Suppression of noise and late reverberation based on blind signal extraction and Wiener filtering

Fine D. Aprilyanti¹*, Jani Even², Hiroshi Saruwatari³, Kiyohiro Shikano¹, Satoshi Nakamura¹ and Tomoya Takatani⁴

¹Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
²Advanced Telecommunications Research Institute International, Intelligent Robotics and Communication Laboratories, Kyoto, Japan
³Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
⁴Toyota Motor Corporation, Aichi, Japan

(Received 1 May 2014, Accepted for publication 26 January 2015)

Abstract: We introduce a new optimized microphone-array processing method for a spoken-dialogue robot in noisy and reverberant environments. The method is based on frequency-domain blind signal extraction, a signal separation algorithm that exploits the sparseness of a speech signal to separate the target speech and diffuse background noise from the sound mixture captured by a microphone array. This algorithm is combined with multichannel Wiener filtering so that it can effectively suppress both background noise and reverberation, given a priori information of room reverberation time. In this paper, first, we develop an automatic optimization scheme based on the assessment of musical noise via higher-order statistics and acoustic model likelihood. Next, to maintain the optimum performance of the system, we propose the multimodal switching scheme using the distance information provided by robot’s image sensor and the estimation of SNR condition. Experimental evaluations have been conducted to confirm the efficacy of this method.

Keywords: Blind signal extraction, Noise suppression, Dereverberation, Higher-order statistics, Acoustic likelihood

PACS number: 43.60.Hj, 43.72.Dv [doi:10.1250/ast.36.302]

1. INTRODUCTION

Recent developments have allowed humans to interact with machines in more flexible ways through a speech input system. However, the performance of such a system often significantly degrades when it is implemented in a real environment, in which it suffers from the adverse effect of background noise and room reverberation. Improving the speech recognition accuracy is particularly difficult in the case of a hands-free spoken dialogue system, where a microphone or a microphone array is utilized to capture the target speech at some distance from the user.

Many studies have focused on enhancing the target speech quality under these conditions. The most common single-microphone techniques include spectral subtraction (SS) [1] and Wiener filtering (WF) [2]. Moreover, the use of a microphone array enables spatial filtering that improves the performance of the speech enhancement method. Delay-and-sum (DS) beamforming [3,4] and adaptive beamforming (ABF) [5] are the most conventional microphone array techniques. Another approach is blind source separation (BSS), in which the observed signals from each microphone are assumed to be linear combinations of the source signals. Many BSS algorithms [6–8] are based on independent component analysis (ICA) [9,10], in which the source signals are separated by utilizing their statistical independence. To improve separation performance, some methods combining ICA and beamforming have also been proposed [11].

Conventional ICA performs well if all the signals can be regarded as point-source signals. However, in a real environment, noise interference is often widespread and diffuse. Takahashi et al. [12] have shown that in the case of a mixture of speech and non-point-source interference, frequency-domain ICA (FD-ICA) performs better in estimating the diffuse background noise than the target
speech. Thus, by combining FD-ICA as a noise estimator and nonlinear postprocessing, such as SS or WF, to suppress the estimated noise, they developed a speech enhancement method that has been proved to be effective in improving the target sound quality. Furthermore, Even et al. [13] have introduced a signal extraction algorithm designed for extracting a target speech from a background noise by exploiting the difference in their modulus sparseness. For this specific problem, this method, called frequency-domain blind signal extraction (FD-BSE), gives better estimates for both the target and the background noise than the conventional FD-BSS.

Dealing with reverberation requires more complicated signal processing because it has unique characteristics. Owing to the precedence effect, the early reverberation tends to reinforce the direct speech and is considered to have no adverse effect on speech intelligibility. However, the late reverberation can mask the original speech and deteriorate the sound quality, depending on its length and strength [14]. Many attempts have been made to compensate the effects of reverberation. An obvious method is to find the inverse of the room impulse response, as proposed in [15]. It has been shown that the exact inverse can be obtained if the channel impulse response has no common zeros. The authors of [16] have shown that under some constraint, the late reverberation components are uncorrelated to the direct signal and the early reverberation components; thus, dereverberation can be carried out by applying amplitude SS. However, these methods work under the assumption that the background noise is weak or well controlled.

The purpose of this research is to develop a method that suppresses the additive noise and the late reverberation components. Recently, a large amount of work has been done in this field, e.g., [17–19]. The method in [20] extends the capability of FD-BSE to enhance the target speech quality by utilizing two sets of quasi-parametric multichannel WF as nonlinear postprocessing. In this method, FD-BSE is used to estimate background noise, which is suppressed using the first multichannel WF. The dereverberation process then utilizes the output of noise suppression stages to synthesize and suppress the late reverberation component.

The FD-BSE-based method provides more flexibility in its setting, which gives it the potential to be implemented under various acoustical conditions, assuming that the room reverberation time $T_{60}$ is known. However, the parameters in this method are still manually selected, which makes it impractical to be implemented in real environments. In this paper, we develop an automatic optimization scheme based on higher order statistics and acoustic model likelihood to solve this problem. Also, provided image sensor that can assess the user distance information, we develop a more robust semi-blind method.

The remainder of this paper is organized as follows. In Sect. 2, we briefly review the sound mixing model and several related works, such as FD-BSE and the suppression of the late reverberation effect. In Sect. 3, we present the FD-BSE-based noise suppression and dereverberation method and the proposed optimization scheme. We evaluate our proposed method experimentally and discuss the result in Sect. 4. An additional semi-blind method to improve the performance of the proposed method is described in Sect. 5. Finally, we summarize and conclude our work in Sect. 6. Throughout the paper, the vectors and matrices are indicated in bold face. The matrix $A^{H}$ is the Hermitian transposition of the matrix $A$ and $\delta (\cdot)$ denotes the time-averaging operator.

2. DATA MODEL AND RELATED WORK

2.1. Sound Mixture Model at Microphone Array

Signals captured by an $M$-channel microphone array are composed of a clean speech signal and interference sounds. This interference can be additive, that is background noise, and also convolutive, that is room reverberation. Mathematically, a mixture of signals at a microphone array can be modeled by

$$x(t) = x_S(t) + x_N(t),$$

$$x_S(t) = (h_E(t) + h_L(t)) * s(t) = x_E(t) + x_L(t),$$

where $x_N(t)$ denotes the contribution from background noise and $x_S(t)$ denotes the contribution from speech and its reverberation. The signal $s(t)$, $h_E(t)$ and $h_L(t)$ are the clean speech source and the early and late parts of the room impulse response, respectively, with $*$ denoting the convolution operation. Since the early reverberation can be handled by hidden Markov model (HMM)-based automatic speech recognition (ASR) if it lies within certain values of $\tau_d$, the early and late impulse responses can be defined as

$$h_E(t) = \begin{cases} h(t) & \text{for } \tau \leq \tau_d \\ 0 & \text{for } \tau > \tau_d \end{cases}$$

$$h_L(t) = \begin{cases} 0 & \text{for } \tau \leq \tau_d \\ h(t) & \text{for } \tau > \tau_d, \end{cases}$$

and the dereverberation method is focused on obtaining the estimated early reverberant speech signal $\hat{x}_E(t)$.

In the time-frequency domain, the mixture model in each frequency bin can be approximated by

$$X(f, k) \approx H(f)S(f, k) + X_N(f, k),$$

where $f$ and $k$ denote the frequency bin and time frame, respectively. We can rewrite Eq. (5) in matrix/vector form as.
\[ X(f, k) = [H(f)]_M \begin{bmatrix} S(f, k) \\ X_N(f, k) \end{bmatrix}, \] (6)

where \( I_M \) is the identity matrix of size \( M \). Without loss of generality, we can reformulate Eq. (6) as

\[ X(f, k) = A(f) S(f, k). \] (7)

Here, we assume that \( S_i(f, k) = S(f, k) \) and \( [S_2(f, k), \ldots, S_{M+1}(f, k)] = X_N(f, k) \). It can also be assumed that the speech component is statistically independent of the noise component in each frequency bin.

### 2.2. Frequency-Domain Blind Signal Extraction

We will discuss the FD-BSE algorithm and its comparison to the conventional FD-BSS in this section. In ICA-based FD-BSS method, the estimated separated signals \( Y(f, k) \) in the \( f \)th frequency bin are obtained by applying an unmixing matrix \( B(f) \) to the observed signals, as given by

\[ Y(f, k) = B(f)X(f, k) = B(f)A(f)S(f, k). \] (8)

\( B(f) \) is updated so that the output signals in \( Y(f, k) \) are mutually independent. For example, the algorithm that based on higher-order statistics \([8]\) searches the optimal mutually independent. For example, the algorithm that

\[ \text{Here, we assume that } S_i(f, k) = S(f, k) \text{ and } [S_2(f, k), \ldots, S_{M+1}(f, k)] = X_N(f, k). \]

It can also be assumed that the speech component is statistically independent of the noise component in each frequency bin.

Since the above calculations are carried out independently in each frequency bin, FD-ICA suffers from two problems, i.e., source permutation and scaling indeterminacy. The scaling indeterminacy problem can be solved by applying a projection back (PB) of the separated independent components to the microphone array input \([7]\). If \( Y_m(f, k) \) is the estimated speech component, the projection back of the background noise components can be defined by

\[ \hat{X}_N(f, k) = B(f)^{-1}DY(f, k), \] (10)

where \( D \) is a diagonal unit matrix with \( d_{mn} = 0 \). The remaining permutation problem requires the matching of the components belonging to the same signal across all the frequency bins. This is carried out by applying a permutation resolution. The methods of permutation resolution often utilize the DOA or temporal structure of signals. However, although the scaling and permutation problem have been solved, each of source signals cannot be recovered perfectly because the number of sources exceeds the number of microphones.

The authors of \([12]\) showed that the unmixing matrix \( B_{ICA}(f) \) works as a DS beamformer in the direction of the speech’s DOA for the row corresponding to the estimated speech component. The other rows corresponding to the estimates of the noise components are null beamformers at the speech’s DOA. In the case of nonpoint noise sources, the DS beamformer only slightly enhances the speech and much of the diffuse noise still remains. On the other hands, although the noise components cannot be estimated perfectly, the null beamformers can efficiently suppress the speech (a point source) from the estimated noise components. For this reason, it is preferable to utilize FD-ICA as a background noise estimator and then apply a nonlinear postfilter, such as SS or WF, to suppress the estimated noise.

In contrast to the conventional FD-BSS, the FD-BSE algorithm only extracts the speech components from noise components rather than separate the signal from each source. Given the same observation signal \( X(f, k) \), BSE estimates only the components of \( S_i(f, k) \) in each frequency bin by applying an extracting vector

\[ Y(f, k) = W(f)X(f, k). \] (11)

The extracting vector \( W(f) \) is updated to minimize the cost function \([20]\)

\[ J(W(f)) = \frac{1}{2} \mathbb{E}[|Y(f, k)|^2] \] (12)

under the constraint

\[ \mathbb{E}[|Y(f, k)|^2] = 1. \] (13)

Under the assumption that the target speech is statistically independent of the background noise, the FD-BSE method extracts the target speech in all frequency bins. The reason is that, in the time-frequency domain, the modulus of the speech component in each frequency bin has a spiky distribution compared with that of the diffuse background noise components and the spikier is a distribution, the smaller is the FD-BSE cost under the normalization constraint (see \([13]\) for details).

After extracting the target speech component, the residual noise components can be obtained by subtracting the orthogonal projection of the extracted signal (speech PB) from the observed signals (noise PB), as given by

\[ X_N(f, k) = W_R(f)X(f, k), \] (14)

where

\[ W_R(f) = I_M - \Gamma_X(f)W^H(f)W(f), \]

\[ \Gamma_X(f) = \mathbb{E}[X(f, k)X^H(f, k)]. \] (15)
2.3. Suppression of Late Reverberation Effect

In this paper, we deal with the reverberation using the framework presented in [16,22]. The speech signal has a strong correlation within each local time frame according to the articulatory movement. However, because of the nonstationary characteristics of speech, it loses its correlation after a certain time delay, as given by

\[ \gamma_s(\tau) \triangleq \mathbb{E}\{s(t)s(t-\tau)\}, \]  
\[ |\gamma_s(\tau)| \approx 0 \text{ iff } \tau \geq \tau_d. \]  

By this assumption, the late reverberation components can be suppressed in the same manner as the noise component in the power spectrum domain.

The method proposed in [23] uses prior knowledge of the room impulse response to efficiently estimate the late reverberation components. Assuming that the late room impulse response \( h_L(\tau) \) does not vary significantly, the late reverberation components is estimated directly from the observed signal, and the error of estimation is suppressed by using SS with parameters optimized in an offline training stage.

3. PROPOSED METHOD 1: OPTIMIZED FD-BSE-BASED NOISE SUPPRESSION AND DEREVERBERATION

3.1. Main Algorithm

The block diagram of FD-BSE-based method proposed in [20] is shown in Fig. 1. Using a similar approach to that in [12], this method combines FD-BSE as a noise estimator. But contrary to [12], the postprocessing is carried out in multichannel to maintain the spatial quality of the sound sources. The suppression of the diffuse background noise and the late reverberation components is performed by two sets of multichannel WF. The output signals are merged by applying a DS beamformer at the end of the processes.

3.1.1. Background noise suppression

This method uses the FD-BSE algorithm to estimate the background noise component. This is done by subtracting the orthogonal projection of the extracted speech component \( Y(f, k) \) obtained from Eq. (11), as given by

\[ \hat{X}_N(f, k) = (I_M - \Gamma_X(f)\lambda^H \hat{W}(f)\lambda \hat{W}(f))X(f, k), \]  

where \( \lambda \) is a scalar such that \( Q(f, k) = \lambda \hat{W}(f)X(f, k) \) verifies \( \mathbb{E}\{|Q(f, k)|^2\} = 1 \). Then, the estimated noise is suppressed using multichannel WF, as given by

\[ \hat{X}_S(f, k) = G|X(f, k)|e^{j\text{arg}(X(f, k))}, \]  
\[ G = \frac{|X(f, k)|^2}{|X(f, k)|^2 + \beta N |\hat{X}_N(f, k)|^2}, \]

where \( \beta_N \) is a parameter used to control the strength of noise suppression.

3.1.2. Dereverberation stage

Assuming that the noise suppression stage is effective, the estimated \( \hat{X}_S(f, k) \) contains only the early reverberant speech \( X_E(f, k) \) and late reverberant speech \( X_L(f, k) \). The next step is to estimate and suppress the late reverberation components. The estimation of the late reverberation components can be separated into two tasks: estimating the late impulse response \( h_L(\tau) \) and the clean speech signal \( s(t) \).

The estimation of \( h_L(\tau) \) is quite simple owing to the characteristics of late reverberation, which has little correlation with the direct sound and is rather room-dependent. In this method, the estimated \( h_L(\tau) \) is approximated by generating a synthetic tail from a Gaussian random variable \( u(\tau) \) with an exponential decay [16], and is given by

\[ h_L(\tau) = au(\tau)e^{-d(\tau-\tau_d)}, \]  
\[ d = \frac{\ln 10^6}{2(T_{60} - \tau_d)}, \]

where \( a \) is a scaling factor. Consequently, the synthesis of \( h_L(\tau) \) requires a priori information on \( T_{60} \) and \( a \).

The direct speech \( s(t) \) must be estimated as no clean speech is provided that can be used for reference. In our proposed method, \( s(t) \) is approximated by projecting back the output of the noise suppression stage \( \hat{X}_S(f, k) \) to the truncated FD-BSE filter \( W_{\text{trunc}}(f) \), as given by

\[ \hat{S}(f, k) = \Gamma_{\hat{X}_S(f)}\lambda^H W_{\text{trunc}}(f)\lambda W_{\text{trunc}}(f)\hat{X}_S(f, k). \]  

The scaling factor \( a \) is also obtained from the energy difference between \( \hat{X}_S(f, k) \) and \( \hat{S}(f, k) \). Then, according to Eq. (2), \( \hat{X}_L(f, k) \) is obtained by applying a convolution in the time domain, given by

\[ \hat{X}_L(t) = h_L(\tau) * \hat{S}(t), \]  

then it transformed back to the time-frequency domain by
where the microphone spacing, for these parameters. In this paper, we focus on developing an optimization scheme for the WF parameters under the assumption that $T_{\theta 0}$ is provided.

The optimization scheme is based on two control parameters, namely, the amount of musical noise assessed via higher-order statistics for noise suppression stage and acoustic model likelihood for dereverberation stage. The use of these two parameters is based on the following reasons:

- The late reverberation components are synthesized from the output of the noise suppression stage. Therefore, it is important to optimize the parameter in order to improve the quality of the output waveform of this stage.
- This method is designed to be implemented in hands-free robot dialogue system. Therefore, the final stage of this method must be optimized to improve the speech recognition accuracy.

The block diagram of the proposed method is shown in Fig. 3.

3.2. Optimization Scheme

By combining FD-BSE with multichannel WF, this method provides more flexibility in suppressing the interference, thus making it more robust to various acoustical conditions. On the other hand, an expected problem is how to predict the best combination of the parameters, namely, $\beta_N$ and $\beta_R$, so that the proposed method yields the optimum performance. In [20], a very large number of parameter combinations were tested before selecting the value giving the best results a posteriori. This is very impractical for to be implemented in real environment with changing conditions. Therefore, it is of great interest to realize a good automatic optimization scheme for the calculation of the frequency subband-wise kurtosis [28], as given by

$$
\text{kurt}^{(i)} = \frac{(1/L) \sum_{f \in F_i} \sum_{k \in T} |X(f,k)|^4}{(1/L) \sum_{f \in F_i} \sum_{k \in T} |X(f,k)|^2},
$$

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 $L$ is the total number of grids in each subband. Here, a 250-Hz-wide $F$, and a $T$ of 5 s are used, which are taken from a noise-only time-frequency region preceding a speech utterance. Then, the generated musical noise is assessed by applying the kurtosis ratio (KR), given by

$$KR = \text{kurt}_{\text{proc}} / \text{kurt}_{\text{org}}, \quad (33)$$

where kurt$_{\text{proc}}$ is the kurtosis of the processed signal and kurt$_{\text{org}}$ is the kurtosis of the observed signal. Using this assessment, $\beta_N$ is updated in an iterative manner to achieve the optimum NRR under a KR constraint, as given by

$$\hat{\beta}_N = \arg\max_{\beta_N} \text{NRR}(\beta_N), \quad (34)$$

$$\frac{\text{kurt}_{\text{proc}}(\beta_N)}{\text{kurt}_{\text{org}}} \leq \text{KR}_{\text{lim}}, \quad (35)$$

where KR$_{\text{lim}}$ is the constraint value of KR. The procedure is described as follows.

**Step 0**: First, set initial $\beta_N$.

**Step 1**: Next, apply the WF for noise suppression using the value of $\beta_N$.

**Step 2**: Apply DS beamformer to the output signal, then calculate the subband-kurtosis using Eq. (32). Obtain kurtosis ratio by dividing kurtosis of the output signal by the kurtosis of observed signal.

**Step 3**: Increase the value of $\beta_N$ 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.

### 3.2.2. Optimization based on acoustic model likelihood

Since KR is calculated in a noise-only region of the observed signal, parameter optimization based on higher-order statistics is not suitable in the dereverberation stage. Therefore, we choose another control parameter that maximizes the likelihood of the acoustic model on an ASR system. This approach has been previously applied in several speech enhancement methods such as beamforming [29] and SS [30].

The ASR system is a statistical pattern classifier that is based on an extracted feature vector of the speech waveform. It hypothesizes the correct transcription by finding the sentence that has the maximum likelihood of generating the extracted feature vector given the statistical models of the recognizers. In ASR, a series of fixed-size acoustic feature vectors $o = [o_1, \ldots, o_T]$ is first extracted from the speech waveform through some transformation. During decoding, the ASR system attempts to find the word sequence $Z = [z_1, \ldots, z_K]$ that is most likely to generate the sequence $o$, as expressed by

$$\hat{Z} = \arg\max_P P(Z|o). \quad (36)$$

By applying Bayes’ theorem, the above expression can be written as

$$\hat{Z} = \arg\max_Z \frac{P(o|Z)P(Z)}{P(o)}, \quad (37)$$

where $P(o|Z)$ is the acoustic likelihood or acoustic score, representing the probability that acoustic feature sequence $o$ is observed given that word sequence $Z$ was spoken, and $P(Z)$ is the language score, i.e., the a priori probability of a particular word sequence $Z$, which is calculated using a language model.

For a fixed feature vector sequence $o$, the denominator $P(o)$ can be ignored. The language score also can be ignored as it does not related to the acoustic feature of the signal. Thus, the WF parameter in the dereverberation stage is optimized to maximize

$$\hat{\beta}_R = \arg\max_{\beta_R} P(o(\beta_R)|Z). \quad (38)$$

For an HMM-based speech recognition system, the solution of this problem can be computed using the Viterbi algorithm [31]. In practice, the calculation of (38) requires the correct transcription $Z$, but if it is already known, the speech recognizer is not required anymore. Therefore, the optimization procedure is carried out as follows.

**Step 0**: First, set initial $\beta_R$.

**Step 1**: Next, apply the WF for dereverberation using the value of $\beta_R$.

**Step 2**: Apply DS beamformer to the output signal, then calculate the log likelihood $P(o(\beta_R)|Z)$ using Viterbi algorithm.

**Step 3**: Increase the value of $\beta_R$ by a certain amount.
Δpₖ. Return to Step 1 and compare the log likelihood score, using Z from initial βᵣ as the correct transcription. Repeat the process until maximum score achieved.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

We conducted experiments to evaluate the performance of our proposed method. In the first experiment, we investigate the effectiveness of using two different optimization parameter by comparing the performance with other optimization schemes, namely:

- Kurtosis-ratio-based method in which both βₙ and βᵣ are optimized based on higher-order statistics criterion.
- Acoustic-likelihood-optimized method in which both βₙ and βᵣ are optimized to maximize the acoustic score.

In the second experiment, we compared the performance of our proposed optimization scheme with currently known method, namely, BSSA [12], multi-step linear prediction based dereverberation method (MSLP) [32], and method adapted from full-rank spatial covariance model (FRSC) [33]. Since MSLP works under noise-free assumption, we also investigate the performance of MSLP combined with WF for noise suppression (Denoised-MSLP). For this combined method, background noise is estimated with a priori SNR [34]. The number of filter coefficient and step-size for MSLP are set to 750 and 360, respectively, according to the setting in [32]. The maximum iteration for EM algorithm in FRSC is set to 100.

For the experiment, an 8-channel microphone array was used to record the room impulse response with the configuration shown in Fig. 4. The estimated T₆₀ of the room is 500 ms. We used 100 utterances from the female JNAS database [35] as the source signals. Speech was convoluted with the room impulse response and mixed with a noise signal at SNRs of 0, 10 and 20 dB. To simulate the real acoustic condition, we used real environment noise recorded using the same microphone array as one in impulse response recording. The τₜ value was set to 75 ms, which corresponds to the effect of a room impulse response that can still be handled by the HMM-based speech recognizer used for experimental evaluation.

The time-frequency domain processing was done by implementing the short-time Fourier transform (STFT) with a 1,024 point FFT size, a Hanning window and 50% overlap. The FD-BSE algorithm was performed for 600 iterations with an adaptation step of 0.3, which was halved every 200 iterations. For the recognition task, Julius 4.2 [36] was used as the decoder with the configurations as shown in Table 1. Noise-free speech convoluted with early room impulse response was used as reference.

4.2. Results and Discussion

The noise reduction performance (NRR) is used to evaluate the performance of each method. NRR is defined as the difference in the SNR of the signal before and after processing. The SNR of a signal is given by

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

(39)

where s(t) and n(t) are the speech and noise components of the signal, respectively. A higher NRR value indicates a better suppression of the noise and the late reverberation components. The experimental results of each method is depicted in Fig. 5. It is shown that the optimized method based on higher-order statistics achieves the best results under almost every condition, as this optimization scheme works in the noise region of the signal.

The results for word recognition accuracy is shown in Fig. 6. On the contrary to the NRR results, it shows that the proposed optimization scheme performs better than the optimization scheme based on only higher-order statistics. This illustrates the fact that improving the NRR does not directly improve the recognition accuracy. On the other hands, optimization scheme based only on acoustic model likelihood is not very effective in improving the performance, particularly under low input SNR condition. This is due to the structure of the FD-BSE-based method, in which good output signal quality from noise suppression stage is

![Fig. 4 The room setting for the experiment.](image)

Table 1 Specifications setting for the experiment.

| 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), 2,000 states, GMM 64 mixtures |
| Language model | Standard word trigram model |
| Training data | Adult JNAS database |
Fig. 5 NRR results of each method in various input SNR condition: (a) 0 dB, (b) 10 dB and (c) 20 dB.

Fig. 6 Word accuracy results of different optimization scheme in various input SNR condition: (a) 0 dB, (b) 10 dB and (c) 20 dB.
required to synthesize the late reverberation components at the dereverberation stage. Moreover, for a short user distance, FD-BSE achieves better recognition results than the proposed method.

Performance comparison with other method is depicted in Fig. 7. It is shown that the proposed method outperforms other method under almost all conditions. We can see that BSSA performs well under severe SNR conditions and short user distance, since this method only suppress background noise. On the other hand, MSLP fails to perform under severe SNR conditions due to noise-free assumption. Combination with WF for noise suppression do not improve the recognition accuracy, probably because the noise suppression causes more distortion to the target speech. The poor performance of FRSC may be caused by failure in initialize the parameters, as this method is sensitive to initialization process.

The experimental results show that the proposed method 1 fails to achieve optimum speech recognition accuracy at a short user distance. This may have occured because of the following factors:

- FD-BSE, which is a linear filter, results in an output signal with less distortion than the proposed method, which includes nonlinear processing, when the interference effect is not so severe.
- At a shorter user distance, the effect of room reverberation is light to moderate. The late reverberation becomes overestimated in the proposed method owing to the simple approximation of the late room impulse response.
- Some of the late reverberation may have been treated as background noise since it shares similar characteristics.

5. PROPOSED METHOD 2: SEMI-BLIND METHOD

Although the proposed method 1 performs better under heavily reverberant conditions, it is still difficult to achieve the optimum performance when the level of interference is not severe. Therefore, we need a system that can select the signal processing method to be applied according to the environmental conditions. Assuming that a spoken-dialogue robot system usually has an image sensor, in this paper, we utilize the user position information provided by the robot’s camera and develop a multimodal switching scheme based on distance information. For the current method, an RGB camera with depth sensor is utilized to obtain the speaker distance information.

The block diagram of the proposed semi-blind scheme is shown in Fig. 8. Here we assume that the user distance corresponds to the severity of the interference. First, an offline training stage is conducted to estimate the distance
at which the method should be switched. Both FD-BSE and the proposed method 1 are applied to signals at various speaker-to-microphone distances and the results are compared. Next, after the switching distance has been decided, the system applies two different schemes according to the user position:

- For a short user distance, only FD-BSE estimation is applied to the input signal. The extracted speech from FD-BSE becomes the output signal.
- For a longer user distance, the proposed method 1 is applied instead of only FD-BSE.

The switching point also depends on the SNR condition of the signal. The SNR can be easily approximated by utilizing the noise estimation from FD-BSE. The calculation is carried out channel-wise, as given by

$$\text{SNR}_{\text{est}} = 10 \log_{10} \frac{E[x(t)]^2 - E[x_N(t)]^2}{E[x_N(t)]^2}.$$  (40)

The lower average SNR indicates not only more severe background noise but also late reverberation, because the noise estimation from FD-BSE may also contain the late reverberation components since they have similar characteristics. In this case, the proposed method 1 is more preferable, thus the switching distance is shorter than in the case of higher SNR.

The average word recognition performance is depicted in Fig. 9. We compare the performance of the proposed method 2 with the estimated speech from FD-BSE and the proposed method 1. On average, the proposed method 2 achieves better word recognition accuracy than other methods. This shows that the semi-blind approach can maintain optimum performance regardless of the interference condition, because it can compensate the low recognition accuracy of the proposed method 1 at a shorter user distance.

6. CONCLUSION AND FUTURE WORK

We have proposed an automatic optimized method to suppress diffuse background noise and late reverberation, combining FD-BSE and multichannel WF. We apply an optimization scheme that is based on musical noise assessment via higher-order statistics and acoustic model likelihood. The complicated optimization problem is simplified by successively updating the parameter of multichannel WF. An experimental evaluation showed that the proposed method achieves the best performance in terms of speech recognition accuracy under severe acoustical environments, while FD-BSE alone performs better when the level of interference is light to moderate.

Under the assumption that a spoken-dialogue system is usually integrated with an image sensor, we have developed a semi-blind method by utilizing user distance information obtained from the image sensor to determine which speech enhancement method should be applied to the input signal. This semi-blind scheme can compensate the performance of the proposed method in less severe acoustical environments, resulting in optimum recognition accuracy regardless of the level of interference.

Future work may include greater utilization of the image sensor for better initialization of the FD-BSE filter and user tracking to increase the robustness of the method. The parameter update strategy may also be improved, particularly under less severe acoustic conditions.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process., ASSP-27, 113–120 (1979).
[2] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proc. IEEE, 67, 1586–1604 (1979).
[3] J. L. Flanagan, J. D. Johnston, R. Zahn and G. W. Elko, “Computer-steered microphone arrays for sound transduction in large rooms,” J. Acoust. Soc. Am., 78, 1508–1518 (1985).
[4] H. F. Silverman and W. R. Patterson, “Visualizing the performance of large-aperture microphone arrays,” Proc. ICASSP, pp. 962–972 (1999).
[5] O. L. Frost, “An algorithm for linear constrained adaptive array processing,” Proc. IEEE, 60, 926–935 (1972).
[6] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” Neural
higher-order statistics,” *Proc. ICASSP*, pp. 5076–5079 (2011).

[27] F. D. Apriliyanti, H. Saruwatari, K. Shikano and T. Takatani, “Optimization scheme of joint noise suppression and dereverberation based on higher-order statistics,” *Proc. APSIPA ASC*, 6 pages (2012).

[28] H. Saruwatari, Y. Ishikawa, T. 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.*, 19, 1457–1466 (2011).

[29] M. L. Seltzer and R. M. Stern, “Subband likelihood-maximizing beamforming for speech recognition in reverberant environments,” *IEEE Trans. Audio Speech Lang. Process.*, 14, 2109–2121 (2006).

[30] B. BabaAli, H. Sameti and M. Safayani, “Likelihood-maximizing-based multiband spectral subtraction for robust speech recognition,” *EURASIP J. Adv. Signal Process.*, 2009(12), 15 pages (2009).

[31] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Inc., Upper Saddle River, 1993), pp. 339–342.

[32] K. Kinoshita, M. Delcroix, T. Nakatani and M. Miyoshi, “Suppression of late reverberation effect on speech signal using long-term multistep linear prediction,” *IEEE Trans. Speech Audio Process.*, 17, 534–545 (2009).

[33] N. Q. K. Duong, E. Vincent and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Trans. Speech Audio Process.*, 18, 1830–1840 (2010).

[34] P. C. Loizou, *Speech Enhancement Theory and Practice* (CRC, Taylor & Francis Group, Boca Raton, FL, 2007).

[35] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Iitahashi, “INAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” *J. Acoust. Soc. Jpn. (E)*, 20, 196–206 (1999).

[36] A. Lee, T. Kawahara and K. Shikano, “Julius—An open source realtime large vocabulary recognition engine,” *Proc. Ear. Conf. Speech Commun. Technol.*, pp. 1691–1694 (2001).
source separation, and statistical signal processing. He received paper awards from the IEICE in 2001 and 2006, from the Telecommunications Advancement Foundation in 2004, 2009 and 2011, and at the IEEE-IROS2005 in 2006. He won the first prize at the IEEE MLSP2007 Data Analysis Competition for BSS. He is a member of IEEE, IEICE, Japan VR Society, and ASJ.

Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1970, 1972, and 1980, respectively. He is a Professor emeritus at the Nara Institute of Science and Technology (NAIST), where he was directing the Speech and Acoustics Laboratory. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990 Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, the IPSJ Yamashita SIG Research Award in 2000, and the Paper Award from the Virtual Reality Society of Japan in 2001, the IEICE paper award in 2005 and 2006, and the Inose award in 2005. He is a fellow of IEEE, IEICE and Information Processing Society of Japan, and a member of ASJ and ISCA.

Satoshi Nakamura received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was a director of ATR Spoken Language Communication Research Laboratories in 2000–2008, and a vice president of ATR in 2007–2008. He was a director general of Keihanna Research Laboratories, National Institute of Information and Communications Technology, Japan in 2009–2010. He is currently a professor and a director of Augmented Human Communication laboratory at NAIST. He is interested in modeling and systems of spoken dialog system, speech-to-speech translation. He is one of the leaders of speech-to-speech translation research projects including C-STAR, IWSLT and A-STAR. He headed the world first network-based commercial speech-to-speech translation service for 3-G mobile phones in 2007 and VoiceTra project for iPhone in 2010. He received LREC Antonio Zampoli Award, the Commendation for Science and Technology by the Ministry of Science and Technology in Japan. He is an elected board member of ISCA and an elected member of IEEE SPS, speech and language TC.

Tomoya Takatani received the B.E. degree in electronics from Doshisha University in 2001. He received the M.E. degree and Ph.D. degrees in engineering from Nara Institute of Science and Technology (NAIST) in 2003 and 2006, respectively. Since 2006, he has been working at Toyota Motor Corporation. Since 2012, he is member of KIBO ROBOT PROJECT. He received 1st Itakura Prize Innovative Young Researcher Award in 2006 from ASJ, Paper awards in 2006 from IEICE, Best Application Paper Award in 2006 from IEEE/RSJ IROS2005. His research interests include speech enhancement, array signal processing, blind source separation, machine learning, robot audition. He is a member of IEICE and ASJ.