Joint Noise Suppression and Dereverberation Combining Frequency-Domain Blind Signal Extraction and Multichannel Wiener Filter for Hands-Free Spoken Dialogue System *

○Fine Dwinita Aprilyanti, Hiroshi Saruwatari, Satoshi Nakamura (NAIST), Tomoya Takatani (TOYOTA)

1 Introduction

Hands-free spoken dialogue system provides convenience and flexibility for user in human-machine interaction. This system uses an array of microphones to capture user's utterance from the distance, and it is preferable to be implemented in real environment rather than close-talking system, where the user must use hand-held or body-mounted microphone. However, the effect of interference sound, such as background noise and reverberation, is more severe in this system, thus degrades its performance[1].

To solve this problem, many speech enhancement methods utilizing microphone array has been studied. The blind signal separation (BSS) aims to separate the signal component, for example by emphasizing on the statistical characteristics of the component of each signal, also known as independent component analysis (ICA)[2]. The blind spatial subtraction array (BSSA)[3] method has shown that frequency-domain ICA (FD-ICA)[4] performs better in estimating the diffuse background noise than the target speech. Hence, a combination technique, which consists of FD-ICA as a noise estimator and nonlinear postprocessing to suppress the estimated noise, has been proved to be effective in improving the target sound quality.

Several methods have been proposed to suppress the effect of both background noise and reverberation. Some of the authors have proposed a method that is based on frequency domain blind signal extraction (FD-BSE)[3] combined with two stages of multichannel Wiener filter (WF) as nonlinear postfilters[6]. An optimization scheme has also been proposed for this method that is based on the assessment of generated musical noise as an effect of nonlinear postprocessing[7]. The method performs well in heavily reverberant environment, but fails to achieve optimum speech recognition performance when interferences are not so severe.

In this paper, we propose improvement for the joint noise suppression and dereverberation method based on FD-BSE and multichannel WF. This joint method takes advantage of the FD-BSE capability to separate speech component from noise mixture. Assume that the speech extraction is effective, the late reverberation is synthesized and suppressed directly from the speech estimation of FD-BSE. Therefore, the processing path becomes shorter and target speech distortion due to processing can be mitigated.

The remaining of this paper will be organized as follows; Section 2 shows the sound mixing model and review on FD-BSE. Section 3 describes the proposed method including optimization scheme. The experimental result and ASR performance evaluation are presented in Section 4. Finally, we conclude our work in Section 5.

2 Related Work

2.1 Sound Mixing Model at Microphone Array

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

$$\boldsymbol{x}(t) = \boldsymbol{x}_S(t) + \boldsymbol{n}(t), \quad (1)$$

$$\boldsymbol{x}_S(t) = \left(\boldsymbol{h}_E(\tau) + \boldsymbol{h}_L(\tau)\right) * \boldsymbol{s}(t), \quad (2)$$

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 responses. Most hidden Markov

*周波数領域ブラインド信号抽出と多チャネルウィーナフィルタを用いたハンズフリー音声対話向けブラインド雑音・残響抑圧。
○アプリリヤンティフィネ、猿渡洋、中村智（奈良先端大）、高谷智哉（トヨタ）

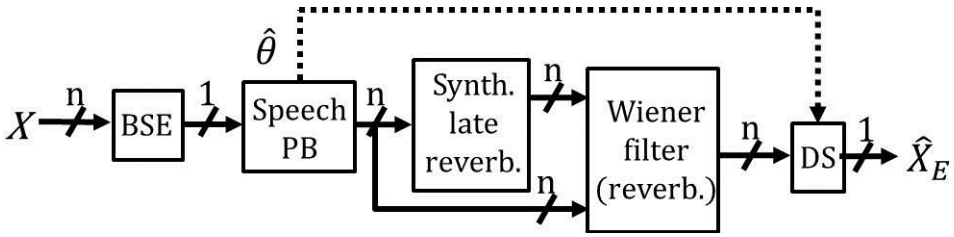

Figure 1 Block diagram of FD-BSE-based joint method

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. Therefore, many speech dereverberation methods focus on suppressing the effect of late reverberation instead of attempting to recover the non-reverberant signal.

In time-frequency domain, we can simplify the mixture model in each frequency bin without explicitly separate the early and late reverberation component, as given by

$$\boldsymbol{X}(f,t) \approx \boldsymbol{H}_\theta S(f,t) + \boldsymbol{N}(f,t), \quad (3)$$

where $S(f,t)$ is the clean speech component and $\boldsymbol{N}(f,t)$ is a vector containing n components of diffuse background noise.

With no loss of generality, we can reformulate the Eq. 2 as given by

$$\boldsymbol{X}(f,t) = \boldsymbol{A}(f)\boldsymbol{S}(f,t). \quad (4)$$

Here, we assume that $S_1 = S(f,t)$ and $[S_2(f,t), \dots, S_n(f,t)] = \boldsymbol{N}(f,t)$. It is also realistic to assume that, in frequency bin, the speech component is statistically independent to noise component.

2.2 Frequency-Domain Blind Signal Extraction

Given the observation $\boldsymbol{X}(f,t)$, FD-BSE works by applying extracting vector $\boldsymbol{B}(f)$ in each frequency bin to obtain $Y(f,t)$, as given by

$$Y(f,t) = \widehat{X_S}(f,t) = \boldsymbol{B}(f)\boldsymbol{X}(f,t), \quad (5)$$

On the contrary to the conventional FD-BSS, FD-BSE algorithm only aims at extracting the speech components from noise mixture, rather than separating signal according to each source. Therefore, the statistical independence within each noise source is not required.

The extracting vector is updated using gradient descent method to minimize the cost function given by

$$J(\boldsymbol{B}(f)) = \frac{1}{2} E\{|Y(f,t)|\}^2, \quad (6)$$

$$E\{|Y(f,t)|^2\} = 1 \ . \quad (7)$$

This cost function is minimized if the extracted component has a modulus with a small mean and a large variance. This implies the sparseness of the modulus of extracted component, in the sense that only a few of the values are significantly large and the rest is close to zero. This is also applied in speech and noise mixture, where speech component is considered sparser than diffuse background noise that has more flat distribution. Consequently, Eq. (6) is minimized when the target speech component is extracted.

3 Proposed Joint Method

3.1 Main Algorithm

The capability of FD-BSE in suppressing diffuse background noise has been confirmed in previous work[5]. However, this method cannot perform well in heavily reverberant environment. Therefore, an additional postprocess is required to improve its performance.

The block diagram of the proposed joint method is shown in Fig. 1. In this proposed method, FD-BSE is used to extract the target speech component from the noise mixture. Assume that the process is effective, the extracted component will only consist of reverberant speech. The next step is to synthesize and suppress the late reverberation component from the target speech. The estimation of late reverberation can be separated into two task: estimating late impulse response $\boldsymbol{h}_L(\tau)$ and clean speech signal $\boldsymbol{s}(t)$.

In this method, $\boldsymbol{h}_L(\tau)$ is approximated by generating synthetic tail from decayed Gaussian random variable $\mathrm{u}(\tau)$[8], given by

$$\boldsymbol{h}_L(\tau) = au(\tau)e^{-d(\tau-\tau_d)}, \quad (8)$$

$$d = \frac{\ln 10^6}{2(T_{60}-\tau_d)}, \quad (9)$$

where a is the scaling factor. Consequently, the synthesize of $\boldsymbol{h}_L(\tau)$ requires *a priori* information of T_{60}.

The direct speech $\boldsymbol{s}(t)$ is approximated by projecting back the ouput of FD-BSE $\widehat{X_S}(f,t)$ to the truncated FD-BSE filter. Then, according to equation (2), the estimate $\widehat{\boldsymbol{X}_L}(f,t)$ is obtained by applying convolution in the time domain. After that, a set of multichannel Wiener filter is applied to suppressed the late reverberation component, as given by

$$\widehat{X_E}(f,t) = G\left|\widehat{X_S}(f,t)\right| e^{jarg\left(\widehat{X_S}(f,t)\right)}, \qquad (10)$$

$$G = \frac{|\widehat{X_S}(f,t)|^2}{|\widehat{X_S}(f,t)|^2 + \beta_R|\widehat{X_L}(f,t)|^2}, \qquad (11)$$

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

3.2 Optimization Strategy for Joint Method

By combining FD-BSE and WF filters in the proposed joint method, we gain more flexibility in suppressing the interferences, thus making the method more robust to various acoustical conditions. On the other hand, it is required to find the value of parameter β_R to gain the optimum performance of this method.

Since the proposed method is intended to be implemented in hands-free spoken dialogue system, the method should achieve best performance in terms of speech recognition accuracy. One common measure to evaluate the performance is word accuracy (WA), given by

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

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.

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

$$\widehat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} P(\boldsymbol{W}|\boldsymbol{o}(\beta_R)). \qquad (13)$$

However, the recognition system cannot compute the posterior probability $P(\boldsymbol{W}|\boldsymbol{o}(\beta_R))$ directly. Instead, Eq. (13) is transformed into the following form based on Bayes' theorem:

$$\widehat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \frac{P(\boldsymbol{o}(\beta_R)|\boldsymbol{W})P(\boldsymbol{W})}{P(\boldsymbol{o}(\beta_R))}, \qquad (14)$$

where $P(\boldsymbol{o}(\beta_R)|\boldsymbol{W})$ 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. (14) is maximized with respect to the word sequence $\boldsymbol{W}$ for a given observed sequence $\boldsymbol{o}$ that is fixed, the denominator term $P(\boldsymbol{o}(\beta_R))$ can be ignored. Thus, the parameter β_R is optimized by maximizing the likelihood of acoustic model of speech recognizer, as written by

$$\widehat{\beta_{\mathrm{R}}} = \arg\max_{\boldsymbol{\beta_{\mathrm{R}}}} P(\boldsymbol{o}(\beta_{\mathrm{R}})|\boldsymbol{W}). \qquad (15)$$

4 Experiment and Result

We used 25 utterances from female JNAS database[8] as the source signals. 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.

Speech is convoluted with room impulse response, and was mixed with noise signal at SNR 10 dB. 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.

The objective evaluation and recognition results are shown in Fig. 2. For the recognition task, Julius 4.2[9] is used as the decoder and noise-free early reverberated speech is used as reference speech. We compare the output of proposed scheme with the estimated speech from FD-BSE and the former 2-stages joint method, which is optimized via higher-order statistics and

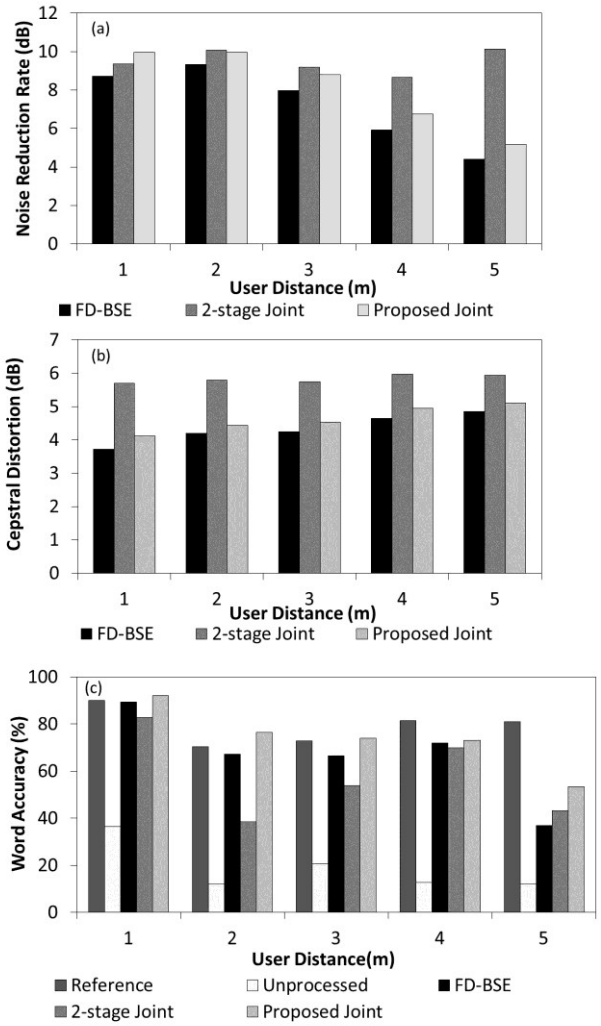

Figure 2 Experimental result of proposed method, input signal SNR = 10 dB. (a) NRR result, (b) cepstral distortion, and (c) word accuracy of recognition task.

acoustic likelihood[7]. We also include the recognition result from unprocessed signal.

From the result it is shown that while the NRR result of proposed method is not optimum in most user distances, it succeeded to achieve the best recognition result compared to other methods. The shorter processing in the proposed method results in lower distortion to the target signal, thus the proposed method can achieve higher recognition performance compared to 2-stage joint method.

5 Conclusion

We have proposed a method to jointly suppress diffuse background noise and late reverberation by combining FD-BSE and multichannel WF. Experimental result confirmed the effectiveness of the proposed method in improving the recognition accuracy in various acoustical condition.

Acknowledgement

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

References

[1] R. Prasad, H. Saruwatari, K. Shikano, "Robots that can hear, understand and talk," *Advanced Robotics*, vol.18, no.5, 2004.

[2] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.

[3] Y. Takahashi, *et al.*, "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.

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

[5] J. Even, *et al.*, "Blind signal extraction based speech enhancement in presence of diffuse background noise," *Proc. IEEE SSP*, pp. 513-516, 2009.

[6] J. Even, *et al.*, "Blind signal extraction based joint suppression of diffuse background noise and late reverberation," *Proc. EUSIPCO*, pp. 1534-1538, 2010.

[7] F. D. Aprilyanti, *et al.*, "Optimization scheme of joint noise suppression and dereverberation based on higher-order statistics," *Proc. APSIPA ASC*, Dec. 2012.

[8] K. Lebart and J.M. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol.87, pp.359--356, 2001.

[9] 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.

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