Speech Enhancement Based on Distribution of Sound Source Localization Results

Naoki Shinohara and Kenji Suyama
School of Engineering, Tokyo Denki University,
5 Senju-Asahi-cho, Adachi-ku, Tokyo, 120-8551, Japan

Abstract: In this paper, a target sound enhancement method by using an adaptive microphone array is studied. Then, a supervised signal is required for learning filters. Under a stationary noise environment, it is easy to generate the supervised signal from a power envelope of the received signal. Otherwise, it is difficult to generate the supervised signal under a nonstationary noise environment by the same way. In the proposed method, the supervised signal is generated based on the distribution of sound source localization results. Several experimental results in a real environment are shown to present the effectiveness of the proposed method.

1. Introduction

Target sound enhancement is an important acoustic signal processing technique applied to a voice recognition system. In this application, a small size of system and a low cost implementation are required. Therefore, just two microphones are used in the proposed system. In the system, 2 channel linear filter is used to enhance the target signal. Then, a learning of filters is needed.

AMNOR (Adaptive Microphone array for NOise Reduction)[1] and SP-SDR-MWF (Spatially Pre–processed Speech Distortion Regularized Multichannel Wiener Filter)[2] have been proposed as the target sound enhancement method using the linear filters. In these methods, a detection of a noise section is required.

In the proposed method, the supervised signal is generated based on the cue signal method[3]. It is considered that an accuracy of sound source localization results has a strong relationship with S/N of the received signal. Therefore, a distribution of sound source localization results is used to generate the supervised signal. Several experimental results in a real environment are shown to present the effectiveness of the proposed method.

2. Problem formulation

The target sound source, $s(n)$, where n is discrete time, is received by two microphones. The received signal of the m–th microphone, $x_m(n), m = 1, 2$, can be written in a time domain as below,

$$x_m(n) = h_m^s(n) * s(n) + \gamma_m(n), \tag{1}$$

where $*$ denotes the convolution operation, $h_m^s(n)$ is the impulse response between the m–th microphone and the target sound source, $\gamma_m(n)$ is the noise signal. The aim of target sound enhancement is to emphasize $s(n)$ while suppressing $\gamma_m(n)$.

It is easily assumed that a vowel part has a periodic structure and occupies a large part of a voice power. Therefore, $s(n)$ can be written as $s(n) = a(n)z(n)$, where $z(n)$ is a periodic signal and $a(n)$ is an envelope. Furthermore, it is assumed that $a(n)$ is slower than $z(n)$. Then, Equation (1) can be written as below,

$$\begin{aligned} x_m(n) &= h_m^s(n) * s(n) + \gamma_m(n) \\ &= a(n)\left(h_m^s(n) * z(n)\right) + \gamma_m(n) \\ &= a(n)z^{'}(n) + \gamma_m(n) \end{aligned} \tag{2}$$

where $z^{'}(n) = h_m^s(n) * z(n)$.

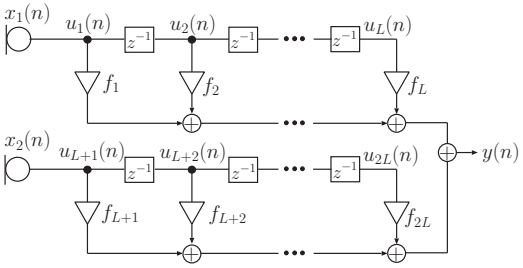

Figure 1. 2ch linear filter

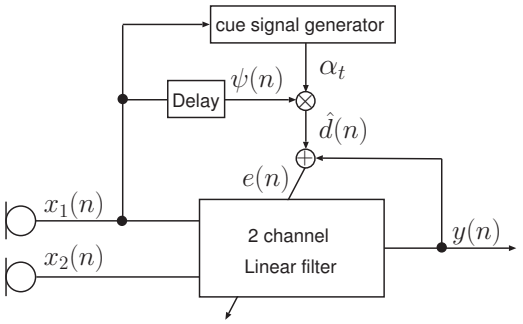

Figure 2. A procedure of cue signal method

3. Cue signal method

In general, filter coefficients of 2 channel linear filter are learned so as to minimize $\overline{e^2(n)}$, where $e(n)$ is an error between the supervised signal $d(n)$ and the output signal $y(n)$, $\bar{}$ denotes a time average operation. In our method, $d(n)$ is generated based on the cue signal method, and a procedure of the cue signal method is shown in Figure 2. Under a stationary noise environment, the cue signal α_t is generated as a signal that satisfies following three conditions.

1. α_t has a correlation with the power envelope of $s(n)$.
2. α_t does not have a correlation with $\gamma_m(n)$.
3. Time average of α_t is 0.

Then, $\hat{d}(n)$ is generated as

$$\hat{d}(n) = \alpha_t \psi(n) \tag{3}$$

where $\psi(n)$ is the $L/2$ samples delayed version of $x_1(n)$, L is a filter length. Moreover, $\psi(n)$ can be written as following,

$$\begin{aligned}\psi(n) &= a\psi_z(n) + \psi_\gamma(n), \\ &= \psi_s(n) + \psi_\gamma(n) \end{aligned} \tag{4}$$

where $\psi_z(n)$ is a periodical component of the target sound, $\psi_\gamma(n)$ is the noise. The filter coefficients can be calculated as a solution of following equation,

$$\boldsymbol{f} = {\boldsymbol{R_c}}^{-1}\boldsymbol{p} \tag{5}$$

where $\boldsymbol{f}$ is a vector of filter coefficients, $\boldsymbol{R_c}$ is an autocorrelation matrix of $u(n)$, $\boldsymbol{p}$ is a cross correlation vector between tap input $u(n)$ and $\hat{d}(n)$. Equation (5) shows that the supervised signal depends on only $\boldsymbol{p}$. Therefore, if $p_l(d)$, $l = 1, 2, \cdots, 2L$, and $p_l(\hat{d})$ are equivalent each other, where $p_l(d)$ and $p_l(\hat{d})$ are the l–th element of $\boldsymbol{p}$ calculated using $d(n)$ and $\hat{d}(n)$, the resultant filters have a common characteristic. Then, $p_l(d)$ and $p_l(\hat{d})$ can be written as following,

$$\begin{aligned} p_l(d) &= \overline{u_l(n)d(n)} \\ &= \overline{u_l(n)\psi_s(n)} \\ &= \overline{\left(a(n)z_l'(n) + \gamma_l(n)\right)\left(a(n)\psi_z(n)\right)} \\ &= \overline{a^2(n)z_l'(n)\psi_z(n)} + \overline{a(n)\gamma_l(n)\psi_z(n)} \\ &= \overline{a^2(n)} \cdot \overline{z_l'(n)\psi_z(n)} + \overline{a(n)} \cdot \overline{\gamma_l(n)\psi_z(n)} \\ &= \overline{a^2(n)} \cdot \overline{z_l'(n)\psi_z(n)}, \end{aligned} \tag{6}$$

$$\begin{aligned} p_l(\hat{d}) &= \overline{u_l(n)\hat{d}(n)} \\ &= \overline{u_l(n)\alpha_t\psi(n)} \\ &= \overline{\left(a(n)z_l'(n) + \gamma_l(n)\right)\alpha_t\left(a(n)\psi_z(n) + \psi_\gamma(n)\right)} \\ &= \overline{\alpha_t a^2(n)z_l'(n)\psi_z(n)} + \overline{\alpha_t a(n)\gamma_l(n)\psi_z(n)} \\ &\quad + \overline{\alpha_t a(n)z_l'(n)\psi_\gamma(n)} + \overline{\alpha_t\gamma_l(n)\psi_\gamma(n)} \\ &= \overline{\alpha_t a^2(n)} \cdot \overline{z_l'(n)\psi_z(n)} + \overline{\alpha_t a(n)} \cdot \overline{\gamma_l(n)\psi_z(n)} \\ &\quad + \overline{\alpha_t a(n)} \cdot \overline{z_l'(n)\psi_\gamma(n)} + \overline{\alpha_t} \cdot \overline{\gamma_l(n)\psi_\gamma(n)} \\ &= \overline{\alpha_t a^2(n)} \cdot \overline{z_l'(n)\psi_z(n)} \\ &= K \cdot \overline{a^2(n)} \cdot \overline{z_l'(n)\psi_z(n)}, \end{aligned} \tag{7}$$

where $u_l(n)$ is the l–th tap input, $z'_l(n)$ is the periodic component in $u_l(n)$, $\gamma_l(n)$ is the noise signal in the l–th tap and K is a constant. K can be written as following,

$$K = \frac{\overline{\alpha_t a^2(n)}}{\overline{a^2(n)}}. \tag{8}$$

The difference between Equation (6) and Equation (7) is just K. Although the power envelope of the received signal can be utilized to hold the conditions under the stationary noise situation, it is difficult to hold them under the nonstationary noise situation. Therefore, the condition 1 and the condition 2 are extended to that α_t has a correlation with S/N of the received signals.

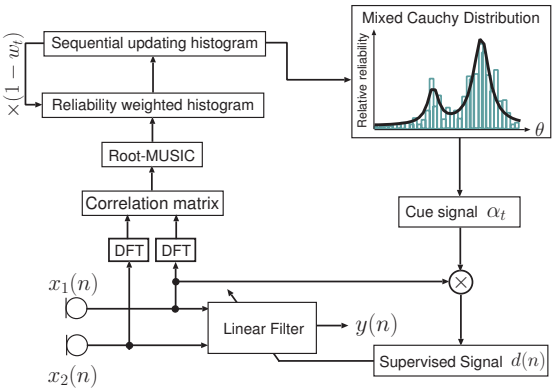

Figure 3. A procedure of the proposed method

4. Proposed method

A procedure of the proposed method is shown in Figure 3. First, the received signals are transformed into the frequency domain by DFT (Discrete Fourier Transform). For applying Root-MUSIC (MUltiple SIgnal Classification), a correlation matrix $\boldsymbol{R}(t,k)$ is calculated as below,

$$\boldsymbol{R}(t,k) = \boldsymbol{X}(t,k)\boldsymbol{X}^H(t,k) + \beta\boldsymbol{R}(t-1,k) \tag{9}$$

$$\boldsymbol{X}(t,k) = [X_1(t,k),\ X_2(t,k)]^T, \tag{10}$$

where t is a frame index, k is a frequency index, $X_m(t,k)$ is DFT of $x_m(n)$, H is a Hermitian transposition, T is a transposition and β is a forgetting factor. In each the time-frequency region, DOA (Direction-Of-Arrival) is estimated using Root-MUSIC. The power ratio $w_p(t,k)$ are weighted as a reliability of estimated result, and the reliability weighted histogram is calculated. $w_p(t,k)$ is calculated as below,

$$w_p(t,k) = \frac{I(t,k)}{\sum_k I(t,k)}, \tag{11}$$

where $I(t,k) = (|X_1(t,k)|^2 + |X_2(t,k)|^2)/2$. The reliability weighted histogram is updated using the updating weight $w_t(t,k)$, and the updated histogram is called as a sequential updating histogram[4]. The mixed Cauchy distribution is fitted to the sequential updated histogram by EM (Expectation Maximization) algorithm for sound source localization. The mixed Cauchy distribution is defined as below,

$$F(\theta) = \sum_{i=1}^{N} \rho_i \left[\frac{1}{\pi}\left(\frac{\lambda_i}{(\theta-\mu_i)^2 + \lambda_i^2}\right)\right], \tag{12}$$

where N is the number of sound sources, ρ_i is a mixture ratio, λ_i is a half width at a half maximum, μ_i is a mode value. In the case of high S/N, ρ_i tends to be large and λ_i tends to be narrow. On the other hand, in the case of low S/N, ρ_i tends to be small and λ_i tends to be wide. Relationship between S/N and ρ_i, between S/N and λ_i were verified, and the sounds (speech1 : male voice, speech2 : female voice) were used as a target sound. These results are shown from Figure 4 to Figure 7, respectively. The verification results have revealed that S/N

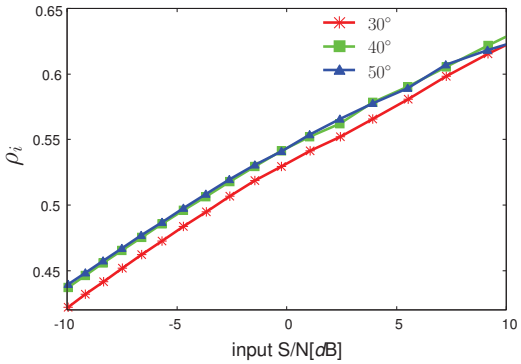

Figure 4. Relationship between S/N and ρ_i (speech1)

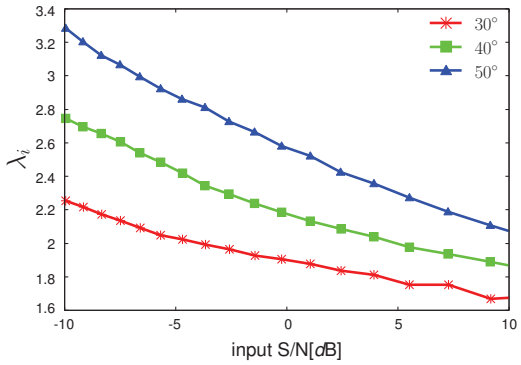

Figure 5. Relationship between S/N and λ_i (speech1)

and ρ_i indicate the relationship of the monotonous increasing and S/N and λ_i indicate the relationship of the monotonous decreasing. Therefore, ρ_i and λ_i can be used for an index of S/N. α_t is generated as following,

$$\alpha_t = \frac{\rho_i}{\lambda_i}. \tag{13}$$

5. Experiments

Several experiments were conducted in an actual room to evaluate the effectiveness of the proposed method. The

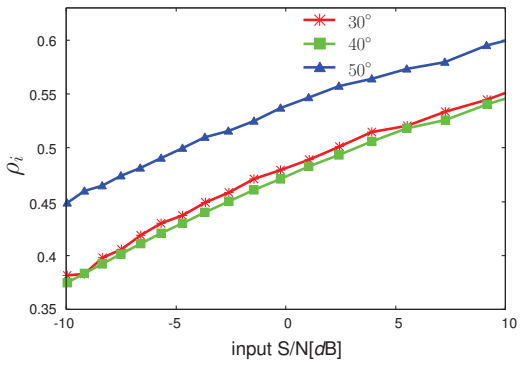

Figure 6. Relationship between S/N and ρ_i (speech2)

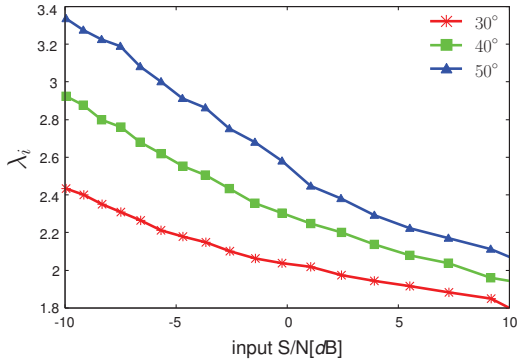

Figure 7. Relationship between S/N and λ_i (speech2)

Table 1. Experimental conditions

sampling frequency	8000 [Hz]
frame size	512
microphone width	40 [mm]
w_t	0.15
β	0.05
frequency band for sound source localization	500–3500 [Hz]
signal length	10 [s]

sounds (speech1 : English male voice, speech2 : English female voice, speech3 : German male voice) were used as the target sound and the Gaussian noise modulated by the sine wave of 2[Hz] was used as the nonstationary noise. Table1 shows parameters used in the experiments. The direction of the target sound was set to $0°$ and the direction of the noise was set to $40°$. The effectiveness of the proposed method was measured by the normalized S/N as following,

$$G = 10\log\frac{\overline{d^2(n)}}{\left(d(n) - K\hat{d}(n)\right)^2}. \tag{14}$$

The eigenvalue ratio method was used as the compared method[5]. In the compared method, cue signal α_t is generated as following,

$$\alpha_t = 1 - \frac{\sigma_2}{\sigma_1}, \tag{15}$$

where σ_1 is the eigenvalue corresponding to the signal subspace, σ_2 is the eigenvalue corresponding to the noise subspace.

The experimental results are shown from Figure 8 to Figure 10. From these results, it can be confirmed that the proposed method worked better than the compared method in the low S/N situation and the case of high S/N in speech2 and speech3.

The performance of the proposed method depends on a precision of the cue signal. Therefore, it is considered that the performance of the proposed method has a limitation because the precision of the cue signal decreases in the silent period.

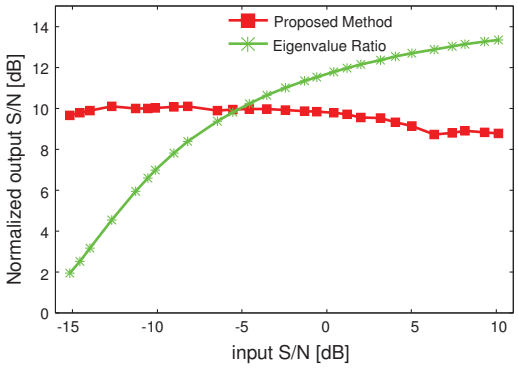

Figure 8. Relationship between input S/N and normalized output S/N (speech1)

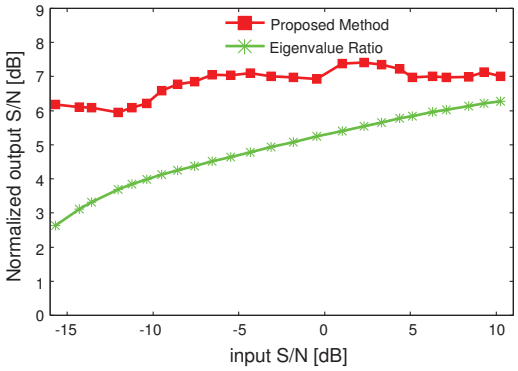

Figure 9. Relationship between input S/N and normalized output S/N (speech2)

Therefore, it can be considered that the proposed method degraded in the situation more than -5dB in the speech1. The compared method degraded particularly in the low S/N. It is supposed that this is a limitation of the compared method caused by a room reverberation.

6. Conclusion

In this paper, the target sound enhancement method by using a adaptive microphone array was proposed. The distribution of sound source localization results was used for generating the supervised signal. The experimental results in the actual room showed the effectiveness of the proposed method.

Acknowledgment

This work was supported by the Research Institute for Science and Technology, Tokyo Denki University, Q16J-03.

References

[1] Y. Kaneda and J. Ohga, “Adaptive Microphone-array System for Noise Reduction,” IEEE Trans, Acoustics, Speech and Signal Process., vol.ASSP-34, no.6, pp.1391–1400, 1986.

[2] A. Spriet, M. Moonen, and J. Wouters, “Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction,” Signal Processing, vol.84, no.12, pp.2367-2387, 2004.

[3] K. Takahashi and H. Yamasaki, “Audio-Visual Sensor Fusion System for Intelligent Sound Sensing,” Proc. of IEEE MFI’94, pp.493–500, 1994.

[4] T. Suzuki and K. Suyama, “An Extension to Multiple Sound Source Tracking of Sequential Updating Histogram Method,” Technical Report of IEICE, vol.112, no.486, pp.81–86, 2013.(in Japanese)

[5] K. Suyama and K. Takahashi, “A Target Sound Extraction Using Linear Filters Based on the Eigenvalues of Signal Space” Journal of signal processing, vol.9, no.3, pp.221-229. 2005.(in Japanese)

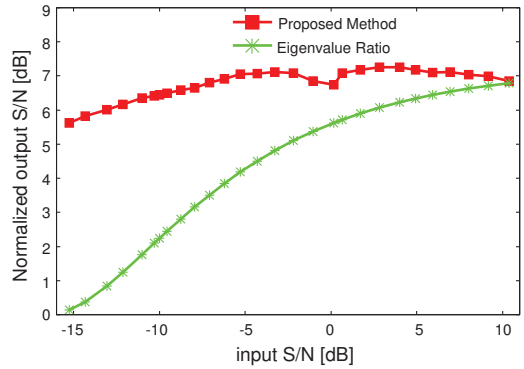

Figure 10. Relationship between input S/N and normalized output S/N (speech3)

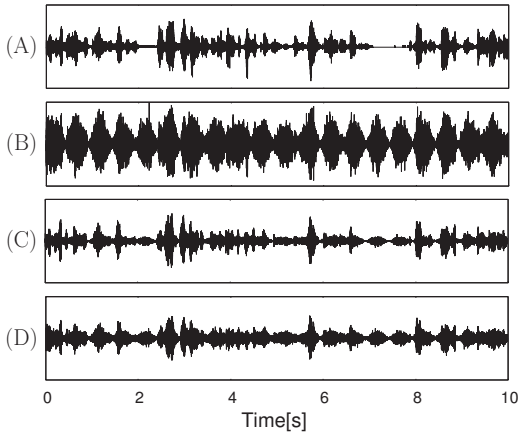

Figure 11. The result of target sound enhancement(speech1) (A) : target source, (B) : received signal, (C) : output signal(proposed method), (D) : output signal(compared method)