Semi-Blind Algorithm for Joint Noise Suppression and Dereverberation Based on Higher-Order Statistics and Acoustic Model Likelihood

Fine Dwinita Aprilyanti*, Hiroshi Saruwatari*, Kiyohiro Shikano*, Satoshi Nakamura* and Tomoya Takatani†
*Graduate School of Information Science, 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 propose an automatic optimization scheme of FD-BSE-based joint suppression of noise and late reverberation to improve the speech recognition accuracy for spoken-dialogue system. First, we optimize the parameter of conventional FD-BSE-based method using the assessment of musical noise measured by higher-order statistics and acoustic model likelihood. Next, to maintain the optimum performance of the system, we proposed the switching scheme using the distance information provided by image sensor. The experimental results show that the proposed approach improves the word recognition accuracy.**

I. Introduction

Hands-free robot dialogue system captures the user's utterances at a distance using a microphone array, resulting in more natural human-machine interaction. In this system, however, it is difficult to achieve accurate speech recognition because of the adverse effect from background noise and room 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 [1]. Therefore, a method that can suppress these interferences is required to improve the recognition accuracy.

Many speech enhancement method have been studied to solve these problems. In [2], it is shown that frequency-domain independent component analysis (FD-ICA) [3] performs better in estimation of diffuse background noise than that of target speech. Hence, combining FD-ICA as noise estimator and a nonlinear postprocessing to suppress the estimated noise has been proved to be effective in improving the target sound quality. Some of the authors have proposed an improved diffuse noise estimator, namely, frequency-domain blind signal extraction (FD-BSE) [4]. Then, by also considering the effect of room reverberation, they combine the FD-BSE with two stages of multichannel Wiener filtering (WF) to jointly suppress diffuse background noise and late reverberation [5] (hereafter this method is referred to as joint method). To synthesize the late reverberation component, this method uses *a priori* knowledge of the room reverberation time (T_{60}).

Joint method proposed in [5] has been proved to be effective to improve the recognition accuracy under noisy and reverberant condition. However, some of the parameters still require manual setting. Therefore, in this paper, first, we address the optimization problem of FD-BSE-based joint method. Motivated by previous works of the authors [6], [7], we have proposed an optimization scheme based on the assessment of musical noise, which is the artificial distortion generated as an effect of nonlinear signal processing [8]. In this paper, we integrate the current optimization scheme with automatic speech recognizer (ASR) under maximum likelihood criterion to improve the recognition accuracy performance.

Next, to achieve high recognition accuracy regardless of the level of interferences, we utilize information on the user distance. This information is obtained under the assumption that a robot has it own camera that can immediately detect the position of the target user. Finally, we confirm the effectiveness of proposed method through experimental evaluations.

II. Related Work: FD-BSE-based Joint Noise Suppression and Dereverberation

In this section, we will review the previously proposed joint method. The architecture of this method is shown in Fig. 1. Generally, the method can be divided into two stages, namely, noise suppression stage and dereverberation stage. Two WFs are utilized separately in each stages, providing flexibility in adjusting filter strength according to the level of each interferences. The dynamic characteristic of these nonlinear filters help to improve the quality of captured speech in real environment due to nonstationary characteristics of interferences.

The observed signal $\boldsymbol{x}(t)$ captured at microphone array is given by

$$\boldsymbol{x}(t) = (\boldsymbol{h}_E(\tau) + \boldsymbol{h}_L(\tau)) * \boldsymbol{s}(t) + \boldsymbol{n}(t), \qquad (1)$$

where $\boldsymbol{s}(t)$ and $\boldsymbol{n}(t)$ are the clean speech source and noise, respectively, and $\boldsymbol{h}_E(\tau)$ and $\boldsymbol{h}_L(\tau)$ indicate the early and late room impulse response. Most hidden Markov model (HMM) based speech recognizers are capable to handle the effect of $\boldsymbol{h}_E(\tau)$ up to certain time delay τ_d, for example by applying cepstral mean normalization.

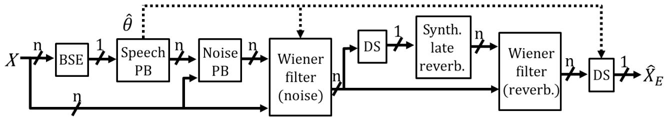

Fig. 1. Block diagram of joint blind noise suppression and dereverberation method.

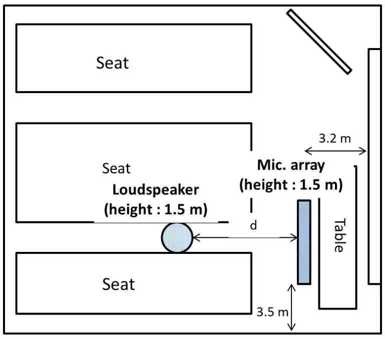

Fig. 2. Room configuration for the experiment.

A. Noise Suppression Stage

Conventional joint method uses FD-BSE algorithm to estimate both direction of arrival (DOA) $\hat{\theta}$ and the background noise component. Unlike conventional FD-ICA, FD-BSE exploits the sparseness of the modulus of the target speech signal [4]. In frequency domain, the extracted output $\boldsymbol{Y}(f,t)$ is obtained by applying extracting vector to the observed signal, as given by

$$\boldsymbol{Y}(f,t) = \boldsymbol{A}(f)\boldsymbol{X}(f,t). \tag{2}$$

The vector $\boldsymbol{A}(f)$ is updated using a gradient descent method to minimize the cost function $J(\boldsymbol{A}(f))$ given by,

$$\begin{aligned} J(\boldsymbol{A}(f)) &= \frac{1}{2}E\{|\boldsymbol{Y}(f,t)|\}^2, & (3)\\ E\{|\boldsymbol{Y}(f,t)|^2\} &= 1. & (4) \end{aligned}$$

In the case of a target speech within background noise, the speech modulus may be considered sparser than that of the diffuse background noise components in the sense that most of its values are close to zero and only a few are significantly large. Thus the cost function is minimum when the target speech component is extracted. In this way, it is not required to confirm the selection of noise components, which means the permutation problem as in frequency-domain blind signal separation (FD-BSS) can be avoided.

The estimated noise $\hat{\boldsymbol{N}}(f,t)$ is obtained by applying projection back to the residual output $\boldsymbol{Y}^{(noise)}(f,t)$ which only contains noise component, given by

$$\hat{\boldsymbol{N}}(f,t) = \boldsymbol{A}^{-1}(f)\boldsymbol{Y}^{(noise)}(f,t). \tag{5}$$

This noise estimate is then suppressed using a set of multichannel WF as given by

$$\begin{aligned} \hat{\boldsymbol{X}}_S(f,t) &= G|\boldsymbol{X}(f,t)|\mathrm{e}^{jarg(\boldsymbol{X}(f,t))}, & (6)\\ G &= \frac{|\boldsymbol{X}(f,t)|^2}{|\boldsymbol{X}(f,t)|^2 + \beta_N|\hat{\boldsymbol{N}}(f,t)|^2}, & (7) \end{aligned}$$

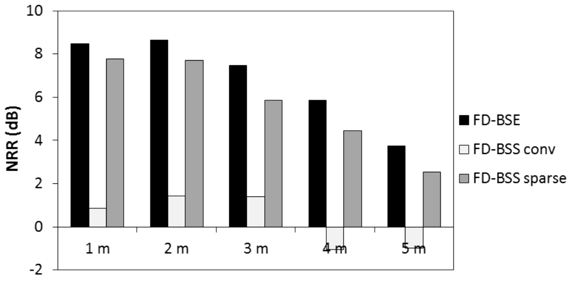

Fig. 3. Performance comparison of FD-BSE and FD-BSS.

TABLE I
SYSTEM SPECIFICATION OF ASR

Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCC, 1-order ΔE
Acoustic model	HMM phonetic tied mixture (PTM), 2000 states, GMM 64 mixtures
Language model	standard word trigram model
Training data	Adult JNAS database

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

B. Dereverberation Stage

Assuming that the noise suppression stage is effective, the estimated $\hat{\boldsymbol{X}}_S(f,t)$ contains only early reverberant speech $\boldsymbol{X}_E(f,t)$ and late reverberant speech $\boldsymbol{X}_L(f,t)$. In this stage, first the late reverberation component is synthesized according to equation (1), given by

$$\hat{\boldsymbol{x}}_L(t) = \boldsymbol{h}_L(\tau) * \hat{s}(t), \tag{8}$$

where $\boldsymbol{h}_L(\tau)$ is approximated by generating channel-wise synthetic tail from decayed Gaussian random variable. The direct speech $s(t)$ estimation requires more strategy as there is no clean speech can be used as reference. In this method, $s(t)$ is estimated using by projecting back the ouput of noise suppression stage $\hat{\boldsymbol{X}}_s(f,t)$ to the truncated FD-BSE filter. After that, dereverberation process is done in the same manner as noise suppression stage, using multichannel WF given by,

$$\begin{aligned} \hat{\boldsymbol{X}}_E(f,t) &= G|\hat{\boldsymbol{X}}_S(f,t)|\mathrm{e}^{jarg(\hat{\boldsymbol{X}}_S(f,t))}, & (9)\\ G &= \frac{|\hat{\boldsymbol{X}}_S(f,t)|^2}{|\hat{\boldsymbol{X}}_S(f,t)|^2 + \beta_R|\hat{\boldsymbol{X}}_L(f,t)|^2}, & (10) \end{aligned}$$

where β_R is a parameter for controlling the strength of dereverberation. Consequently, this stage requires *a priori* information of T_{60}.

III. AUTOMATIC OPTIMIZATION SCHEME OF JOINT METHOD

A. Motivation and Strategy

In order to confirm the effectiveness of FD-BSE to cope with diffuse background noise and reverberation, we conduct

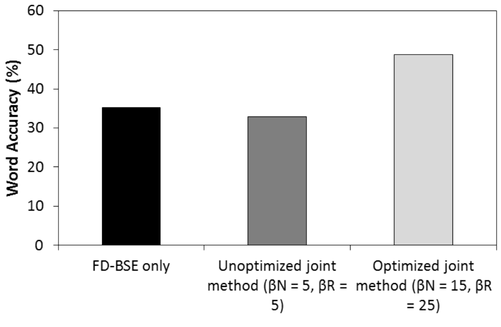

Fig. 4. Effect of parameter settings in joint method to recognition accuracy result for input signal at 5m distance and SNR of 10 dB.

preliminary experiment to compare the output from FD-BSE and FD-BSS method. Here we use ICA algorithm in [9] for FD-BSS. In addition, the modulus-sparseness-based permutation solver is also applied to the FD-BSS method, so the performance of main algorithm in both FD-BSE and FD-BSS can be compared.

An 8-channel microphone array was use to record the room impulse response with the configuration as shown in Fig. 2. The estimated T_{60} is 500 ms. For the input signals, 10 utterances (average length $\approx$ 5 s) were convoluted with real recorded impulse response at various distance between speaker and microphone array, and then were mixed with 10 dB SNR noise.

The performance of each methods is evaluated using noise reduction rate (NRR), which is defined as the difference of signal-to-noise ratio (SNR) of signal before and after processing. The SNR of a signal is represented by

$$\text{SNR} = 10 \log_{10} \frac{E[s(t)]^2}{E[n(t)]^2}, \qquad (11)$$

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.

Figure 3 shows the NRR result of estimated output speech from each methods. It is clearly shown that the conventional FD-BSS cannot perform well in the presence of diffuse background noise and reverberation. The performance is significantly improved in modified FD-BSS due to the use of modulus-sparseness-based permutation solver. However, we can see that the FD-BSE still outperforms in every condition. Moreover, the computation time of FD-BSE is greatly reduced in comparison to that of FD-BSS, with the ratio of 0.37 in average. It is possible due to the fact that the update rule in FD-BSE only involves $n \times 1$ vector, while it is computed for $n \times n$ matrix for FD-BSS. Thus, the use of FD-BSE is preferable.

By combining FD-BSE and WF filters in joint method as described in Sect. II, we gain more flexibility in suppressing the interferences, thus making the method more robust to various acoustical condition. However, one expected problem to arise is the complex parameter optimization and prediction of the best parameters, namely, β_N and β_R, for speech recognition performance. The common way to evaluate this performance by using word accuracy, described as

$$\text{WA} = 100 \times \frac{N - (I + S + D)}{N}, \qquad (12)$$

where N is the number of words in the reference, I is the number of insertions, S is the number of substitutions, and D is the number of deletions.

Another preliminary experiment has been done to analyze the effect of parameter settings in joint method to recognition accuracy. For the recognition task, we use similar condition as the previous experiment, and a 20K-word Japanese dictation task from the JNAS database [10] is used as performance measure. Here, the reference signal is clean speech convoluted with early impulse response. We use JULIUS [11] as speech recognizer. The specification for speech recognition performance evaluation is shown in Table I. As depicted in Fig. 4, the correct setting of both parameters of joint method will result in better word accuracy compared to speech estimation from FD-BSE alone. On the other hand, bad setting of the parameters will result in worse performance. In the conventional joint method, the optimization is still done manually, which makes it impossible to be implemented in real environment. Therefore, it is our concern to build an efficient method to control these parameters automatically.

The previous optimization scheme is based on higher-order statistics which corresponds to the assessment of generated musical noise [8]. In this paper, we propose an optimization scheme that utilizes two control parameters, namely, higher-order statistics and acoustic model likelihood. The proposed optimization scheme is focused on setting of parameters of WFs, under the assumption that T_{60} is known.

B. Parameter Optimization Based on 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, it is possible to assess the amount of musical noise using statistics measures of their pdf. Kurtosis has been introduced in our studies to evaluate the tail of the pdf [12], successfully showing the effectiveness of using the higher-order statistics.

In this paper, first, we calculate the frequency subband-wise kurtosis [7] as given by

$$\text{kurt}^{(i)} = \frac{(1/M) \sum_{f \in F_i} \sum_{t \in T} (|\boldsymbol{X}(f,t)|^2)^4}{\{(1/M) \sum_{f \in F_i} \sum_{t \in T} (|\boldsymbol{X}(f,t)|^2)^2\}^2}, \qquad (13)$$

where $\text{kurt}^{(i)}$ is the i-th subband kurtosis of a signal x. F_i and T represent the evaluated subband time-frequency grid indexes, while M indicates the total number of grids in each subband. Here 250-Hz-width F_i and T of 5 s are used, which are taken from noise-only time-frequency region preceding a

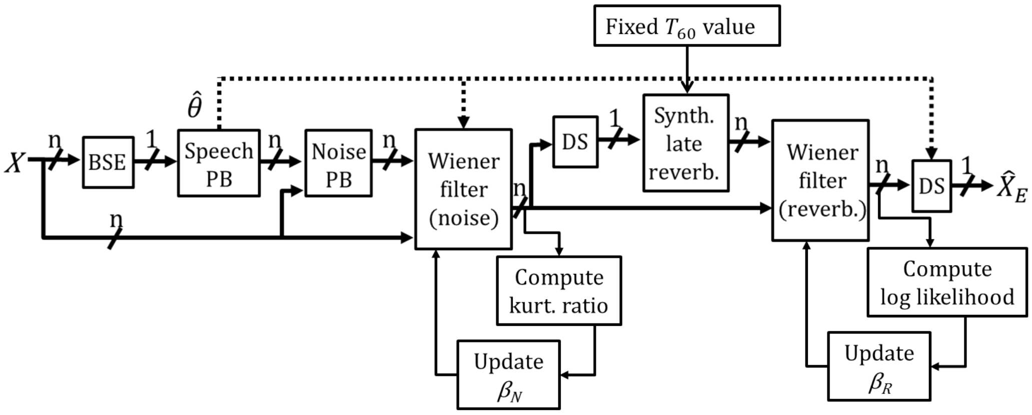

Fig. 5. Optimization scheme of joint method via higher-order statistics and acoustic model likelihood.

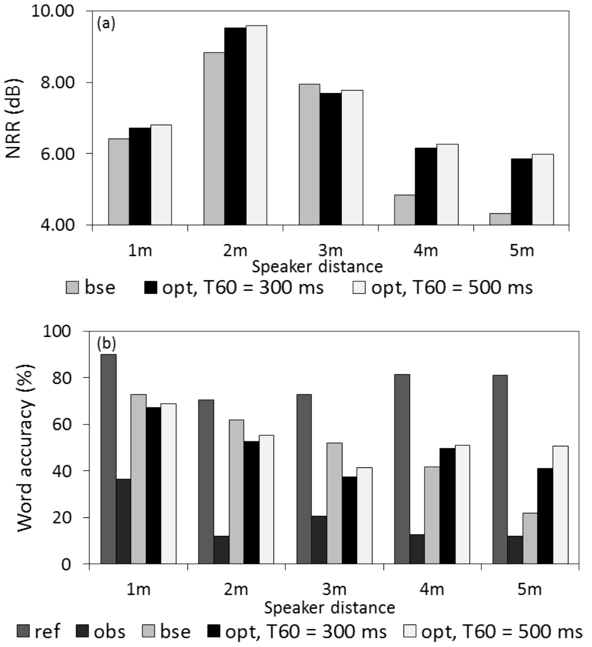

Fig. 6. Experimental evaluation of optimization scheme; (a) NRR, and (b) Word accuracy.

speech utterance. Then, the assessment of generated musical noise is done by applying the *kurtosis ratio*, given by

$$\text{kurtosis ratio} = \text{kurt}_{\text{proc}}/\text{kurt}_{\text{org}}, \tag{14}$$

where $\text{kurt}_{\text{proc}}$ is the kurtosis of the processed signal and kurt_{org} is the kurtosis of the observed signal.

C. Parameter Optimization Based on Acoustic Model Likelihood

In spoken-dialogue system, the speech enhancement method can only be expected to improve recognition performance if it generates results that improve the likelihood of the correct transcriptions. Therefore, we also optimize the parameter to maximize the likelihood of acoustic model of speech recognizer. Previous researches have confirmed the effectiveness of this solution in optimizing several speech enhancement methods, for example on beamforming method [13].

In speech recognizer, a series of fixed size acoustic vectors $\boldsymbol{o}(\beta) = [o_1, ..., o_T]$ is extracted from the output of speech enhancement with parameter β through some feature extraction process. During decoding, it attempts to hypothesize the word sequence $\boldsymbol{W} = [w_1, ..., w_K]$ which is the most probable to generate the sequence $\boldsymbol{o}(\beta)$, as stated by

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} P(\boldsymbol{W}|\boldsymbol{o}(\beta)). \tag{15}$$

However, the recognition system cannot compute the posterior probability $P(\boldsymbol{W}|\boldsymbol{o}(\beta))$ directly. Instead, the above expression is transformed into the following form based on Bayes' theorem:

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \frac{P(\boldsymbol{o}(\beta)|\boldsymbol{W})P(\boldsymbol{W})}{P(\boldsymbol{o}(\beta))}, \tag{16}$$

where $P(\boldsymbol{o}(\beta)|\boldsymbol{W})$ is the acoustic likelihood or acoustic score, representing the probability that feature sequence $\boldsymbol{o}$ is observed given that word sequence $\boldsymbol{W}$ was spoken, and $P(\boldsymbol{W})$ is the language score, i.e., the *a priori* probability of a particular word sequence $\boldsymbol{W}$. The former term is calculated from acoustic model, while latter term is computed using a language model.

Since Eq. (16) is maximized with respect to the word sequence $\boldsymbol{W}$ for a given observed sequence $\boldsymbol{o}$ that is fixed, the denominator term $P(\boldsymbol{o}(\beta))$ can be ignored. Thus, the parameter β can be optimized by maximizing the likelihood of acoustic model of speech recognizer, as written by

$$\hat{\beta} = \arg\max_{\beta} P(\boldsymbol{o}(\beta)|\boldsymbol{W}). \tag{17}$$

D. Algorithm and Experimental Evaluation

The block diagram of the proposed optimization scheme is shown in Fig. 5. Provided the *a priori* knowledge of T_{60}, the WF parameters are updated consecutively. First, the residual noise quality is assessed from silence part of signals preceeding the utterances, using voice activity detection based on noise estimation from FD-BSE. Using this assessment, β_N

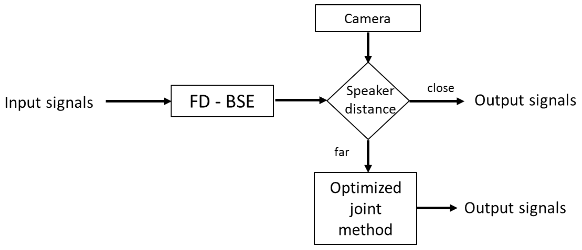

Fig. 7. Block diagram of semi-blind optimized joint method.

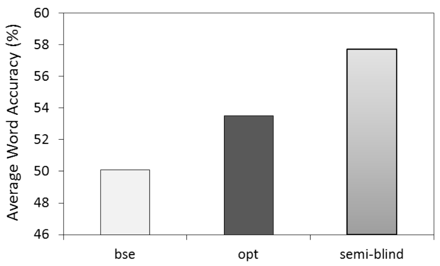

Fig. 8. Word recognition accuracy result of proposed scheme.

is updated to achieve optimum NRR under a kurtosis ratio constraint, as given by

$$\hat{\beta_{\mathrm{N}}} = \arg\max_{\beta_{\mathrm{N}}} \mathrm{NRR}(\beta_{\mathrm{N}}), \quad \frac{\mathrm{kurt}_{\mathrm{proc}}(\beta_{\mathrm{N}})}{\mathrm{kurt}_{\mathrm{org}}} \leq \mathrm{KR}_{\mathrm{lim}}, \tag{18}$$

where $\mathrm{KR}_{\mathrm{lim}}$ is a constraint value of KR. Next, β_R is updated according to acoustic model likelihood according to Eq. (17), to optimize

$$\hat{\beta_R} = \arg\max_{\beta_R} P(\boldsymbol{o}(\beta_R)|\boldsymbol{W}). \tag{19}$$

This two-step scheme has a great advantage that the complex optimization of two parameters β_N and β_R can be reasonably decomposed into two simple optimization for each parameters. By using noise kurtosis ratio constraint in noise suppression stage, the unwanted background noise can be suppressed effectively, thus the late reverberation synthesis becomes more accurate. Moreover, the acoustic model likelihood criterion for optimization in dereverberation stage will ensure optimum recognition accuracy of the output speech.

We conduct experiments to evaluate the performance of the proposed optimization scheme. For this experiment, we use 200 utterances from JNAS database as input signals, while other settings are set similar to preliminary experiment. The performance of the optimized joint method (**opt**) is compared with the estimated speech from FD-BSE (**bse**). The experiment is also conducted on underestimated T_{60} of 300 ms, to analyze the effect of mismatched value to the performance of the proposed scheme. The experimental results are shown in Fig. 6. For the recognition performance, the reference is indicated by (**ref**) and the unprocessed observed signal is indicated by (**obs**).

From the figure, it is shown that the proposed scheme is superior to FD-BSE in terms of noise reduction performance. However, the word recognition accuracy result shows that the speech estimate from FD-BSE gives best accuracy at close distance compare to the proposed method. The effect of mismatched T_{60}, as one can expect, is not very clear in case of NRR, but it is significant at recognition accuracy, particularly under heavily reverberant condition indicated by far speaker to microphone distance.

IV. Semi-blind Joint Noise Suppression and Dereverberation Using Image Information

A. Motivation and Strategy

The experimental result shows that the proposed optimization scheme fails to achieve optimum speech recognition accuracy at close user distance. This may be happened because of the following reasons:

- FD-BSE as linear filter results in output signal with less distortion when the interferences effect is not so severe, compare to the proposed scheme that includes nonlinear processing.
- At closer user distance, the effect of room reverberation is light to moderate. The late reverberation becomes overestimated in the proposed scheme.
- Some part of late reverberation may have been treated as background noise due to similar characteristics.

In order to achieve the optimum performance, we need a system that can select the signal processing method to be applied according to the condition of environments. In this paper, we utilize the user position information, provided by robot's camera, and develop a multimodal switching scheme according to the distance information, under assumption that the user distance corresponds to the effect of interferences.

B. Algorithm and Experimental Evaluation

The block diagram of the proposed multimodal scheme is shown in Fig. 7. First, the system is trained to estimate the distance limit to switch method. Both FD-BSE and optimized joint method are applied to the signals at various speaker to microphone distance, and the results are compared. Next, after the switching point is decided, the system will apply two different schemes according to the user position:

- For close user distance, only FD-BSE estimation is applied to the input signal. The extracted speech from FD-BSE becomes the output signal.
- For far user distance, the optimized joint method is applied in addition to the FD-BSE.

The average result of recognition performance is depicted in Fig. 8. We compare the performance of semi-blind joint method with estimated speech from FD-BSE (**bse**) and the optimized joint method (**opt**). In average, semi-blind method achieves better word recognition accuracy compare to the conventional method. This shows that the semi-blind joint approach can maintain optimum performance regardless the interference condition, in contrast with the other two methods.

V. Conclusions

In this paper, first, we have proposed an optimization scheme of FD-BSE based joint noise suppression and dereverberation via higher-order statistics and acoustic model likelihood. Next, to maintain optimum word recognition accuracy performance, we developed a semi-blind method selection scheme using image information of user position. The experimental result confirms the effectiveness of our proposed method.

VI. Acknowledgements

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

References

[1] Y. Huang, J. Benesty, and J. Chen, "Dereverberation," *Handbook of Speech Processing*, J. Benesty, et al. [Eds], pp.929-942, Springer, 2008.

[2] 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, 2009, pp.650-664.

[3] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proc. ICA99*, Jan. 1999, pp.365-371.

[4] J. Even, H. Saruwatari, and K. Shikano, "Blind signal extraction based speech enhancement in presence of diffuse background noise," in *Proc. IEEE SSP 2009*, pp.513-516, 2009.

[5] J. Even, H. Saruwatari, K. Shikano, and T. Takatani, "Blind signal extraction based joint suppression of diffuse background noise and late reverberation," in *Proc. EUSIPCO*, pp.1534-1538, 2010.

[6] 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, 2011, pp.1770-1779.

[7] H. Saruwatari, Y. Ishikawa, Y. Takahashi, T. Inoue, K. Shikano, and K. Kondo, "Musical noise controllable algorithm of channelwise spectral subtraction and adaptive beamforming based on higher order statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol.19, no.6, 2011, pp.1457–1466.

[8] F. D. Aprilyanti, H. Saruwatari, K. Shikano, and T. Takatani, "Optimization Scheme of Joint Noise Suppression and Dereverberation Based on Higher-Order Statistics," in *Proc. APSIPA ASC 2012*, Dec. 2012.

[9] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol.22, no.1-3, pp.21-34, 1998.

[10] K. Ito, et al., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of Acoust. Soc. Of Japan*, vol.20, 2006, pp.196–206.

[11] Julius, an open-source large vocabulary csr engine - http://julius.sourceforge.jp.

[12] 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, 2012, pp.2080–2094.

[13] M. L. Seltzer and R. M. Stern, "Subband likelihood-maximizing beamforming for speech recognition in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol.14, no.6, 2006, pp.2109–2121.