Joint Suppression of Background Noise and Late Reverberation Combining Blind Speech Extraction and Generalized MMSE-STSA Estimator*

☆Fine Dwinita Aprilyanti, Hiroshi Saruwatari, Satoshi Nakamura (NAIST)

1 Introduction

Hands-free automatic speech recognition (ASR) system provides more convenience and flexibility to the user under many conditions. However, the presence of background noise and room reverberation, particularly the late reverberation component, significantly degrade the speech recognition performance. Thus, array signal processing is required as the front-end for such a system.

One of the microphone array processing techniques is frequency-domain blind source separation (FD-BSS), which separates signal components according to the statistical independence of each source, e.g., by using independent component analysis (ICA)[1]. FD-BSS does not require *a priori* information about the input signal; however, it cannot perform well in the presence of non-point source noise[2]. Frequency-domain blind signal extraction (FD-BSE)[3] to effectively extracts speech from a mixture of speech and noise by utilizing the difference in sparseness between their modulus. However, the reverberation effect is not considered in this method.

In this paper, we utilize a nonlinear postprocessing method to extend the capability of FD-BSE for dereverberation. We introduce the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator[4] and its generalized method[5] to suppress the late reverberation component in addition to diffuse background noise. The related works are reviewed in the next section. In Section 3, we describe our proposed method including the optimization strategy. The experimental evaluation and a discussion of the results are given in Section 4, followed by conclusion in Section 5.

2 Related Works

2.1 Frequency-Domain Blind Signal Extraction

The time-frequency domain model of the signal captured by a microphone array $\boldsymbol{X}(f,k)$ at the fth frequency bin is given by

$$\boldsymbol{X}(f,k) = \boldsymbol{A}(f)\boldsymbol{Z}(f,k), \quad (1)$$

where $\boldsymbol{A}(f)$ is the mixing matrix. Without loss of generality, we may assume that the first-row component of $\boldsymbol{Z}(f,k)$ is the contribution of speech and its reverberation $\boldsymbol{X}_S(f,k)$, and the remainder is the contribution of diffuse noise $\boldsymbol{X}_N(f,k)$. We may also assume that the speech component is statistically independent of the noise component.

In FD-BSE, the extracted output $Y(f,k)$ is obtained by applying an extracting vector to the observed signal, as given by

$$Y_{\mathrm{BSE}}(f,k) = B(f)\boldsymbol{X}(f,k) = B(f)\boldsymbol{A}(f)\boldsymbol{Z}(f,k). \quad (2)$$

The vector $B(f)$ is updated using the gradient descent method to minimize the cost function $J(B(f))$ given by

$$J\big(B(f)\big) = \tfrac{1}{2}E\{|Y_{\mathrm{BSE}}(f,k)|\}^2, \quad (3)$$

$$E\{|Y_{\mathrm{BSE}}(f,k)|^2\} = 1. \quad (4)$$

The speech modulus has a sparser distribution than the diffuse background noise as most of its values are close to zero and only a few are significantly large. Thus, the cost function becomes minimum when the target speech component is extracted.

2.2 Generalized MMSE-STSA Estimator

The generalized MMSE-STSA estimator is also referred to as MMSE estimation with optimizable speech model and inhomogeneous error criterion (MOSIE) estimator[5]. In MOSIE, the probability density function of the clean speech power spectrum is modeled by a chi

---

* ブラインド信号抽出と一般化 MMSE-STSA 推定器を用いた背景雑音・後部残響抑圧
☆アプリリヤンティフィネ、猿渡洋、中村哲（奈良先端大）

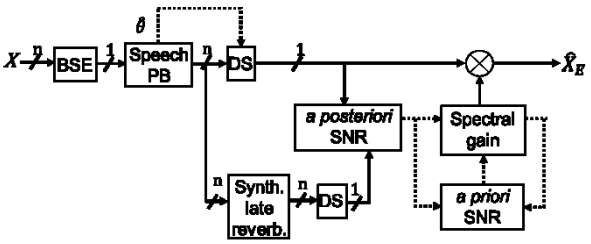

Figure 1 Block diagram of FD-BSE combined with single channel MOSIE estimator.

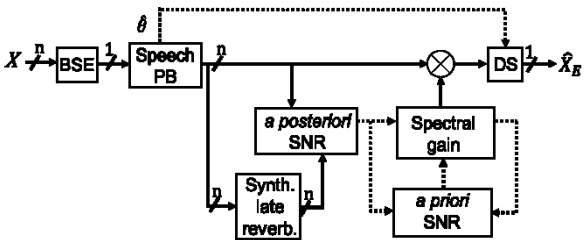

Figure 2 Block diagram of FD-BSE combined with multichannel MOSIE estimator.

distribution as

$$p(s) = \frac{2}{\Gamma(\rho)}\left(\frac{\rho}{P_S(f)}\right)^{\rho} s^{2\rho-1} \exp\left(-\frac{\rho}{P_S(f)}s^2\right), \quad (5)$$

where ρ is the shape parameter of the speech model, $P_S(f)$ is the mean of speech spectral amplitude, and $\Gamma(a)$ is the complete Gamma function. Also, $\rho = 1$ indicates a Gaussian speech signal[4]. In this paper, we employ $0 < \rho < 1$ to model the speech power spectrum as a super-Gaussian signal.

The MOSIE estimator is mainly used for single-channel noise suppression by applying the gain function $G(f,k)$, as given by

$$Y_{\mathrm{MOSIE}}(f,k) = G(f,k)X(f,k), \quad (6)$$

$$G(f,k) = \frac{\sqrt{\upsilon(f,k)}}{\hat{\gamma}(f,k)}\cdot \left[\frac{\gamma(\rho+\frac{\beta}{2})}{\gamma(\rho)}\cdot\frac{\Phi(1-\rho-\frac{\beta}{2},1,-\upsilon(f,k)}{\Phi(1-\rho,1,-\upsilon(f,k))}\right]^{1/\beta} \quad (7)$$

$$\upsilon(f,k) = \frac{\hat{\xi}(f,k)}{1+\hat{\xi}(f,k)}\hat{\gamma}(f,k), \quad (8)$$

where $\Phi(a,b,c)$ is the confluent hypergeometric function. $\hat{\xi}(f,k)$ and $\hat{\gamma}(f,k)$ are the estimated *a priori* and *a posteriori* signal-to-noise ratio (SNR), respectively, as given by

$$\hat{\xi}(f,k) = \alpha\hat{\gamma}(f,k-1)G^2(f,k-1) + (1-\alpha)\max[\hat{\gamma}(f,k)-1,0], \quad (9)$$

$$\hat{\gamma}(f,k) = \frac{|X(f,k)|^2}{|\widehat{X_N}(f,k)|^2}, \quad (10)$$

where α is the forgetting parameter in the decision-directed approach, and β is the compression parameter of the error function given by

$$\mathrm{e}\left(S_o(\mathrm{f},\mathrm{k}), S_p(\mathrm{f},\mathrm{k})\right) = |S_o(\mathrm{f},\mathrm{k})|^{\beta} - \left|S_p(\mathrm{f},\mathrm{k})\right|^{\beta}, \quad (11)$$

where $S_o(\mathrm{f},\mathrm{k})$ and $S_p(\mathrm{f},\mathrm{k})$ are the original and processed speech spectral amplitude, respectively.

3 Proposed Joint Method

The reverberant speech in the mixture in Eq. (1) is composed of clean speech with the room impulse response, as given by

$$X_S(f,k) = X_E(f,k) + X_L(f,k) = S(f,k)H_E(f) + S(f,k)H_L(f), \quad (12)$$

where $H_E(f)$ and $H_L(f)$ indicate the early and late room impulse responses, respectively. This is because the speech signal tends to loses its correlation after some delays, owing to its nonstationary characteristics. Therefore, the late reverberation component can be suppressed in the same manner as additive noise.

3.1 Main Algorithm

Assuming that the diffuse background noise has been suppressed effectively, the extracted component $\widehat{X_S}(f,k)$ will only consist of clean speech and its reverberation. First, we synthesize the late reverberation in the time domain by applying convolution according to Eq. (12). The late room impulse response is estimated by

$$\boldsymbol{h}_L(\tau) = u(\tau)e^{-d(\tau-\tau_d)}, \quad (13)$$

$$d = \frac{\ln 10^6}{2(T_{60}-\tau_d)}, \quad (14)$$

where $u(\tau)$ is a Gaussian random function, τ_d is the cutoff time between early and late impulse responses, and T_{60} is the reverberation time. The clean speech estimate $\hat{s}(t)$ is approximated by projecting back $\widehat{X_S}(f,k)$ to the truncated FD-BSE filter.

Next, we will compare the performance of single-channel and multichannel MOSIE estimator postprocessing, as shown in Fig. 1 and Fig.2, respectively. Parametric postprocessing allows flexible control of the level of dereverberation. However, it is important to set the parameter to obtain the optimum output. In this paper, we only focus on optimizing the shape

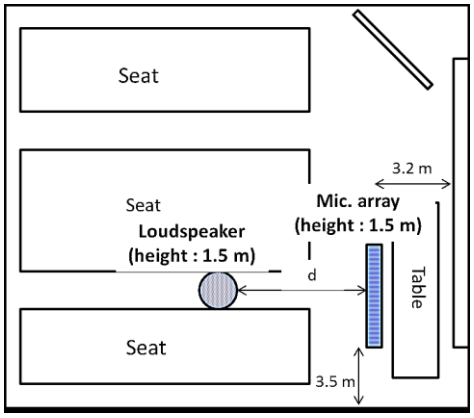

Figure 3 Room configuration for the experiment.

parameter ρ. Other parameters, such as T_{60}, are assumed to be known.

3.2 Optimization Scheme Based on Acoustic Likelihood

Signal processing front-end in ASR system can only be expected to improve the recognition performance if it generates outputs that maximize the probability of the correct transcription relative to other possible candidates. Therefore, we can optimize the parameter of the front-end based on the likelihood in the acoustic model of ASR.

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

$$\widehat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} P(\boldsymbol{W}|\boldsymbol{o}(\rho)). \quad (15)$$

However, the posterior probability $P(\boldsymbol{W}|\boldsymbol{o}(\rho))$ cannot be computed directly. Thus, Bayes' theorem is applied, as given by

$$\widehat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \frac{P(\boldsymbol{o}(\rho)|\boldsymbol{W})P(\boldsymbol{W})}{P(\boldsymbol{o}(\rho))}. \quad (16)$$

$P(\boldsymbol{o}(\rho)|\boldsymbol{W})$ is the acoustic likelihood (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 $\boldsymbol{W}$.

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 $P(\boldsymbol{o}(\rho))$ can be ignored. Thus, the parameter

Table 1 Speech recognizer specification

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

ρ is optimized by maximizing the likelihood of acoustic model, as written by

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

This optimization problem requires the correct transcription $\boldsymbol{W}$. However, if it is already known, the speech recognizer is not required anymore. Therefore, in practice, we apply MOSIE estimator postprocessing using a set of ρ parameters, from which the output with the highest acoustic score is chosen. $\boldsymbol{W}$ is selected from the result of the first iteration.

4 Experimental Evaluation

Two experiments have been carried out for evaluation purposes. An eight-channel microphone array (inter-microphone spacing of 2.1 cm) was used to record the room impulse response with the configuration shown in Fig. 3. The estimated T_{60} is 500 ms. The observed signal was created by convolution of the clean speech with the impulse response, and recorded real noise was added at SNR of 10 dB.

Recognition performance was evaluated using Julius with the specifications shown in Table 1. Since the goal of the proposed method is to suppress diffuse background noise and late reverberation, we used clean speech convoluted with the early impulse response as the baseline. The word accuracy measure used in the evaluation is calculated as

$$\mathrm{WA} = 100 \times \frac{N-(I+S+D)}{N}, \quad (18)$$

where N, I, S, and D are the number of words in the correct transcription and the number of insertions, substitutions, and deletions,

Table 2 Word accuracy results for Experiment 1

Distance	1 m	2 m	3 m	4 m	5 m
Baseline	97.20	90.22	90.11	92.91	92.85
FD-BSE	89.65	61.55	56.32	70.59	41.24
FD-BSE + SC-MOSIE $\alpha = 0.98$	74.44	56.36	55.97	61.94	36.04
FD-BSE + MC-MOSIE $\alpha = 0.96$	91.79	75.00	73.13	85.07	58.56
FD-BSE + MC-MOSIE $\alpha = 0.98$	95.50	65.91	72.07	82.84	52.25
FD-BSE + MOSIE-LSA $\alpha = 0.98$	71.64	60.23	61.26	73.87	48.86

respectively. The frequency domain processing was carried out with a 512-point Hamming window and 50% overlap of the STFT. FD-BSE was performed in 600 iterations with an adaptation step of 0.3. The parameter τ_d was set to 75 ms, corresponding to the delay that can still be handled by the speech recognizer.

The first experiment was carried out using 5-male and 5-female utterances from JNAS database. We employed several parameter sets, e.g., β of 0.001 to represent the MOSIE-LSA estimator and 1 to represent the MOSIE-STSA estimator[5]. We manually selected the best result among these combinations.

The results are shown in Table 2. It is shown that the multichannel MOSIE estimator improves the recognition accuracy compared with FD-BSE. On the other hand, the single-channel MOSIE-STSA estimator's performance is inferior. This is understandable as single-channel processing tends to result in higher speech output distortion.

It can be observed that the MOSIE-STSA estimator performs better than the MOSIE-LSA estimator for dereverberation. It is also shown that the $\alpha = 0.96$ results in better word accuracy than $\alpha = 0.98$, which is known to give the optimum result for speech enhancement such as that for hearing aid system. This may be because a high quality output signal waveform is less important for speech recognition purposes.

The optimization scheme was evaluated in the second experiment. We used the utterances from 50-male and 50-female speakers. Parameters α and β were set to 0.96 and 1, respectively. The results in Table 3 show that the optimized methods outperform FD-BSE with an average

Table 3 Word accuracy results for Experiment 2

Distance	1 m	2 m	3 m	4 m	5 m
Baseline	90.86	78.20	80.30	83.96	84.97
FD-BSE	77.50	52.99	47.59	60.63	33.80
FD-BSE+ MC-MOSIE	83.82	65.75	64.02	73.38	50.05

improvement of 12.9%. This implies that optimization based on acoustic likelihood is effective for our method.

5 Conclusion

We combined FD-BSE and MOSIE estimator to suppress the diffuse background noise and late reverberation for a hands-free ASR system, which was optimized based on the acoustic likelihood. Experimental results show that the proposed method improves the word recognition accuracy compared with the FD-BSE method. Future work may include the utilization of FD-BSE in estimating the speech and late reverberation statistical model.

References

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

[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, pp. 650–664, 2009.

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

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 6, pp. 1109.1121, 1984.

[5] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient condition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 2, pp. 277–289, 2011.