.

Step1: Next, apply the WF for noise suppression to the input signal (after BSE processing) using the value of β_N.

TABLE I
OBJECTIVE EVALUATIONS FOR DIFFERENT METHODS

SNR (dB)	0 dB														
distance (m)	1 m			2 m			3 m			4 m			5 m		
	BSE	BSE + w	opt	BSE	BSE + w	opt	BSE	BSE + w	opt	BSE	BSE + w	opt	BSE	BSE + w	opt
NRR (dB)	4.67	8.27	**9.05**	1.79	7.42	**7.62**	4.02	6.13	**6.17**	2.01	6.90	**8.06**	4.73	7.41	**8.06**
CD	6.26	6.41	**5.96**	8.15	6.34	**6.31**	6.51	6.48	**6.23**	8.23	6.32	**6.01**	6.48	6.61	**6.36**
KR	1.18	1.36	1.39	1.16	1.31	1.51	1.16	1.37	1.45	1.23	1.36	1.51	1.22	1.42	1.39
SNR (dB)	10 dB														
distance (m)	1 m			2 m			3 m			4 m			5 m		
	BSE	BSE + w	opt	BSE	BSE + w	opt	BSE	BSE + w	opt	BSE	BSE + w	opt	BSE	BSE + w	opt
NRR (dB)	5.37	**10.20**	10.05	4.56	**12.28**	9.44	3.73	7.19	**8.02**	2.75	7.13	**7.71**	3.20	6.18	**7.06**
CD	**4.70**	4.91	4.75	**4.91**	5.39	5.06	**5.08**	5.37	5.21	5.09	5.33	**5.06**	**5.40**	5.77	5.71
KR	1.15	1.32	1.35	1.16	1.37	1.38	1.13	1.30	1.36	1.14	1.28	1.34	1.19	1.33	1.48

TABLE II
OBJECTIVE EVALUATIONS OF DIFFERENT METHODS FOR INPUT SIGNALS WITH SNR 20 DB

SNR (dB)	20 dB																			
distance (m)	1 m				2 m				3 m				4 m				5 m			
	BSE	BSE + w	BSE + r	opt	BSE	BSE + w	BSE + r	opt	BSE	BSE + w	BSE + r	opt	BSE	BSE + w	BSE + r	opt	BSE	BSE + w	BSE + r	opt
NRR (dB)	2.08	6.27	0.68	**6.39**	1.55	5.68	0.68	**6.15**	1.64	1.15	1.82	**2.98**	0.48	0.95	1.38	**2.02**	(6.32)	2.73	7.24	**10.42**
CD	**3.69**	3.94	4.06	3.90	**4.22**	4.42	4.40	4.39	4.65	4.75	4.44	**4.25**	4.87	4.88	4.60	**4.53**	6.96	5.42	**4.83**	5.30
KR	1.05	1.17	1.03	1.18	1.04	1.15	1.03	1.17	1.02	1.01	1.05	1.07	1.02	1.02	1.07	1.16	1.22	1.13	1.50	1.54

Step 2: Apply DS beamformer to the output signal, then calculate the kurtosis using (4). Obtain kurtosis ratio by dividing kurtosis of the output signal by the kurtosis of observed signal.

Step 3: Increase the value of β_N by a certain amount $\Delta_{\beta N}$. Return to Step 1 until the kurtosis ratio value reaches the given limit, or until the difference between updated value and previous value of kurtosis ratio is below certain threshold.

Step 4: Apply the WF for dereverberation to the output signal of noise suppression stage (before DS beamformer is applied), and update the β_R in the same manner as updating the β_N.

IV. EXPERIMENTAL EVALUATIONS

A. Experimental Setup

We used a 20K-word Japanese dictation from the database JNAS [11] as the source signals, with Julius 4.2 [12] as the decoder for recognition task. An eight-channel microphone array (inter microphone spacing of 2.0 cm) was used to record the diffuse background noise from the railway station, and room impulse response at various distance between microphone and speaker at a large lecture room. The estimated T_{60} value is 500 ms.

The clean speech with sampling frequency of 16000 Hz is convoluted with room impulse response, and was mixed with noise signal at SNR 0 dB, 10 dB and 20 dB. For the time-frequency domain processing, the short time Fourier transform is implemented with 1024 point FFT size, hanning window, and 50% overlap. The separation is performed by 600 iterations of BSE method with adaptation step of 0.3. For every 200 iterations, the adaptation step will be updated into half of its current value. For dereverberation stage, the τ_d value is set to 75 ms. This corresponds to the effect of room impulse response that still can be handled by most HMM-based speech recognizer.

We compare the output signals of proposed scheme with the speech estimation from FD-BSE and the output of FD-BSE with noise suppression WF proposed in [4]. For input signals with high level of SNR, we also compare the performance of the proposed scheme with the simple

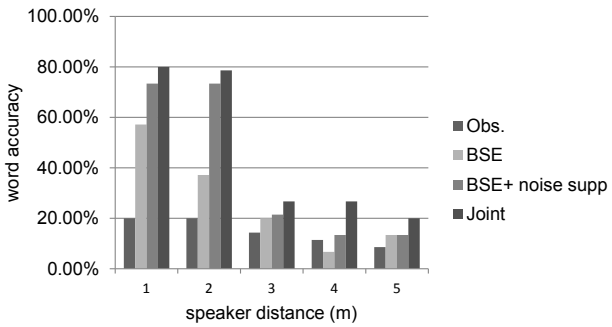

Fig. 6 Recognition result of input signals with SNR 10 dB

combination of FD-BSE and dereverberation. This is done under assumption that for high SNR signals, the late reverberation is stronger than the background noise, therefore we would like to seek for the possibility to skip the noise suppression stage. In this method, the late reverberation speech part is synthesized from the speech estimation of FD-BSE, assuming that even if the separation process is not effective, WF in dereverberation stage still can suppress some amount of noise according to our previous study.

B. Experimental Results

The comparison of objective evaluation results for input signals with SNR level 0 dB and 10 dB is given in Table I. The results from FD-BSE are displayed in column '**BSE**', while the objective evaluations of FD-BSE with noise suppression WF are displayed in column '**BSE + w**' and the proposed scheme in column '**opt**'. The 'NRR', 'CD', and 'KR' indicate noise reduction rate, cepstral distortion, and kurtosis ratio, respectively.

For input signal with SNR level of 0 dB, the optimized method gives the best performance in each distance, indicated by higher NRR and lower CD than that of other methods. However, the results for input signal with SNR level of 10 dB are varied. While for far speaker distances, the proposed scheme outperforms other method, for close speaker distances FD-BSE with noise suppression WF shows better performance in terms of NRR. But in terms of CD, the FD-BSE gives best results in almost every condition. This may be because the value of β_N is too strong that the original speech is already distorted.

Table II shows the performance comparison for input signals with SNR level 20 dB. In addition to method displayed in Table I, the objective evaluation results of FD-BSE with dereverberation WF are displayed in column '**BSE + r**'. Again, the proposed scheme gives the best performance in every condition in terms of NRR. However, the results of CD are varied correspond to the speaker distance. For close speaker distances, the FD-BSE still gives better CD than other method, while for far speaker distances, the proposed scheme and FD-BSE with dereverberation WF give the best results.

The word accuracy results with comparison to other method are shown in Fig. 6. In this figure, 'Obs.' represents the observed signal that is left unprocessed, while 'BSE', 'BSE + noise supp.', and 'Joint' are speech estimation from FD-BSE, output of FD-BSE with noise suppression stage, and output of the proposed scheme, respectively. It is shown that our proposed method improved the recognition performance more than other methods.

Although in almost all condition the proposed scheme gives higher KR compare to other method, it will not have significant effect in output quality, since we have set the limit that minimalizes the amount of generated musical noise.

V. CONCLUSIONS

In this paper, we proposed an optimization scheme to obtain the best combination of parameters used in BSE and WF based joint suppression of noise and late reverberation. We confirmed the effectiveness of this scheme with objective evaluations. The experimental results show that this algorithm, which is based on the measure of higher order statistics, performs well in almost every condition compare to previous methods. Also, the proposed scheme still opens many possibilities to improve the parameter setting strategy for further development. After all, this proposed scheme has great potential to be implemented in real environment.

ACKNOWLEDGMENT

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

REFERENCES

[1] 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, Lang. Process.,* vol. 17, no. 4, pp. 650-664, May 2009.

[2] Y. Huang, J. Benesty, and J. Chen, "Dereverberation," *Springer Handbook of Speech Processing,* J. Benesty, M. M. Sondhi, and Y. Huang, Eds., Springer: Berlin London, 2008, pp.929-942.

[3] J. Even, H. Saruwatari, K. Shikano, T. Takatani, "Blind Signal Extraction Based Joint Suppression of Diffuse Background Noise and Late Reverberation," in *Proc. EUSIPCO*, pp. 1534-1538, August 2010.

[4] J. Even, H. Saruwatari, K. Shikano, "Blind Signal Extraction Based Speech Enhancement in Presence of Diffuse Background Noise," in *Proc. IEEE SSP 2009,* pp. 513-516, 2009.

[5] A. Schwarz, K. Reindl, W. Kellerman, "A Two-Channel Reverberation Suppression Scheme Based on Blind Signal Separation and Wiener Filtering," in *Proc. ICASSP*, pp. 113-116, March 2012.

[6] E. K. Kokkinis, A. Tsilfidis, E. Georganti, and J. Mourjopoulos, "Joint noise and reverberation suppression for speech applications," in *Proc. AES 130th Convention*, May 2011.

[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," in *Proc. IWAENC,* 2008.

[8] T. Inoue, H. Saruwatari, K. Shikano, Y. Takahashi, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher-order statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol.19, no.6, pp.1770-1779, 2011.

[9] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, "Musical Noise Generation Analysis for Noise Reduction Methods Based on Spectral Subtraction and MMSE STSA Estimation", in *Proc. ICASSP,* pp.4433--4436, April 2009.

[10] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol.20, No.7, pp.2080-2094, 2012.

[11] K. Ito et al., "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research," *Journal of Acoust. Soc. Of Japan,* vol. 20, pp. 196-206, 1999.

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