A TWO-STAGE MULTICHANNEL ACOUSTIC NOISE REDUCTION UNDER REVERBERANT ENVIRONMENTS

Masahito Togami and Yohei Kawaguchi

Central Research Laboratory, Hitachi Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, 185-8601, Tokyo, Japan
phone: +81-42-323-1111, fax: +81-42-327-7823,
email: { masahito.togami.fe, yohei.kawaguchi.xk }@hitachi.com

ABSTRACT

In this paper, there are two contributions for multichannel noise reduction techniques under reverberant environments. The first contribution is a theoretical analysis of previously proposed multichannel spatial prediction based noise reduction technique (MSP-BF), and discussion about the relationship between MSP-BF and MWF. In MSP-BF, noise sources can be reduced effectively at the first stage by multichannel spatial filtering. The desired sources are distorted after the first stage. The distortion of the desired sources are restored at the second stage by multichannel distortion-restoration filtering. The advantageous point of MSP-BF is that this method can reduce noise sources without the steering vectors of the desired sources. However, there is not a sufficient theoretical analysis in the previous paper. The second contribution is that an alternative noise reduction technique based on the subspace filtering at the first stage, NSR-BF, is proposed. The experimental result under a reverberant environment is shown.

1. INTRODUCTION

For recording systems such as IC recorders or video cameras, the noise reduction problem for the noisy microphone input signal which is recorded under highly reverberant environments is one of the hot research topics. Stationary noise sources that have time-invariant spectrum can be reduced by the conventional noise cancellers. However, nonstationary noise sources cannot be reduced by the noise cancellers. In this paper, the noise sources are assumed to be nonstationary point noise sources. Multichannel noise reduction techniques with a microphone array have been studied [1][2][3][4] for nonstationary noise sources. Blind source separation techniques such as independent component analysis (ICA) which uses mutual independence between sources have been widely studied. However, the mutual independence between sources degrades under highly reverberant environments, and the separation performance degrades under these environments. Generalized sidelobe canceller (GSC) [2] is one of the major beamforming techniques. The number of the desired sources is assumed to be 1. GSC is composed of two stages. At the first stage, a desired source is reduced by using the steering vector of the desired source, and only the noise sources are estimated. At the second stage, the noise sources in the microphone input signal are reduced by using the estimated noise sources. When the steering vector of the desired source can be correctly estimated, then the noise sources can be reduced accurately. However, when the steering vector of the desired source is far from the correct value, the desired source in the output signal is distorted (signal cancellation). To overcome the signal cancellation problem, various robust GSC methods have been studied [5][3], and these methods are successful under less reverberant environments. However, due to the large estimation error of the steering vector, robust GSC methods are not successful under reverberant environments. It is pointed out that the noise reduction framework based on Multichannel Wiener Filtering (MWF) is robust against the estimation error of the steering vector [6]. The cost function of the original MWF is a weighted average of the cost function of desired-source distortion and that of noise reduction. MWF can be executed without the steering vector of the desired source. Furthermore, the number of the desired sources are not limited. Some of the authors proposed an alternative multichannel noise reduction technique which is a combination method of the multichannel spatial prediction based noise reduction and MWF (MSP-BF). MSP-BF can also reduce the noise sources without the steering vector of the desired source. MSP-BF is composed of the two stages. On contrary to the GSC based approaches, only noise sources are reduced without any preknowledge about the desired sources at the first stage. The distortion of the desired sources after the first stage is restored in the second stage. The previous paper [7] is lack of theoretical analysis of the MSP-BF. In this paper, a theoretical analysis is complemented by comparison with MWF. From the analysis, we propose an alternative noise reduction method based on the subspace processing on the first stage, which is more suitable to the purpose of the first stage. The most important thing for the distortion-restoration at the second stage is detection of the period when there are desired sources (the desired-source period). It is difficult to detect the desired-source period, when there are nonstationary noise sources. The advantageous point of the proposed two-stage noise reduction method to MWF is that the proposed method can detect the desired-source period accurately after the first stage, because signal-to-noise ratio after the first stage is higher than that of the microphone input signal. The detection algorithm for the desired-source period driven by the energy based voice activity detection to the output signal after the noise reduction filtering at the first stage is shown. By using this algorithm, the distortion-restoration filter can be trained more correctly. The experimental results under a reverberant environment indicate that the proposed method can reduce noise sources with less distortion of the desired sources than MWF and the previously proposed MSP-BF.

2. PROBLEM STATEMENT

2.1 Setting

We focus on the noise reduction problem under the situation that the noise-only period is obtained in advance but the preknowledge of the desired sources is not assumed to be obtained. The image of the noise reduction problem is shown in Fig. 1.

2.2 Input signal model

In this paper, the noise sources are assumed to be nonstationary point sources. The desired sources are also assumed to be nonstationary point sources. Therefore, the m-th microphone input signal at the sample time t, $x_m(t)$, can be defined as follows:

$$x_m(t) = \sum_{i=0}^{N_s-1} \boldsymbol{h}_{i,m}^T \boldsymbol{s}_i(t) + \sum_{i=0}^{N_n-1} \boldsymbol{g}_{i,m}^T \boldsymbol{n}_i(t) + v_m(t), \qquad (1)$$

where T is the transpose operator of a matrix/vector, N_s is the number of the desired sources, N_n is the number of the noise sources, $v_m(t)$ is the background noise, $\boldsymbol{h}_{i,m}$ is the time-invariant impulse response of the i-th desired source at the m-th microphone, $\boldsymbol{g}_{i,m}$ is the

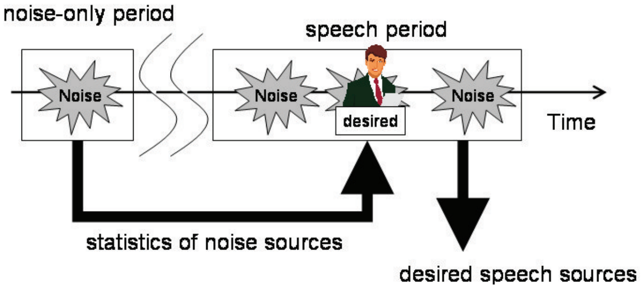

Fig. 1. Noise reduction focused in this paper

time-invariant impulse response of the i-th noise source at the m-th microphone, $\boldsymbol{s}_i(t)$ is $[\ s_i(t) \quad s_i(t-1) \quad \dots \quad s_i(t-L+1)\ ]^T$, $s_i(t)$ is the source signal of the i-th desired source, $\boldsymbol{n}_i(t)$ is defined similarly to $s_i(t)$, and L is the length of impulse responses. The goal of the noise reduction is approximation of the noise reduction signal $y(t)$ to the desired sources at the c-th microphone, $\sum_{i=0}^{N_s-1} \boldsymbol{h}_{i,c}^T \boldsymbol{s}_i(t)$.$c$ is set to be the first microphone in this paper.

2.3 Subband processing

The length of impulse responses are quite long under reverberant environments. The noise reduction at the time domain requires high computational cost. To reduce computational cost, the microphone input signal is converted into the subband domain by discrete Fourier transform (DFT) modulated filterbank [8]. The microphone input signal at the subband domain can be depicted as follows:

$$x_m(k,r) = \sum_{i=0}^{N_s-1} \boldsymbol{h}_{i,m}(k)^T \boldsymbol{s}_i(k,r) + \sum_{i=0}^{N_n-1} \boldsymbol{g}_{i,m}(k)^T \boldsymbol{n}_i(k,r) + v_m(k,r), \tag{2}$$

where k is the bin index at the subband domain, all variables with the suffix (k,r) is the r-th variables at the k-th bin, and the variables with the suffix (k) is time-invariant variables at the k-th bin. The downsampling ratio is set to be R, and $t = Rr$. To simplify the formulation of the noise reduction methods, the microphone input signal $x_m(k,r)$ is depicted with the matrix form as follows:

$$\boldsymbol{x}(k,r) = [\ \boldsymbol{x}_1(k,r)^T \quad \boldsymbol{x}_2(k,r)^T \quad \dots \quad \boldsymbol{x}_M(k,r)^T\ ]^T, \tag{3}$$

where M is the number of the microphones, $\boldsymbol{x}_m(k,r)$ is $[\ x_m(k,r) \quad x_m(k,r-1) \quad \dots \quad x_m(k,r-L_f+1)\ ]^T$, and L_f is the length of the noise reduction filter.

3. MULTICHANNEL BEAMFORMING TECHNIQUES

3.1 Multichannel Wiener Filter

A. Spriet pointed out the noise reduction performance of MWF based methods are extremely free from the steering vector error of the desired sources [6]. MWF reduces the noise source by the multichannel noise reduction filter $\boldsymbol{w}_{MWF}(k,r)$ as follows:

$$y(k,r) = \boldsymbol{w}_{MWF}(k,r)^H \boldsymbol{x}(k,r), \tag{4}$$

where H is the Hermite transpose, and $y(k,r)$ is the output signal. The cost function of the MWF, $L_{MWF}(\boldsymbol{w}(k,r))$, is defined as follows:

$$\begin{array}{rcl} L_{MWF}(\boldsymbol{w}(k,r)) & = & \mu \boldsymbol{w}(k,r)^H \mathbb{E}[\boldsymbol{x}(k,r)\boldsymbol{x}(k,r)^H]_n \boldsymbol{w}(k,r) \\ & + & \mathbb{E}[||x_c(k,r) - \boldsymbol{w}(k,r)^H \boldsymbol{x}(k,r)||^2]_s, \end{array} \tag{5}$$

where $\mathbb{E}[x]$ is the mathematical expectation of the variable x, $\mathbb{E}[x]_n$ is the expectation at the noise-only period, $\mathbb{E}[x]_s$ is the expectation at the desired-only period when there are only desired sources, the first term in Eq. 5 is the cost function for the residual noise power, the second term is the cost function for the desired-source distortion, and μ is the tradeoff parameter between the residual noise power and the desired-source distortion. $\boldsymbol{w}_{MWF}(k,r)$ which fulfills $\frac{\partial L_{MWF}(\boldsymbol{w}(k,r))}{\partial \boldsymbol{w}(k,r)} = 0$ is obtained as follows:

$$\boldsymbol{w}_{MWF}(k,r) = (\mu \boldsymbol{R}_n(k) + \boldsymbol{R}_s(k))^{-1} \boldsymbol{R}_{s,c}(k), \tag{6}$$

where $\boldsymbol{R}_n(k)$ is the second order statistics (SOS) of the noise sources and is $\mathbb{E}[\boldsymbol{x}(k,r)\boldsymbol{x}(k,r)^H]_n$, $\boldsymbol{R}_s(k)$ is the SOS of the desired sources and is $\mathbb{E}[\boldsymbol{x}(k,r)\boldsymbol{x}(k,r)^H]_s$, and $\boldsymbol{R}_{s,c}(k)$ is $\mathbb{E}[\boldsymbol{x}(k,r)x_c(k,r)^*]_s$ ($*$ is the operator for complex conjugate). Under the assumption that the noise-only period is given, the SOS of the desired sources $\boldsymbol{R}_s(k)$ is approximated as $\boldsymbol{R}_s(k) \approx \boldsymbol{R}_x(k) - \boldsymbol{R}_n(k)$, where $\boldsymbol{R}_x(k) = \mathbb{E}[\boldsymbol{x}(k,r)\boldsymbol{x}(k,r)^H]_{mix}$ is the SOS of the microphone input signal at the desired-source period when there are both the desired sources and the noise sources. $\boldsymbol{R}_{s,c}(k)$ is also approximated by $\boldsymbol{R}_{s,c}(k) \approx \boldsymbol{R}_{x,c}(k) - \boldsymbol{R}_{n,c}(k)$. In these operations, the statistics of the noise sources at the desired-source period are assumed to be the same as that at the noise-only periods. Actually, the expectation of each variable is replaced by the simple average. In this paper, the noise-only period are assumed to be selected to identify the noise sources to be reduced.

3.2 Multichannel spatial prediction based beamforming

Recently, some of the authors have proposed a two-stage beamforming on the subband domain, a multichannel spatial prediction based beamforming (MSP-BF) [7]. At the first stage, noise reduction filter which reduces the noise sources at each microphone is obtained without any preknowledge about the desired sources. The noise sources in each microphone is predicted by the noise sources in the other microphones. Residual noises in the output signal after the first stage are little, but the desired sources are distorted in the output signal. This distortion can be restored at the second stage by using a multichannel distortion-restoration filter. Noise reduction filter at the m-th microphone, $\boldsymbol{a}_m(k)$, can be obtained as follows:

$$\begin{array}{rcl} \boldsymbol{a}_m(k) & = & \underset{\boldsymbol{a}_m(k)}{\mathrm{argmin}} \mathbb{E}[||x_m(k,r) - \boldsymbol{a}_m(k)^H \boldsymbol{x}_m^e(k,r)||^2]_n \\ & = & \boldsymbol{V}_m(k)^{-1} \boldsymbol{C}_m(k) \end{array} \tag{7}$$

where $\boldsymbol{x}_m^e(k,r)$ is a multichannel input signal which excludes the m-th microphone input signal, the length of $\boldsymbol{x}_m^e(k,r)$ is set to be $(M-1)L_1$, L_1 is the length of the noise reduction filter, $\boldsymbol{V}_m(k)$ is $\mathbb{E}[\boldsymbol{x}_m^e(k,r)\boldsymbol{x}_m^e(k,r)^H]_n$, and $\boldsymbol{C}_m(k)$ is $\mathbb{E}[\boldsymbol{x}_m^e(k,r)x_m(k,r)^*]_n$. $\boldsymbol{a}_m(k)$ is a prediction filter of the noise source in the m-th microphone. Subtracting the predicted noise component $\boldsymbol{a}_m(k)^H \boldsymbol{x}_m^e(k,r)$ from $x_m(k,r)$, a noiseless signal $y_{f,m}(k,r) = x_m(k,r) - \boldsymbol{a}_m(k)^H \boldsymbol{x}_m^e(k,r)$ can be extracted.

The residual noise after the first filtering is minimized, but the desired sources after the first filtering are distorted because there are no constraint to the desired sources. This distortion is restored at the second stage. From the multichannel distorted desired sources, the less distorted desired sources can be extracted by the multichannel restoration filter $\boldsymbol{w}_{dist}(k)$ as follow:

$$y_c(k,r) = \boldsymbol{w}_{dist}(k)^H \boldsymbol{y}_f(k,r), \tag{8}$$

where $y_c(k,r)$ is the output signal after the distortion-restoration filtering, $\boldsymbol{y}_f(k,r) = [\ \boldsymbol{y}_{f,1}(k,r)^H \quad \boldsymbol{y}_{f,2}(k,r)^H \quad \dots \quad \boldsymbol{y}_{f,M}(k,r)^H\ ]^H$, $\boldsymbol{y}_{f,m}(k,r) = [\ y_{f,m}(k,r), \quad y_{f,m}(k,r-1) \quad \dots \quad y_{f,m}(k,r-L_2+1)\ ]^H$, L_2 is the length of the distortion-restoration filter for each microphone.

$\boldsymbol{w}_{dist}(k)$ is obtained as follows:

$$\begin{aligned}\boldsymbol{w}_{dist}(k) &= \underset{\boldsymbol{w}_{dist}(k)}{\operatorname{argmin}} \mathbb{E}[||x_c(k,r) - \boldsymbol{w}_{dist}(k)^H \boldsymbol{y}_f(k,r)||^2]_s \\ &+ \mu \mathbb{E}[||\boldsymbol{w}_{dist}(k)^H \boldsymbol{y}_f(k,r)||^2]_n \\ &= (\boldsymbol{R}_y(k) + \mu \boldsymbol{R}_{y,n}(k))^{-1} \boldsymbol{R}_{y,c}(k), \end{aligned} \tag{9}$$

where $\boldsymbol{R}_y(k) = \mathbb{E}[\boldsymbol{y}_f(k,r)\boldsymbol{y}_f(k,r)^H]_s$, $\boldsymbol{R}_{y,n}(k) = \mathbb{E}[\boldsymbol{y}_f(k,r)\boldsymbol{y}_f(k,r)^H]_n$, $\boldsymbol{R}_{y,c}(k) = \mathbb{E}[\boldsymbol{y}_f(k,r)x_c(k,r)^*]_s$, and $\mathbb{E}[]_s$ is replaced by $\mathbb{E}[]_{mix} - \mathbb{E}[]_n$. The quality of the output signal after the distortion-restoration filtering depends on the estimation accuracy of the statistics of the desired sources, $\boldsymbol{R}_y(k)$ and $\boldsymbol{R}_{y,c}(k)$.

4. THEORECTICAL ANALYSIS OF MSP-BF

Theoretically, MSP-BF is closely related to MWF. In this section, a theoretical relationship between MSP-BF and MWF is shown.

4.1 Analysis of the first noise reduction filter in MSP-BF

The multichannel input signal (Eq. 2) can be transformed into the matrix form as follows:

$$\boldsymbol{x}_m(k,r) = \boldsymbol{H}_m(k)\boldsymbol{s}(k,r) + \boldsymbol{G}_m(k)\boldsymbol{n}(k,r) + \boldsymbol{v}_m(k,r), \tag{10}$$

where $\boldsymbol{x}_m(k,r)$ is originally defined as the L_f vector, but in this subsection, $\boldsymbol{x}_m(k,r)$ is re-defined as the L_1 vector. $\boldsymbol{s}(k,r) = [\ \boldsymbol{s}_0(k,r)^T \ \ \ldots \ \ \boldsymbol{s}_{N_s-1}(k,r)^T\ ]^T$, $\boldsymbol{n}(k,r)$ is defined similarly to $\boldsymbol{s}(k,r)$, $\boldsymbol{v}_m(k,r)$ is a L_1 dimensional vector from $v_m(k,r)$, $\boldsymbol{H}_m(k) = [\ \boldsymbol{H}_{0,m}(k) \ \ \boldsymbol{H}_{1,m}(k) \ \ \ldots \ \ \boldsymbol{H}_{N_s-1,m}(k)\ ]$, $\boldsymbol{H}_{i,m}(k)$ is a $L_1 \times (L_1 + L - 1)$ Sylvester matrix made by $\boldsymbol{h}_{i,m}(k)$, and $\boldsymbol{G}_m(k)$ is defined similarly to $\boldsymbol{H}_m(k)$. In the noise-only period, $\boldsymbol{x}_m(k,r)$ can be approximated as follows:

$$\boldsymbol{x}_m(k,r) = \boldsymbol{G}_m(k)\boldsymbol{n}(k,r), \tag{11}$$

where the background noise is neglected. Under the condition that $(M-1)L_1$ is larger than $N_n(L+L_1-1)$ and the impulse response of each noise source does not share a common zero [10], the ideal prediction filter at the first stage, $\boldsymbol{a}_{ideal,m}(k)$, can be depicted as follows:

$$\begin{aligned}\boldsymbol{a}_{ideal,m}(k) &= \underset{\boldsymbol{a}_m(k)}{\operatorname{argmin}} \mathbb{E}[x_m(k,r) - \boldsymbol{a}_m(k)^H \boldsymbol{x}^e_m(k,r)]_n \\ &= \underset{\boldsymbol{a}_m(k)}{\operatorname{argmin}} \mathbb{E}[(\boldsymbol{G}_m(k)[1] - \boldsymbol{a}_m(k)^H \boldsymbol{G}^e_m(k))\boldsymbol{n}(k,r)]_n \\ &= (\boldsymbol{G}^e_m(k)^+)^H \boldsymbol{G}_m(k)[1]^H, \end{aligned} \tag{12}$$

where $\boldsymbol{G}_m(k)[1]$ is the first row of $\boldsymbol{G}_m$, and $\boldsymbol{G}^e_m(k)^+$ is the generalized inverse matrix of $\boldsymbol{G}^e_m(k)$. $\boldsymbol{G}^e_m(k)$ is defined as follows:

$$\boldsymbol{G}^e_m(k) = [\ \boldsymbol{G}_1(k)^H \ldots \ \ \boldsymbol{G}_{m-1}(k)^H \ \ \boldsymbol{G}_{m+1}(k)^H \ldots \ \ \boldsymbol{G}_M(k)^H\ ]^H. \tag{13}$$

The estimated prediction filter by Eq. 7, $\boldsymbol{a}_m(k)$, can be expanded as follows:

$$\boldsymbol{a}_m(k) = (\boldsymbol{G}^e_m(k)\boldsymbol{R}_{nn}(k)\boldsymbol{G}^e_m(k)^H)^{-1}\boldsymbol{G}^e_m(k)\boldsymbol{R}_{nn}(k)\boldsymbol{G}_m(k)[1]^H, \tag{14}$$

where $\boldsymbol{R}_{nn}(k) = \mathbb{E}[\boldsymbol{n}(k,r)\boldsymbol{n}(k,r)^H]$. Assuming that each noise source is a white Gaussian signal and each noise source has a same power, $\boldsymbol{R}_{nn}(k)$ can be approximated by $\sigma(k)\boldsymbol{E}$ ($\boldsymbol{E}$ is the identity matrix). In this case,

$$\boldsymbol{a}_m(k) = (\boldsymbol{G}^e_m(k)\boldsymbol{G}^e_m(k)^H)^{-1}\boldsymbol{G}^e_m(k)\boldsymbol{G}_m(k)[1]^H. \tag{15}$$

When $(M-1)L_1$ equals to $N_n(L+L_1-1)$ and the impulse response of each noise source does not share a common zero, $\boldsymbol{a}_m(k)$ equals to $\boldsymbol{a}_{ideal,m}(k)$. When each source is not a white Gaussian signal, $\boldsymbol{a}_m(k)$ is different from $\boldsymbol{a}_{ideal,m}(k)$. To simplify the following discussion, the output signal after the first noise reduction filtering is transformed as follows:

$$y_{f,m}(k,r) = \tilde{\boldsymbol{a}}_m(k)^H \tilde{\boldsymbol{x}}(k,r), \tag{16}$$

where $\tilde{\boldsymbol{x}}(k,r)$ is a ML_1-dimensional multichannel microphone input signal, and $\tilde{\boldsymbol{a}}_m(k)$ is defined as follows:

$$\begin{aligned}\tilde{\boldsymbol{a}}_m(k) \ &= [\ -a_m(k)[1] \ \ -a_m(k)[2] \ \ \ldots \ \ -a_m(k)[(m-1)L_1] \\ &\quad \boldsymbol{I}^T \ \ -a_m(k)[(m-1)L_1+1] \\ &\quad \ldots \ \ -a_m(k)[(M-1)L_1]\ ]^T, \end{aligned} \tag{17}$$

wherer $\boldsymbol{I}$ is a L_1 dimensional vector defined as $[\ 1 \ \ 0 \ \ \ldots \ \ 0\ ]^H$. Furthermore, the L_2 dimensional vector, $\boldsymbol{y}_{f,m}(k,r)^H$, is denoted as follows:

$$\boldsymbol{y}_{f,m}(k,r)^H = \boldsymbol{A}_m(k)\boldsymbol{x}(k,r), \tag{18}$$

where $\boldsymbol{x}(k,r)$ is a L_f dimensional vector, L_f is set to be L_1+L_2-1, and $\boldsymbol{A}_m(k)$ is a Sylvester matrix of $(\tilde{\boldsymbol{a}}_m(k)^H)^T$.

4.2 Analysis of the distortion-restoration filtering at the second stage

$\boldsymbol{y}_f(k,r)$, can be described in a matrix form as follows:

$$\boldsymbol{y}_f(k,r) = \boldsymbol{A}(k)\boldsymbol{x}(k,r), \tag{19}$$

where $\boldsymbol{A}(k)$ is $[\ \boldsymbol{A}_1(k)^T \ \ \ldots \ \ \boldsymbol{A}_M(k)^T\ ]^T$. Furthermore, the noise sources and the desired sources in $\boldsymbol{x}(k,r)$ can be represented as follows:

$$\boldsymbol{x}(k,r) = \mathscr{H}(k)\boldsymbol{s}(k,r) + \mathscr{G}(k)\boldsymbol{n}(k,r), \tag{20}$$

where $\mathscr{H}(k) = [\ \boldsymbol{H}_1(k)^T \ \ \ldots \ \ \boldsymbol{H}_M(k)^T\ ]^T$, and $\mathscr{G}(k) = [\ \boldsymbol{G}_1(k)^T \ \ \ldots \ \ \boldsymbol{G}_M(k)^T\ ]^T$ ($\boldsymbol{H}_m(k)$ and $\boldsymbol{G}_m(k)^T$ are $L_f \times N_s(L_f+L-1)$ matrices). The distortion-restoration filter $\boldsymbol{w}_{dist}(k)$ can be expanded as follows:

$$\begin{aligned}\boldsymbol{w}_{dist}(k) &= (\boldsymbol{B}(k)\boldsymbol{R}_{ss}(k)\boldsymbol{B}(k)^H + \mu\boldsymbol{D}(k)\boldsymbol{R}_{nn}(k)\boldsymbol{D}(k)^H)^{-1} \\ &\quad \boldsymbol{B}(k)\boldsymbol{R}_{ss}(k)\boldsymbol{H}_c(k)[1]^H, \end{aligned} \tag{21}$$

where $\boldsymbol{R}_{ss}(k) = \mathbb{E}[\boldsymbol{s}(k,r)\boldsymbol{s}(k,r)^H]_s$, $\boldsymbol{B}(k) = \boldsymbol{A}(k)\mathscr{H}(k)$, $\boldsymbol{D}(k) = \boldsymbol{A}(k)\mathscr{G}(k)$. $\boldsymbol{D}(k)$ is regarded as a blocking matrix for the noise sources. Therefore, $\boldsymbol{D}(k)\boldsymbol{R}_{nn}(k)\boldsymbol{D}(k)^H$ is a residual noise term after the first filtering. When each desired source is a white Gaussian signal and has a same power and the blocking of the noise sources in the first stage is performed completely, $\boldsymbol{w}_{dist}(k)$ can be transformed as follows:

$$\boldsymbol{w}_{dist}(k) = (\boldsymbol{B}(k)\boldsymbol{B}(k)^H)^{-1}\boldsymbol{B}(k)\boldsymbol{H}_c(k)[1]^H. \tag{22}$$

$\boldsymbol{B}(k)$ is a $ML_2 \times N_s(L_f+L-1)$ matrix and $L_f = L_1+L_2-1$. Under the condition that L_2 is larger than $\frac{N_s(L_1+L-2)}{M-N_s}$ and the impulse response between each source position and the output signal after the first stage at each microphone has no common zero, the ideal distortion-restoration filter, $\boldsymbol{w}_{dist,ideal}(k)$, is $(\boldsymbol{B}(k)^+)^H\boldsymbol{H}_c(k)[1]^H$. Therefore, $\boldsymbol{w}_{dist}(k)$ is also an approximation of the ideal distortion-restoration filter. From Eq. 8, the output signal of the MSP-BF can be represented as follows:

$$\begin{aligned} y_c(k,r) &= \boldsymbol{H}_c(k)[1]\boldsymbol{R}_{ss}(k)\boldsymbol{B}(k)^H(\boldsymbol{B}(k)\boldsymbol{R}_{ss}(k)\boldsymbol{B}(k)^H + \\ &\quad \mu\boldsymbol{D}(k)\boldsymbol{R}_{nn}(k)\boldsymbol{D}(k)^H)^{-1}\boldsymbol{A}(k)\boldsymbol{x}(k,r). \end{aligned} \tag{23}$$

4.3 Relationship between MSP-BF and MWF

The output signal of MWF can be represented as follows:

$$y_{mwf}(k,r) = \boldsymbol{H}_c(k)[1]\boldsymbol{R}_{ss}(k)\mathscr{H}(k)^H(\mathscr{H}(k)\boldsymbol{R}_{ss}(k)\mathscr{H}(k)^H + \mu\mathscr{G}(k)\boldsymbol{R}_{nn}(k)\mathscr{G}(k)^H)^{-1}\boldsymbol{x}(k,r). \quad (24)$$

From comparison Eq. 24 with Eq. 23, the difference between MSP-BF and MWF is existence of the prefiltering structure. When the noise reduction filter $\boldsymbol{A}(k)$ is invertible, MSP-BF equals to MWF. However, the noise reduction filter $\boldsymbol{A}(k)$ is a $ML_2 \times M(L_2+L_1-1)$ matrix. Therefore, when $L_1 \geq 2$, $\boldsymbol{A}(k)$ is not invertible and the output signal of MSP-BF does not equal to that of MWF.

5. AN IMPROVED TWO-STAGE NOISE REDUCTION STRUCTURE

5.1 An alternative noise reduction structure at the first stage

The spatial prediction based noise reduction can be regarded as a null beamforming, the noise reduction filter is orthogonal to the subspace spanned by the noise sources. However, the spatial prediction does not deal with the subspace of the noise sources directly. Motivated by the subspace approaches, an alternative noise reduction structure based on the noise subspace reduction (NSR-BF) is shown. The m-th noise reduction filter, $\boldsymbol{d}_m(k)$, is set to be the eigen vector whose eigen value is the m-th smallest eigen value of the matrix $\boldsymbol{R}_n(k)$. The l_2 norm of each eigen vector is set to be 1. $\boldsymbol{R}_n(k)$ can be transformed as $\boldsymbol{R}_n(k) = \mathscr{G}(k)\boldsymbol{R}_{nn}(k)\mathscr{G}(k)^H$. When each noise source is a white Gaussian signal and has a same power, $\boldsymbol{R}_n(k) = \mathscr{G}(k)\mathscr{G}(k)^H$. When $ML_1 > N_n(L+L_1-1)$, the rank of the $ML_1 \times ML_1$ matrix ,$\mathscr{G}(k)\mathscr{G}(k)^H$, is $N_n(L+L_1-1)$. Therefore, there are some subspaces whose eigenvalues are zero, and $\boldsymbol{d}_m(k)^H\mathscr{G}(k)\mathscr{G}(k)^H\boldsymbol{d}_m(k) = 0$. $\boldsymbol{d}_m(k)^H$ is orthogonal to $\mathscr{G}(k)$, and the noise sources are reduced in the output signal after the filtering by $\boldsymbol{d}_m(k)^H$. The cost function of this process can be interpreted as follows:

$$\boldsymbol{d}_m(k) = \underset{\boldsymbol{d}_m(k)}{\overset{(m)}{\mathrm{argmax}}}\, \frac{\boldsymbol{d}_m(k)^H\tilde{\boldsymbol{R}}_s(k)\boldsymbol{d}_m(k)}{\boldsymbol{d}_m(k)^H\tilde{\boldsymbol{R}}_n(k)\boldsymbol{d}_m(k)},$$

$$\leftrightarrow \quad \tilde{\boldsymbol{R}}_s(k)\boldsymbol{d}_m(k) = \mathbf{SNR}_m\boldsymbol{R}_n\boldsymbol{d}_m(k)$$

$$\leftrightarrow \quad \frac{1}{\mathbf{SNR}_m}\tilde{\boldsymbol{R}}_s(k)\boldsymbol{d}_m(k) = \boldsymbol{R}_n\boldsymbol{d}_m(k) \quad (25)$$

where $\mathbf{SNR}_m$ is the m-th biggest ratio of the desired sources to the noise sources at the output signal of $\boldsymbol{d}_m(k)$, $\tilde{\boldsymbol{R}}_s(k)$ is a tentative covariance matrix of the desired sources and is set to be the identity matrix, argmax$^{(m)}$ returns the argument which has the m-th biggest value. Therefore, the m-th smallest eigen value of $\boldsymbol{R}_n$ is corresponding to the m-th biggest SNR value. When the noise sources are not white signals, the frequency characteristic of $\boldsymbol{d}_m(k)$ tends to reduce the dominant frequencies of the noise sources in the spectral domain (not in the spatial domain). This effect is prominent when the assumed desired sources do not contain these frequencies. However, because each desired source in $\tilde{\boldsymbol{R}}_s(k)$ is assumed to be a white signal, $\tilde{\boldsymbol{R}}_s(k)$ has all frequencies. Therefore, extreme reduction of particular frequencies in the frequency characteristic of $\boldsymbol{d}_m(k)$ is expected to be suppressed. $\boldsymbol{d}_m(k)$ can be used by an alternative of the noise reduction filter $\tilde{\boldsymbol{a}}_m(k)$ at the first stage. When the actual covariance matrix of the desired sources are far from $\tilde{\boldsymbol{R}}_s(k)$, the output signal after the first stage is distorted, but this distortion can be restored at the second stage.

5.2 An energy based VAD after the first filtering

The detection of the desired-source period is important for the estimation of the distortion-restoration filter. When the noise sources are nonstationary, a simple energy based VAD is not adequate. ICA-based VAD [9] requires direction of arrival of the desired sources. To detect the desired-source period without any preknowledge about the desired sources, an energy based VAD by utilizing the output signal after the first filtering is shown in this subsection. The advantageous point of the two stage filtering structure is that SNR (signal to noise ratio) of the output signal after the first filtering is expected to be higher than the unprocessed microphone input signal. Therefore, even when the noise sources are nonstationary sources, the desired-source period can be detected by the simple energy based VAD from the output signal after the first filtering.

The output signal after the first filtering, noise sources are highly reduced and the desired sources are distorted. The noise reduction filter at the first stage makes the spatial nulls toward the transfer function of the noise sources. When the transfer function of the desired sources are different from that of the noise sources, the desired sources are not crucially reduced. Therefore, compared with the noise sources, the remaining power of the desired sources are big. A sample of a waveform after the first filtering by NSR-BF is shown in Fig. 2. It is shown that the desired sources can be extracted. By utilizing this phenomenon, the desired-source period can be extracted via energy based VAD. The proposed algorithm is shown as follows:

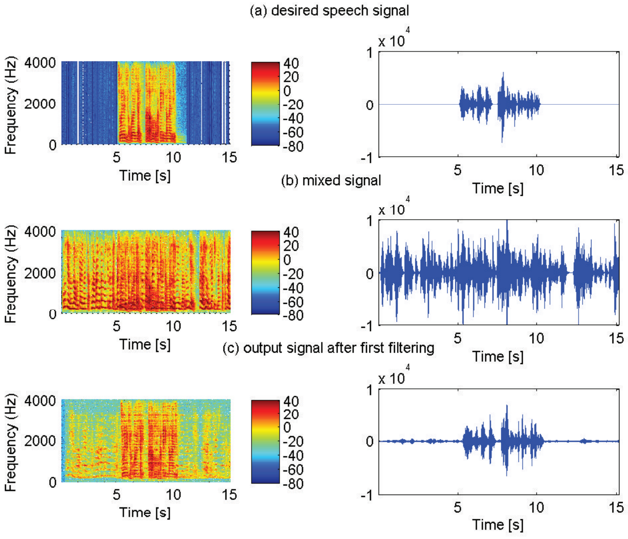

Fig. 2. A sample of a waveform after the first filtering by NSR-BF

1. Log-power of the output signal after the first filtering is divided into two clusters.
2. The bigger cluster is regarded as the desired-source period, and the smaller cluster is regarded as the noise-only period.
3. Each time period is divided into the noise-only period or the desired-source period.
4. In the noise-only period, the SOS of the noise sources, $\boldsymbol{R}_n(k)$, is replaced by $\boldsymbol{R}_{n,new}(k)$.$\boldsymbol{R}_{n,new}(k)$ is the SOS of the detected noise-only period. In the desired-source period, the SOS of the noisy input signal, $\boldsymbol{R}_x(k)$ is updated.

5.3 Flow of proposed method

Flow of the proposed method can be summarized as follows:

1. The first noise reduction filter $\boldsymbol{d}_m(k)$ is estimated by Eq. 25 from the pre-given noise-only period. The output signal after the first noise reduction filtering can be calculated.
2. From the output signal after the first noise reduction filtering, the desired-source period is detected by the energy based VAD shown in sec. 5.2;
3. The distortion-restoration filter is updated by Eq. 9 from the detected-source period and the detected noise-only period.

4. The distortion of the output signal after the first filtering is restored at the whole period.

6. EXPERIMENT

6.1 Condition

The proposed method was evaluated at a reverberant environment. The reverberation time was about 300 ms. Impulse responses were measured at the environment, and the evaluation data was made by the convolution of dry sources with the measured impulse responses. The sampling rate was set to be 8 kHz. The downsampling ratio R was set to be 56 pt (point). The length of the low-pass filter for the analysis/synthesis filterbank was 3584 pt. The number of the microphones was 3. The equilateral triangle microphone array (4 cm on a side) was used. The number of the desired sources was set to be 1 or 2. The number of the noise sources was set to be 2. One noise source was a speech source, and another noise source was a pink noise source. The proposed two stage noise reduction techniques were compared with MWF. The length of the noise reduction filter of MWF, L_f, was set to be 12. In the proposed methods, the length of the first noise reduction filter, L_1 was set to be 8, and L_2 was set to be 5. Therefore, $L_f = L_1 + L_2 - 1$ is 12, and L_f is equivalent to that of MWF. The evaluation measure are MFCC distance, and SNR improvement. MFCC distance is the evaluation measure for the desired-source distortion, and SNR improvement is a performance measure the noise reduction. MFCC distance is $\mathrm{MFCC}(\boldsymbol{k}_{out}, \boldsymbol{k}_{desired}) = \sum_{i=0}^{12} ||k_{out}(i) - k_{desired}(i)||^2$. k_{out} is the MFCC coeficients of the output signal after the noise reduction, and $k_{desired}$ is that of the desired source signal in the c-th microphone.

The dimension of the MFCC coeficients is set to be 13. SNR improvement is $\mathbf{SNR}_{imp} = \mathbf{SNR}_{out} - \mathbf{SNR}_{in}$. $\mathbf{SNR}_{in}$ is the ratio between the desired sources and the noise sources in the c-thmicrophone input signal, and $\mathbf{SNR}_{out}$ is the ratio between the desired sources and the noise sources in the output signal of the noise reduction system. μ is changed from 1.0 to 10.0 at 1.0 intervals, and is changed from 10.0 to 100.0 at 10.0 intervals.

6.2 Results

When the number of the desired sources is 1, the experimental results are shown in Fig. 3. The different point from "NSR-BF" and "MSP-BF" is only the first filtering structure. "NSR-BF+VAD" is the noise reduction which uses the first filtering by NSR and detection of the desired-source period from the output signal after the first filtering. In this result, the parameter μ is regarded as an intermediate variable. SNR$_{in}$ is set to be 0 dB. The noise reduction method reduces noise at the expense of the distortion of the desired sources. The rapid increase of $\mathbf{SNR}_{imp}$ is desirable to the increase of MFCC distance. In this viewpoint, MWF is superior to MSP-BF, but NSR is superior to MWF. This means the noise reduction performance of the first filtering of NSR is superior to MSP-BF. Furthermore, NSR-BF+VAD is superior to NSR. The proposed detection algorithm of the desired-source period is shown to be effective. In Fig. 4, the experimental result when there are 2 desired sources at SNR 0 dB is shown. NSR-BF+VAD also achieves the best performance.

7. CONCLUSION

In this paper, a two-stage noise reduction method which is composed of the noise reduction stage and the distortion-restoration stage is discussed. Theoretical analysis of the previously proposed MSP-BF is shown. From the analysis, an alternative structure for MSP-BF, namely NSR-BF, is proposed, and the energy based voice activity detection to the output signal after the noise reduction stage is proposed. The experimental results under a reverberant environment show that NSR-BF is superior to the conventional methods.

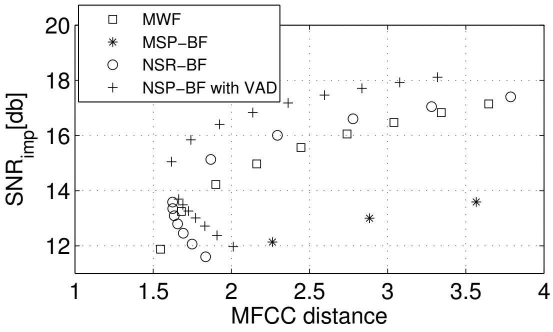

Fig. 3. Experimental result under the condition that the number of the desired sources is set to be 1, the number of the noise sources are set to be 2, and SNR$_{in}$ is set to be 0 dB. MFCC distance of the microphone input signal is about 5.2.

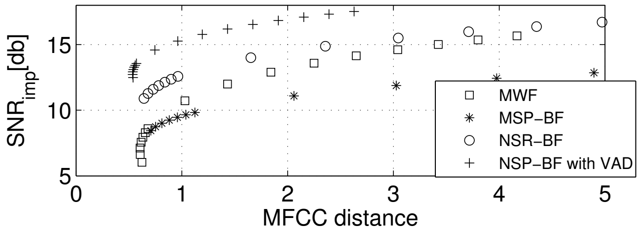

Fig. 4. Experimental result under the condition that the number of the desired sources and the number of the noise sources are set to be 2, and SNR$_{in}$ is set to be 0 dB. MFCC distance of the microphone input signal is about 1.8.

REFERENCES

[1] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," In *Proc. IEEE*, vol.60, no.8, pp.926-935, Aug. 1972.

[2] L. J. Griffith and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Anntenas Propagation*, vol.30, i.1, pp.27-34, Jan. 1982.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and non-stationarity with applications to speech," *IEEE Trans. SP*, vol. 49, no. 8, pp. 1614–1626, 2001.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, "Independent component analysis," John Wiley & Sons, 2001.

[5] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. SP,* vol. 47, no. 10, pp. 2677-2684, Oct. 1999.

[6] 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.

[7] M. Togami, Y. Kawaguchi, and Y. Obuchi, "Subband nonstationary noise reduction based on multichannel spatial prediction under reverberant environments," *Proc. ICASSP 2009*, pp. 133–136, 2009.

[8] R. Crochiere and L. Rabiner, Multirate Digital Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[9] M. Togami, Y. Kawaguchi, H. Kokubo, and Y. Obuchi, "Subband beamformer combined with time-frequency ICA for extraction of target source under reverberant environments," *Proc. EUSIPCO 2009*, pp. 150–154, 2009.

[10] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, vol. 30, no. 2, pp. 145-152, Feb. 1988.