Toward Musical-Noise-Free Blind Speech Extraction: Concept and Its Applications

Ryoichi Miyazaki*, Hiroshi Saruwatari*, Satoshi Nakamura*, Kiyohiro Shikano*,
Kazunobu Kondo†, Jonathan Blanchette‡, and Martin Bouchard‡
* Graduate School of Information Science, Nara Institute of Science and Technology, Japan
E-mail: ryoichi-m@is.naist.jp
† Corporate Research & Development Center, Yamaha Corporation, Japan
‡ School of Information Technology and Engineering, University of Ottawa, Canada

***Abstract*—In this paper, we review a blind musical-noise-free speech extraction method using a microphone array that can be applied to nonstationary noise. In our previous study, it was found that optimized iterative spectral subtraction (SS) results in speech enhancement with almost no musical noise generation, but this method is valid only for stationary noise. The proposed method consists of iterative blind dynamic noise estimation by, e.g., ICA or multichannel Wiener filtering, and musical-noise-free speech extraction by modified iterative SS, where multiple iterative SS is applied to each channel while maintaining the multichannel property reused for the dynamic noise estimators. Also, related to the proposed method, we discuss the justification of applying ICA to such signals nonlinearly distorted by SS. From objective and subjective evaluations simulating real-world hands-free speech communication system, we reveal that the proposed method outperforms the conventional methods.**

I. Introduction

In the past few decades, many applications of speech communication systems, such as hearing aids and mobile phones, have been investigated. It is, however, well known that these systems always suffer from the deterioration of speech quality under adverse noise conditions, and thus noise reduction is a problem requiring urgent attention. Spectral subtraction is a commonly used noise reduction method that has high noise reduction performance with low computational complexity [1], [2], [3], [4], [5]. However, in this method, artificial distortion, so-called musical noise, arises owing to nonlinear signal processing, leading to a serious deterioration of sound quality.

To achieve high-quality noise reduction with low musical noise, an iterative SS method has been proposed [6], [7], [8]. This method is performed through signal processing in which weak SS processes are iteratively applied to the input signal. Also, some of the authors have reported the very interesting phenomenon that this method with appropriate parameters gives equilibrium behavior in the growth of higher-order statistics with increasing number of iterations [9]. This means that almost no musical noise is generated even with high noise reduction, which is one of the most desirable properties of single-channel nonlinear noise reduction methods. Following this finding, the authors have derived the optimal parameters satisfying the no musical noise generation condition by analysis based on higher-order statistics. We have defined this method as musical-noise-free speech enhancement, where no musical noise is generated even for a high SNR in iterative SS [10].

In conventional iterative SS, however, it is assumed that the input noise signal is stationary, meaning that we can estimate the expectation of noise power spectral density from a time-frequency period of a signal that contains only noise. In contrast, under real-world acoustical environments, e.g., a nonstationary noise field, although it is necessary to dynamically estimate noise, this is very difficult. Therefore in this paper, first, we propose a new iterative signal extraction method using a microphone array that can be applied to nonstationary noise [11], [12]. Our proposed method consists of iterative blind dynamic noise estimation by independent component analysis (ICA) [13], [14] and musical-noise-free speech extraction by modified iterative SS, where multiple iterative SS is applied to each channel while maintaining the multi-channel property reused for ICA.

Secondly, related to the proposed method, we discuss the justification of applying ICA to such signals nonlinearly distorted by SS. We theoretically clarify that the degradation in ICA-based noise estimation obeys an amplitude variation in room transfer functions between the target user and microphones. Next, to reduce speech distortion, we introduce a channel selection strategy into ICA, where we automatically choose less varied inputs to maintain a high accuracy of the noise estimation. Furthermore, we introduce a time-variant noise PSD estimator [15] instead of ICA for improvement of the noise estimation accuracy. From objective and subjective evaluations, we reveal that the proposed method outperforms the conventional method.

The rest of the paper is organized as follows. In Sect. II, we describe related works on spectral subtraction and the musical noise metric. In Sect. III, new musical-noise-free blind speech extraction method is proposed. In Sect. IV, an improvement scheme for poor noise estimation is presented. In Sect. V, objective and subjective evaluations are described. Following a discussion on the results of the experiments, we present our conclusions in Sect. VI.

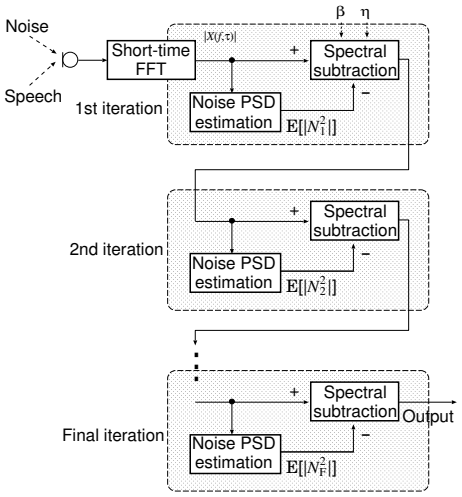

Fig. 1. Block diagram of iterative SS.

II. Related works

A. Conventional non-iterative SS [2]

We apply short-time Fourier analysis to the observed signal, which is a mixture of target speech and noise, to obtain the time-frequency signal. We formulate conventional *non-iterative SS* [2] in the time-frequency domain as follows:

$$y(f,\tau) = \begin{cases} \sqrt{|x(f,\tau)|^2 - \beta \mathrm{E}[|N|^2]}\exp(j\arg(x(f,\tau))) \\ \quad (\text{if} \quad |x(f,\tau)|^2 > \beta \mathrm{E}[|N|^2]), \\ \eta x(f,\tau) \quad (\text{otherwise}), \end{cases} \tag{1}$$

where $y(f,\tau)$ is the enhanced target speech signal, $x(f,\tau)$ is the observed signal, f denotes the frequency subband, τ is the frame index, β is the oversubtraction parameter, and η is the flooring parameter. Here, $\mathrm{E}[|N|^2]$ is the expectation of the random variable $|N|^2$ corresponding to the noise power spectra. In practice, we can approximate $\mathrm{E}[|N|^2]$ by averaging the observed noise power spectra $|n(f,\tau)|^2$ in the first K-sample frames, where we assume the absence of speech in this period and noise stationarity. However, this often requires high-accuracy voice activity detection.

B. Iterative SS [6], [7], [8]

In an attempt to achieve high-quality noise reduction with low musical noise, an improved method based on iterative SS was proposed in previous studies [6], [7], [8]. This method is performed through signal processing, in which the following *weak* SS processes are recursively applied to the noise signal (see Fig. 1). (I) The average power spectrum of the input noise is estimated. (II) The estimated noise prototype is then subtracted from the input with the parameters specifically set for weak subtraction, e.g., a large flooring parameter η and a small subtraction parameter β. (III) We then return to step (I) and substitute the resultant output (partially noise reduced signal) for the input signal.

C. Modeling of input signal

In this paper, we assume that the input signal x in the power spectral domain is modeled using the gamma distribution as

$$P(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\theta^{\alpha}}\exp(-x/\theta), \tag{2}$$

where $x \geq 0,\ \alpha > 0$, and $\theta > 0$. Here, α is the shape parameter, θ is the scale parameter, and $\Gamma(\alpha)$ is the gamma function, defined as $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\exp(-t)dt$.

D. Mathematical metric of musical noise generation via higher-order statistics for non-iterative SS [16]

In this study, we apply the *kurtosis ratio* to a *noise-only time-frequency period* of the subject signal for the assessment of musical noise [16]. This measure is defined as

$$\text{kurtosis ratio} = \mathrm{kurt}_{\mathrm{proc}}/\mathrm{kurt}_{\mathrm{org}}, \tag{3}$$

where $\mathrm{kurt}_{\mathrm{proc}}$ is the kurtosis of the processed signal and $\mathrm{kurt}_{\mathrm{org}}$ is the kurtosis of the observed signal. Kurtosis is defined as

$$\mathrm{kurt} = \mu_4/\mu_2^2, \tag{4}$$

where μ_m is the mth-order moment, given by

$$\mu_m = \int_0^{\infty} x^m P(x)dx, \tag{5}$$

and $P(x)$ is the probability density function (p.d.f.) of a power-spectral-domain signal x. A kurtosis ratio of unity corresponds to no musical noise. This measure increases as the amount of generated musical noise increases.

The mth-order moment after SS, μ_m, is given by [9]

$$\mu_m = \theta_{\mathrm{n}}^m \mathcal{M}(\alpha_{\mathrm{n}}, \beta, \eta, m), \tag{6}$$

where θ_{n} is the noise scale parameter, α_{n} is the noise shape parameter, and

$$\mathcal{M}(\alpha_{\mathrm{n}}, \beta, \eta, m) = \mathcal{S}(\alpha_{\mathrm{n}}, \beta, \eta) + \eta^{2m}\mathcal{F}(\alpha_{\mathrm{n}}, \beta, \eta), \tag{7}$$

$$\mathcal{S}(\alpha_{\mathrm{n}}, \beta, m) = \sum_{l=0}^{m}(-\beta\alpha_{\mathrm{n}})^l \frac{\Gamma(m+1)\Gamma(\alpha_{\mathrm{n}}+m-l, \beta\alpha_{\mathrm{n}})}{\Gamma(\alpha_{\mathrm{n}})\Gamma(l+1)\Gamma(m-l+1)}, \tag{8}$$

$$\mathcal{F}(\alpha_{\mathrm{n}}, \beta, m) = \frac{\gamma(\alpha_{\mathrm{n}}+m, \beta\alpha_{\mathrm{n}})}{\Gamma(\alpha_{\mathrm{n}})}. \tag{9}$$

$\Gamma(b,a)$ and $\gamma(b,a)$ are the upper and lower incomplete gamma functions defined as $\Gamma(b,a) = \int_b^{\infty} t^{a-1}\exp(-t)dt$ and $\gamma(b,a) = \int_0^b t^{a-1}\exp(-t)dt$, respectively. From (4), (6), and (7), the kurtosis after SS can be expressed as

$$\mathrm{kurt} = \frac{\mathcal{M}(\alpha_{\mathrm{n}}, \beta, \eta, 4)}{\mathcal{M}^2(\alpha_{\mathrm{n}}, \beta, \eta, 2)}. \tag{10}$$

Using (3) and (10), we also express the kurtosis ratio as

$$\text{kurtosis ratio} = \frac{\mathcal{M}(\alpha_{\mathrm{n}}, \beta, \eta, 4)/\mathcal{M}^2(\alpha_{\mathrm{n}}, \beta, \eta, 2)}{\mathcal{M}(\alpha_{\mathrm{n}}, 0, 0, 4)/\mathcal{M}^2(\alpha_{\mathrm{n}}, 0, 0, 2)}. \tag{11}$$

Also, as a measure of noise reduction performance, the noise reduction rate (NRR), the output signal-to-noise ratio (SNR)

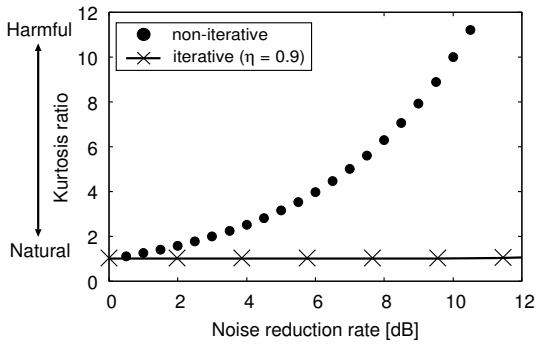

Fig. 2. Relation between NRR and kurtosis ratio obtained from theoretical analysis for Gaussian noise case.

minus the input SNR in dB, can be given in terms of a 1st-order moment as [9]

$$\mathrm{NRR} = 10\log_{10}\frac{\alpha_{\mathrm{n}}}{\mathcal{M}(\alpha_{\mathrm{n}},\beta,\eta,1)}. \tag{12}$$

E. Musical-noise-free speech enhancement [10]

In [10], we have proposed musical-noise-free noise reduction, where no musical noise is generated even for a high SNR in iterative SS. In the study, first, some of the authors discovered an interesting phenomenon that the kurtosis ratio sometimes does not change even after SS via mathematical analysis based on (11) [9]. This indicates that the kurtosis ratio can be maintained at unity even after iteratively applying SS to improve the NRR, and thus no musical noise is generated. Following this finding, the authors have derived the optimal parameters satisfying the musical-noise-free condition by finding a fixed-point status in the kurtosis ratio, i.e., by solving $\mathcal{M}(\alpha_{\mathrm{n}},0,0,4)/\mathcal{M}^2(\alpha_{\mathrm{n}},0,0,2) = \mathcal{M}(\alpha_{\mathrm{n}},\beta,\eta,4)/\mathcal{M}^2(\alpha_{\mathrm{n}},\beta,\eta,2)$ [10]. Given the noise shape parameter α_{n}, we can choose combinations of the over-subtraction parameter β and the flooring parameter η that simultaneously satisfy the musical-noise-free condition using the following equation;

$$\begin{aligned}\eta^4 =& \{\mathcal{F}(\alpha_{\mathrm{n}},\beta,4)(\alpha_{\mathrm{n}}+1)\alpha_{\mathrm{n}} - \mathcal{F}^2(\alpha_{\mathrm{n}},\beta,2)(\alpha_{\mathrm{n}}+3)(\alpha_{\mathrm{n}}+2)\}^{-1} \\ & \Bigg[\mathcal{S}(\alpha_{\mathrm{n}},\beta,2)\mathcal{F}(\alpha_{\mathrm{n}},\beta,2)(\alpha_{\mathrm{n}}+3)(\alpha_{\mathrm{n}}+2) \\ & \pm \Big[\{\mathcal{S}(\alpha_{\mathrm{n}},\beta,2)\mathcal{F}(\alpha_{\mathrm{n}},\beta,2)(\alpha_{\mathrm{n}}+3)(\alpha_{\mathrm{n}}+2)\}^2 \\ & - \big\{\mathcal{F}(\alpha_{\mathrm{n}},\beta,4)(\alpha_{\mathrm{n}}+1)\alpha_{\mathrm{n}} - \mathcal{F}^2(\alpha_{\mathrm{n}},\beta,2)(\alpha_{\mathrm{n}}+3)(\alpha_{\mathrm{n}}+2)\big\} \\ & \big\{\mathcal{S}(\alpha_{\mathrm{n}},\beta,4)(\alpha_{\mathrm{n}}+1)\alpha_{\mathrm{n}} - \mathcal{S}^2(\alpha_{\mathrm{n}},\beta,2)(\alpha_{\mathrm{n}}+3)(\alpha_{\mathrm{n}}+2)\big\}\Big]^{\frac{1}{2}}\Bigg].\end{aligned} \tag{13}$$

Figure 2 shows an example of the kurtosis ratio in optimized iterative SS, where Gaussian noise is assumed. We can confirm the flat trace of the kurtosis, indicating no musical noise generation.

III. Proposed Method: Extension to Microphone Array Signal Processing

A. Conventional blind spatial subtraction array

In the previous section, we assumed that the input noise signal is stationary, meaning that we can estimate the expectation of a noise signal from a time-frequency period of a signal that contains only noise, i.e., speech absence. However, in actual environments, e.g., a nonstationary noise field, it is necessary to dynamically estimate the noise power spectral density.

To solve this problem, we previously proposed blind spatial subtraction array (BSSA) [17], which involves accurate noise estimation by ICA followed by a speech extraction procedure based on SS (see Fig. 3). BSSA improves the noise reduction performance, particularly in the presence of both of diffuse and nonstationary noises; thus, almost all the environmental noise can be dealt with. However, BSSA always suffers from musical noise owing to SS. In addition, the output signal of BSSA degenerates to a *monaural* (not multi-channel) signal, meaning that ICA cannot be reapplied; thus, we cannot iteratively estimate the noise power spectra. Therefore, it is impossible to directly apply iterative SS to the conventional BSSA.

B. Iterative blind spatial subtraction array [11]

In this section, we propose a new multi-iterative blind signal extraction method integrating iterative blind noise estimation by ICA and iterative noise reduction by SS. As mentioned previously, the conventional BSSA cannot iteratively and accurately estimate noise by ICA because the conventional BSSA performs a delay and sum (DS) operation before SS. To solve this problem, we propose a new BSSA structure that performs multiple independent SS in each channel before DS; we call this structure *channel-wise SS* [18], [19], [20]. Using this structure, we can equalize the number of channels of the observed signal to that of the signals after channel-wise SS. Therefore, we can iteratively apply noise estimation by ICA and speech extraction by SS (see Fig. 4). Also, the advantage of the proposed structure is that ICA has the possibility of adaptively estimating the *distorted wavefront* of a speech signal to some extent even after SS, because ICA is a blind signal identification method that does not require knowledge of the target signal direction. Details of this issue will be discussed in Sect. III-C. Hereafter, we refer to this proposed BSSA as *iterative BSSA*.

We conduct iterative BSSA in the following manner, where the superscript $[i]$ represents the value in the ith iteration of SS (initially $i = 0$).

(I) The observed signal vector of the K-channel array in the time-frequency domain, $\boldsymbol{x}^{[0]}(f,\tau)$, is given by

$$\boldsymbol{x}^{[0]}(f,\tau) = \boldsymbol{h}(f)s(f,\tau) + \boldsymbol{n}(f,\tau), \tag{14}$$

where $\boldsymbol{h}(f) = [h_1(f), h_2(f)\ldots, h_K(f)]^{\mathrm{T}}$ is a column vector of the transfer functions from the target signal position to each microphone, $s(f,\tau)$ is the target speech signal, and $\boldsymbol{n}(f,\tau)$ is a column vector of the additive noise.

(II) Next, we perform signal separation using ICA as [13]

$$\boldsymbol{o}^{[i]}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f)\boldsymbol{x}^{[i]}(f,\tau), \tag{15}$$

$$\boldsymbol{W}_{\mathrm{ICA}}^{[i][p+1]}(f) = \mu[\boldsymbol{I} - \langle \boldsymbol{\varphi}(\boldsymbol{o}^{[i]}(f,\tau))(\boldsymbol{o}^{[i]}(f,\tau))^{\mathrm{H}}\rangle_{\tau}] \cdot \boldsymbol{W}_{\mathrm{ICA}}^{[i][p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[i][p]}(f), \tag{16}$$

where $\boldsymbol{W}_{\mathrm{ICA}}^{[i][p]}(f)$ is a demixing matrix, μ is the step-size parameter, $[p]$ is used to express the value of the pth step in the ICA iterations, $\boldsymbol{I}$ is the identity matrix, $\langle\cdot\rangle_{\tau}$ denotes a time-averaging operator, and $\boldsymbol{\varphi}(\cdot)$ is an appropriate nonlinear vector function. Then, we construct a *noise-only vector*,

$$\boldsymbol{o}_{\mathrm{noise}}^{[i]}(f,\tau) = [o_1^{[i]}(f,\tau),\ldots,o_{U-1}^{[i]},0, o_{U+1}^{[i]}(f,\tau),\ldots,o_K^{[i]}(f,\tau)]^{\mathrm{T}}, \tag{17}$$

where U is the signal number for speech, and we apply the projection back operation to remove the ambiguity of the amplitude and construct the estimated noise signal, $\boldsymbol{z}^{[i]}(f,\tau)$, as

$$\boldsymbol{z}^{[i]}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f)^{-1}\boldsymbol{o}_{\mathrm{noise}}^{[i]}(f,\tau). \tag{18}$$

(III) Next, we perform SS independently in each input channel and derive the multiple target-speech-enhanced signals. This procedure can be given by

$$x_k^{[i+1]}(f,\tau) = \begin{cases} \sqrt{|x_k^{[i]}(f,\tau)|^2 - \beta|z_k^{[i]}(f,\tau)|^2}\exp(j\arg(x_k^{[i]}(f,\tau))) \\ (\text{if}\quad |x_k^{[i]}(f,\tau)|^2 > \beta|z_k^{[i]}(f,\tau)|^2), \\ \eta x_k^{[i]}(f,\tau)\quad (\text{otherwise}), \end{cases} \tag{19}$$

where $x_k^{[i+1]}(f,\tau)$ is the target-speech-enhanced signal obtained by SS at a specific channel k. Then we return to step (II) with $\boldsymbol{x}^{[i+1]}(f,\tau)$. When we obtain sufficient noise reduction performance, go to step (IV).

(IV) Finally, we obtain the resultant target-speech-enhanced signal by applying DS to $\boldsymbol{x}^{[*]}(f,\tau)$, where $*$ is the number of iterations after which sufficient noise reduction performance is obtained. This procedure can be expressed by

$$y(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{x}^{[*]}(f,\tau), \tag{20}$$

$$\boldsymbol{w}_{\mathrm{DS}}(f) = [w_1^{(\mathrm{DS})}(f),\ldots,w_K^{(\mathrm{DS})}(f)], \tag{21}$$

$$w_k^{(\mathrm{DS})}(f) = \frac{1}{K}\exp(-2j(f/N)f_{\mathrm{s}}d_k\sin\theta_U/c), \tag{22}$$

$$\theta_U = \sin^{-1}\frac{\arg\left(\frac{\left[\boldsymbol{W}_{\mathrm{ICA}}^{[*]}(f)^{-1}\right]_{kU}}{\left[\boldsymbol{W}_{\mathrm{ICA}}^{[*]}(f)^{-1}\right]_{k'U}}\right)}{2\pi f_{\mathrm{s}}c^{-1}(d_k - d_{k'})}, \tag{23}$$

where $y(f,\tau)$ is the final output signal of iterative BSSA, $\boldsymbol{w}_{\mathrm{DS}}$ is the filter coefficient vector of DS, N is the DFT size, f_{s} is the sampling frequency, d_k is the microphone position, c is the sound velocity, and θ_U is the estimated direction of arrival of the target speech. Moreover, $[\boldsymbol{A}]_{lj}$ represents the entry of $\boldsymbol{A}$ in the lth row and jth column.

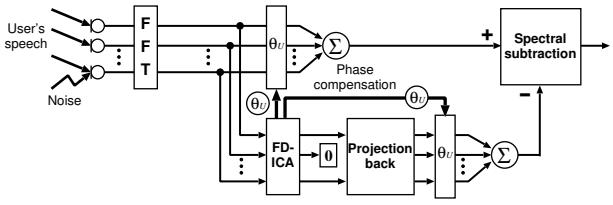

Fig. 3. Block diagram of conventional BSSA [17].

C. Accuracy of wavefront estimated by ICA after SS

In this subsection, we discuss the accuracy of the estimated noise signal in each iteration of iterative BSSA. In actual environments, not only point-source noise but also non-point-source (e.g., diffuse) noise often exists. It is known that ICA is proficient in noise estimation rather than speech estimation under such a noise condition [17]. This is because the target speech can be regarded as a point-source signal (thus, the wavefront is static in each subband) and ICA acts as an effective blocking filter of the speech wavefront even in a time-invariant manner, resulting in good noise estimation. However, in iterative BSSA, we should address the inherent question of whether the distorted speech wavefront after nonlinear noise reduction such as SS can be blocked by ICA or not; thus, the speech component after channel-wise SS can become a point source again or not.

Hereafter, we quantify the degree of point-source-likeness for SS-applied speech signals. For convenience of discussion, a simple two-channel array model is assumed. First, we define the speech component in each channel after channel-wise SS as

$$\hat{s}_1(f,\tau) = h_1(f)s(f,\tau) + \Delta s_1(f,\tau), \tag{24}$$

$$\hat{s}_2(f,\tau) = h_2(f)s(f,\tau) + \Delta s_2(f,\tau), \tag{25}$$

where $s(f,\tau)$ is the original point-source speech signal, $\hat{s}_k(f,\tau)$ is the speech component after channel-wise SS at the kth channel, and $\Delta s_k(f,\tau)$ is the speech component distorted by channel-wise SS. Also, we assume that $s(f,\tau)$, $\Delta s_1(f,\tau)$, and $\Delta s_2(f,\tau)$ are uncorrelated with each other. Obviously, $\hat{s}_1(f,\tau)$ and $\hat{s}_2(f,\tau)$ can be regarded as being generated by a point source if $\Delta s_1(f,\tau)$ and $\Delta s_2(f,\tau)$ are zero, i.e., a valid static blocking filter can be obtained by ICA as

$$\begin{aligned}&[\boldsymbol{W}_{\mathrm{ICA}}(f)]_{11}\hat{s}_1(f,\tau) + [\boldsymbol{W}_{\mathrm{ICA}}(f)]_{12}\hat{s}_2(f,\tau)\\ &= ([\boldsymbol{W}_{\mathrm{ICA}}(f)]_{11}h_1(f) + [\boldsymbol{W}_{\mathrm{ICA}}(f)]_{12}h_2(f))s(f,\tau)\\ &= 0,\end{aligned} \tag{26}$$

where we assume $U = 1$ and, e.g., $[\boldsymbol{W}_{\mathrm{ICA}}(f)]_{11} = h_2(f)$ and $[\boldsymbol{W}_{\mathrm{ICA}}(f)]_{12} = -h_1(f)$. However, if $\Delta s_1(f,\tau)$ and $\Delta s_2(f,\tau)$ become nonzero as a result of SS, ICA does not

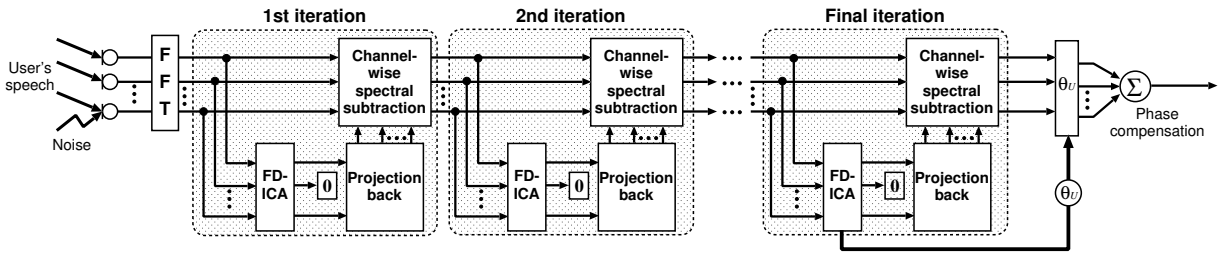

Fig. 4. Block diagram of proposed iterative BSSA.

have a valid speech blocking filter with a static (time-invariant) form.

Second, the cosine distance between speech power spectra $|\hat{s}_1(f,\tau)|^2$ and $|\hat{s}_2(f,\tau)|^2$ is introduced in each frequency subband to indicate the degree of point-source-likeness, as

$$\mathrm{COS}(f) = \frac{\sum_\tau |\hat{s}_1(f,\tau)|^2 |\hat{s}_2(f,\tau)|^2}{\sqrt{\sum_\tau |\hat{s}_1(f,\tau)|^4}\sqrt{\sum_\tau |\hat{s}_2(f,\tau)|^4}}. \tag{27}$$

From (27), the cosine distance reaches its maximum value of unity if and only if $\Delta s_1(f,\tau) = \Delta s_2(f,\tau) = 0$, regardless of the values of $h_1(f)$ and $h_2(f)$, meaning that the SS-applied speech signals $\hat{s}_1(f,\tau)$ and $\hat{s}_2(f,\tau)$ can be assumed to be produced by the point source. The value of $\mathrm{COS}(f)$ decreases with increasing magnitudes of $\Delta s_1(f,\tau)$ and $\Delta s_2(f,\tau)$ as well as the difference between $h_1(f)$ and $h_2(f)$; this indicates the non-point-source state.

Third, we evaluate the degree of point-source-likeness in each iteration of iterative BSSA by using $\mathrm{COS}(f)$. We statistically estimate the distorted speech component of the enhanced signal in each iteration. Here, we assume that the original speech power spectrum $|s(f,\tau)|^2$ obeys a gamma distribution with a shape parameter of 0.1 (this is a typical value for speech) as

$$|s(f,\tau)|^2 \sim \frac{x^{-0.9}}{\Gamma(0.1)\theta_{\mathrm{s}}^{0.1}} \exp(-x/\theta_{\mathrm{s}}), \tag{28}$$

where θ_{s} is the speech scale parameter. Regarding the amount of noise to be subtracted, the 1st-order moment of the noise power spectra is equal to $\theta_{\mathrm{n}}\alpha_{\mathrm{n}}$ when the number of iterations, i, equals zero. Also, the value of α_{n} does not change in each iteration when we use the specific parameters β and η that satisfy the musical-noise-free condition because the kurtosis ratio does not change in each iteration. If we perform SS only once, the rate of noise decrease is given by

$$\mathcal{M}(\alpha_{\mathrm{n}},\beta,\eta,1)/\alpha_{\mathrm{n}}, \tag{29}$$

and thus, the amount of residual noise after the ith iteration is given by

$$\begin{aligned}\mu_1^{[i]} &= \theta_{\mathrm{n}}\alpha_{\mathrm{n}}\left\{\mathcal{M}(\alpha_{\mathrm{n}},\beta,\eta,1)/\alpha_{\mathrm{n}}\right\}^i \\ &= \theta_{\mathrm{n}}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{1-i}.\end{aligned} \tag{30}$$

Next, we assume that speech and noise are disjoint, i.e., there are no overlaps in the time-frequency domain, and that speech distortion is caused by subtracting the average noise from the pure speech component. Thus, the speech component $|\hat{s}_k^{[i+1]}(f,\tau)|^2$ at the kth channel after the ith iteration is represented by subtracting the amount of residual noise (30) as

$$|\hat{s}_k^{[i+1]}(f,\tau)|^2 = \begin{cases} |\hat{s}_k^{[i]}(f,\tau)|^2 - \beta\theta_{\mathrm{n}}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{1-i} \\ (\text{if} \quad |\hat{s}_k^{[i]}(f,\tau)|^2 > \beta\theta_{\mathrm{n}}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{1-i}), \\ \eta^2|\hat{s}_k^{[i]}(f,\tau)|^2 \quad (\text{otherwise}). \end{cases} \tag{31}$$

Here, we define the input SNR as the average of both channel SNRs,

$$\begin{aligned}\mathrm{ISNR}(f) &= \frac{1}{2}\left(\frac{0.1|h_1(f)|^2\theta_{\mathrm{s}}}{\alpha_{\mathrm{n}}\theta_{\mathrm{n}}} + \frac{0.1|h_2(f)|^2\theta_{\mathrm{s}}}{\alpha_{\mathrm{n}}\theta_{\mathrm{n}}}\right) \\ &= \frac{0.1\theta_{\mathrm{s}}}{2\alpha_{\mathrm{n}}\theta_{\mathrm{n}}}(|h_1(f)|^2 + |h_2(f)|^2).\end{aligned} \tag{32}$$

If we normalize the speech scale parameter θ_{s} to unity, from (32), the noise scale parameter θ_{n} is given by

$$\theta_{\mathrm{n}} = \frac{0.1(|h_1(f)|^2 + |h_2(f)|^2)}{2\alpha_{\mathrm{n}}\mathrm{ISNR}(f)}, \tag{33}$$

and using (33), we can reformulate (31) as

$$|\hat{s}_k^{[i+1]}(f,\tau)|^2 = \begin{cases} |\hat{s}_k^{[i]}(f,\tau)|^2 - \beta\frac{0.1(|h_1(f)|^2+|h_2(f)|^2)}{2\mathrm{ISNR}(f)}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{-i} \\ (\text{if } |\hat{s}_k^{[i]}(f,\tau)|^2 > \beta\frac{0.1(|h_1(f)|^2+|h_2(f)|^2)}{2\mathrm{ISNR}(f)}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{-i}), \\ \eta^2|\hat{s}_k^{[i]}(f,\tau)|^2 \quad (\text{otherwise}). \end{cases} \tag{34}$$

Furthermore, we define the transfer function ratio (TFR) as

$$\mathrm{TFR}(f) = |h_1(f)/h_2(f)|^2, \tag{35}$$

and if we normalize $|h_1(f)|^2$ to unity in each frequency subband, $|h_1(f)|^2 + |h_2(f)|^2$ becomes $1 + 1/\mathrm{TFR}(f)$. Finally, we express (34) in terms of the input SNR $\mathrm{ISNR}(f)$ and the transfer function ratio $\mathrm{TFR}(f)$ as

$$|\hat{s}_k^{[i+1]}(f,\tau)|^2 = \begin{cases} |\hat{s}_k^{[i]}(f,\tau)|^2 - \beta\frac{0.1(1+1/\mathrm{TFR}(f))}{2\mathrm{ISNR}(f)}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{-i} \\ (\text{if} \quad |\hat{s}_k^{[i]}(f,\tau)|^2 > \beta\frac{0.1(1+1/\mathrm{TFR}(f))}{2\mathrm{ISNR}(f)}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{-i}), \\ \eta^2|\hat{s}_k^{[i]}(f,\tau)|^2 \quad (\text{otherwise}). \end{cases} \tag{36}$$

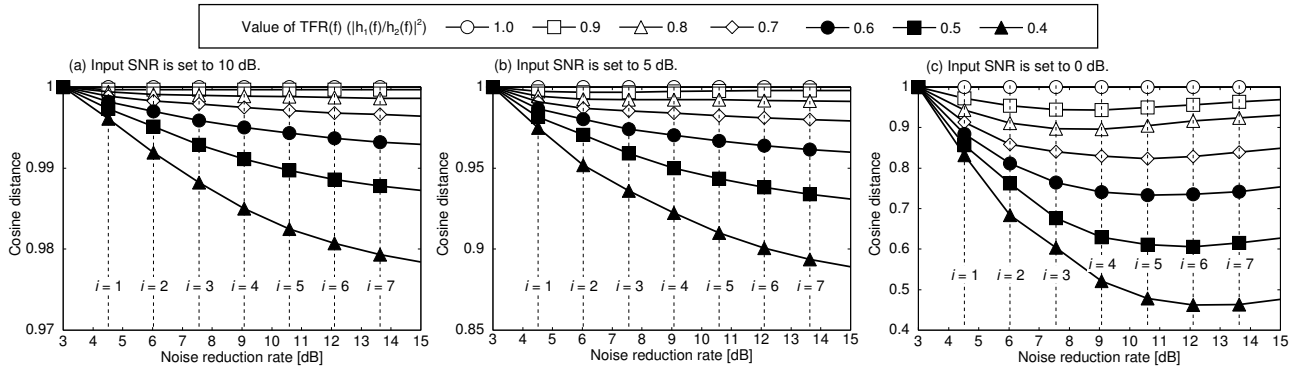

Fig. 5. Relation between number of iterations of iterative BSSA and cosine distance. Input SNR is (a) 10 dB, (b) 5 dB, and (c) 0 dB.

As can be seen, the speech component is subjected to greater subtraction and distortion as $\mathrm{ISNR}(f)$ and/or $\mathrm{TFR}(f)$ decrease.

Figure 5 shows the relation between the TFR and the corresponding value of $\mathrm{COS}(f)$ calculated by (27) and (36). In Fig. 5, we plot the average of $\mathrm{COS}(f)$ over the whole frequency subbands. The noise shape parameter α_{n} is set to 0.2 with the assumption of super-Gaussian noise (this corresponds to the real noises used in Sect. V), the input SNR is set to 10 dB, 5 dB, or 0 dB, and the noise scale parameter θ_{n} is uniquely determined by (33) and the previous parameter settings. The TFR is set from 0.4 to 1.0 ($|h_1(f)|$ is fixed to 1.0). Note that the TFR is highly correlated to the room reverberation and the interelement spacing of the microphone array; we determined the range of the TFR by simulating a typical moderately reverberant room and the array with 2.15 cm interelement spacing used in Sect. V (see the example of the TFR in Fig. 6). For the internal parameters used in iterative BSSA in this simulation, β and η are 8.5 and 0.9, respectively, which satisfy the musical-noise-free condition. In addition, the smallest value on the horizontal axis is 3 dB in Fig. 5 because DS is still performed even when $i = 0$.

From Figs. 5(a) and (b), which correspond to relatively high input SNRs, we can confirm that the degree of point-source-likeness, i.e., $\mathrm{COS}(f)$, is almost maintained when the TFR is close to 1 even if the speech components are distorted by iterative BSSA. Also, it is worth mentioning that the degree of point-source-likeness is still above 0.9 even when the TFR is decreased to 0.4 and i is increased to 6. This means that almost 90% of the speech components can be regarded as a point source and thus can be blocked by ICA. In contrast, from Fig. 5(c), which shows the case of a low input SNR, when the TFR is dropped to 0.4 and i is more than 3, the degree of point-source-likeness is lower than 0.6. Thus, less than 60% of the speech components can be regarded as a point source. However, this is a worst-case scenario; actually when the TFR is dropped to 0.4 and i is more than 3, the degree of point-source-likeness is lower than 0.6. Thus, more than 40% of the speech components cannot be regarded as a point source, and this leads to poor noise estimation.

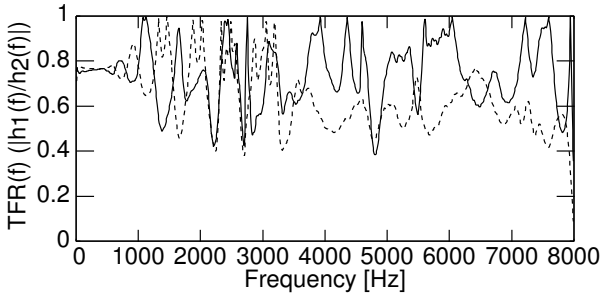

Fig. 6. Typical examples of $\mathrm{TFR}(f)$ ($|h_1(f)/h_2(f)|^2$) in each frequency subband, where solid and broken lines are different combinations of microphones.

IV. Improvement scheme for poor noise estimation

A. Channel selection in ICA

In this subsection, we propose a channel selection strategy in ICA for achieving high accuracy of noise estimation. As mentioned previously, speech distortion is subjected to $\mathrm{ISNR}(f)$ and $\mathrm{TFR}(f)$, and the accuracy of noise estimation is degraded along with its speech distortion. Figure 6 shows a typical example of the TFR. From Fig. 6, we can confirm that the TFRs in different combinations of microphones are not the same in each frequency subband; in a specific frequency, one microphone pair has higher $\mathrm{TFR}(f)$ than another pair, and vice versa in another frequency. Thus, we are able to select the appropriate combination of the microphones to obtain higher TFR.

Therefore, we introduce the channel selection method into ICA in each frequency subband, where we automatically choose less varied inputs to maintain high accuracy of noise estimation. Hereafter, we describe the detail of the channel selection method. First, we calculate the average power of the observed signal $x_k(f,\tau)$ at the kth channel as

$$\mathrm{E}_\tau[|x_k(f,\tau)|^2] = \mathrm{E}_\tau[|s(f,\tau)|^2]|h_k(f)|^2 + \mathrm{E}_\tau[|n_k(f,\tau)|^2]. \tag{37}$$

Here, $\mathrm{E}_\tau[|s(f,\tau)|^2]$ is a constant, and if we assume the diffuse noise field, $\mathrm{E}_\tau[|n_k(f,\tau)|^2]$ is also a constant. Thus, we can

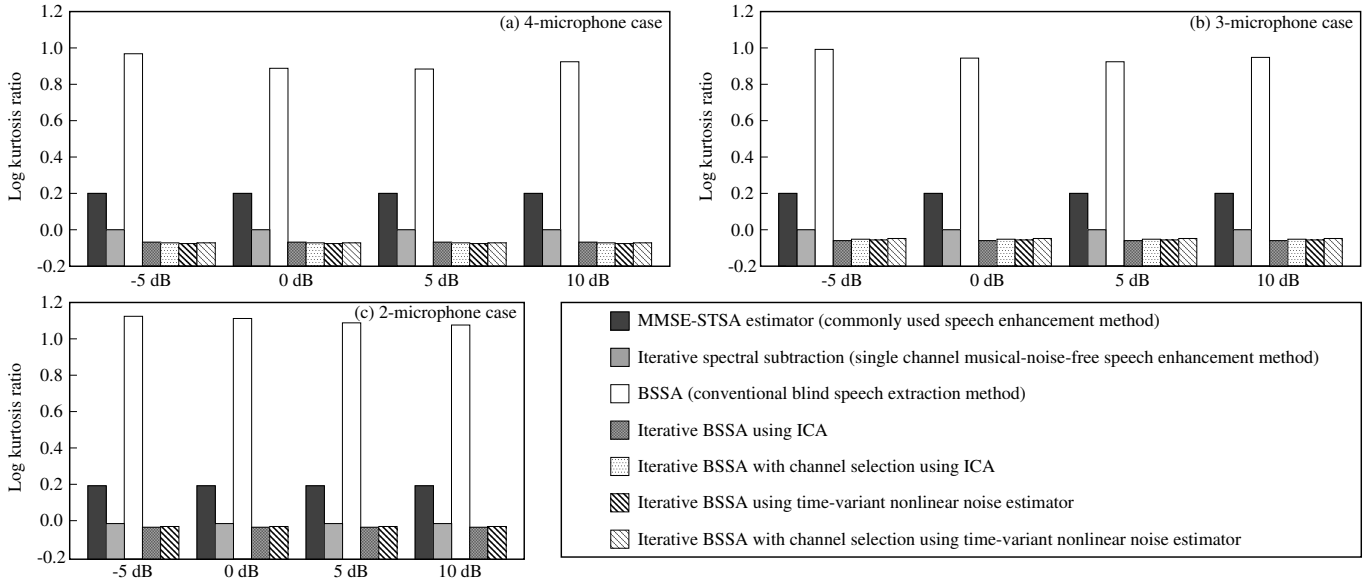

Fig. 7. Kurtosis ratio obtained from experiment for traffic noise under 10-dB NRR condition.

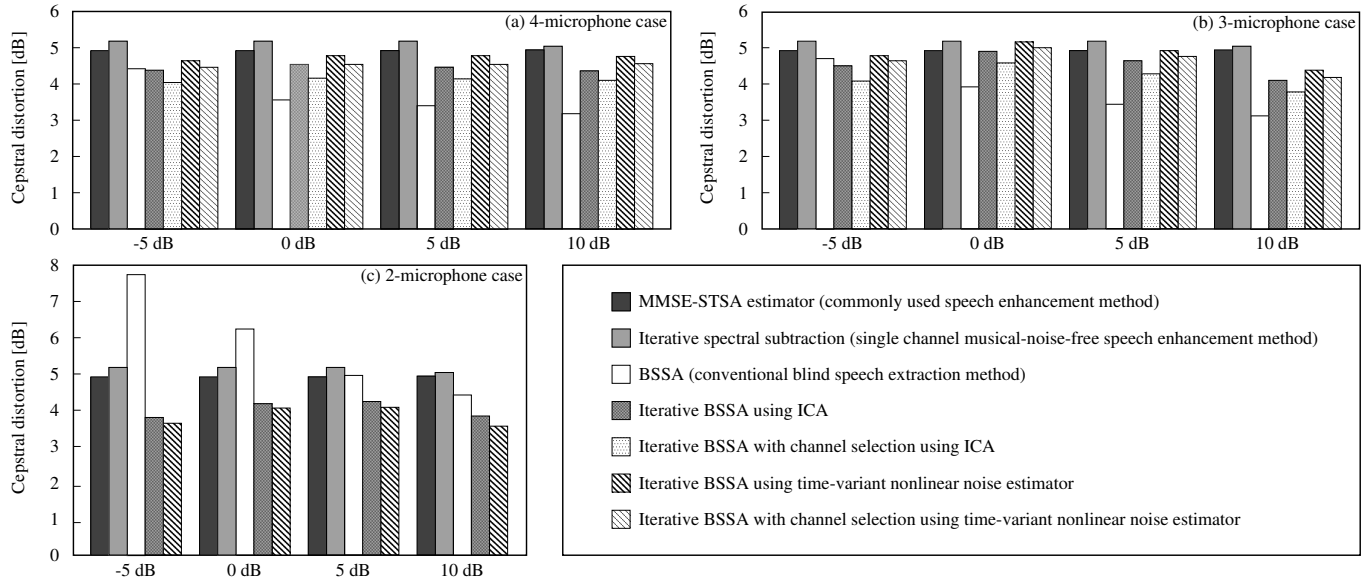

Fig. 8. Cepstral distortion obtained from experiment for traffic noise under 10-dB NRR condition.

estimate the relative order of $|h_k(f)|^2$ by comparing (37) for every k.

Next, we sort $\mathrm{E}_\tau[|x_k(f,\tau)|^2]$ in descending order and select the channels k corresponding to high amplitude of $|h_k(f)|^2$ satisfying the following condition:

$$\max_k \mathrm{E}_\tau[|x_k(f,\tau)|^2] \cdot \xi \leq \mathrm{E}_\tau[|x_k(f,\tau)|^2], \qquad (38)$$

where $\xi(<1)$ is the threshold for the selection.

Finally, we perform noise estimation based on ICA using the selected channels in each frequency subband, and we apply the projection back operation to remove the ambiguity of the amplitude and construct the estimated noise signal.

B. Time-variant noise PSD estimator

In the previous section, we reveal that the speech components cannot be regarded as a point source, and this leads to poor noise estimation in iterative BSSA. To solve this problem, we introduce a time-variant noise PSD estimator [15] instead of ICA for improvement of the noise estimation accuracy. This method has been developed for future high-end binaural hearing aids and performs a prediction on the left noisy signal from the right noisy signal via an Wiener filter, followed by an auto-PSD of the difference between the left noisy signal and the prediction. By applying the noise PSD estimated from

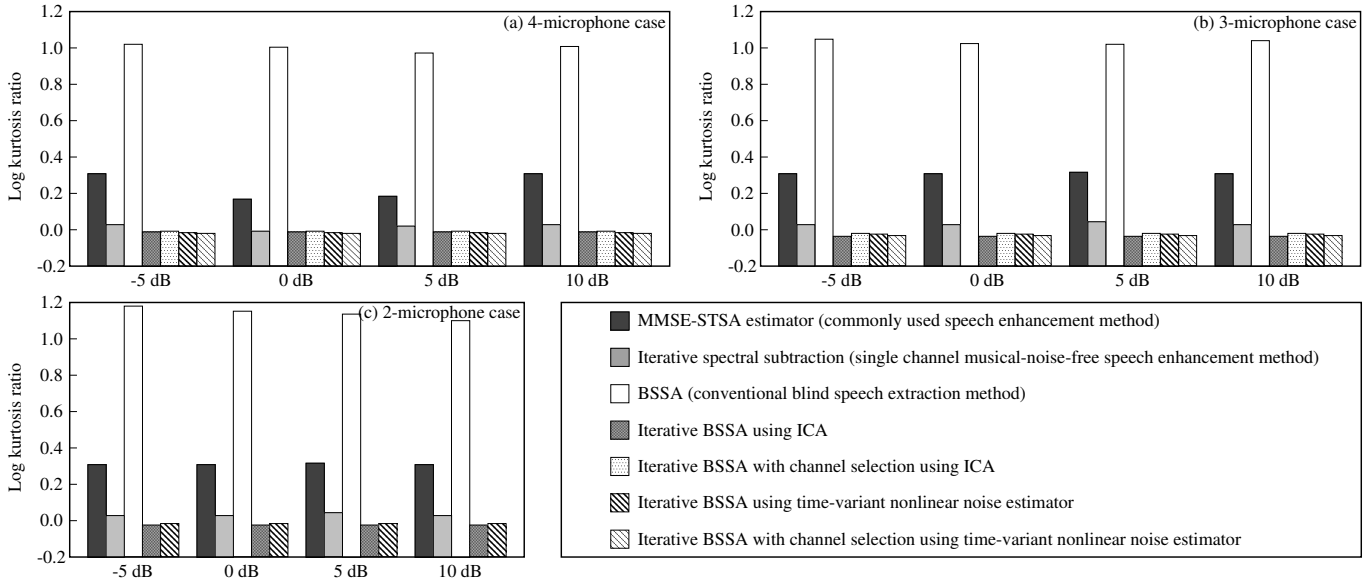

Fig. 9. Kurtosis ratio obtained from experiment for railway station noise under 10-dB NRR condition.

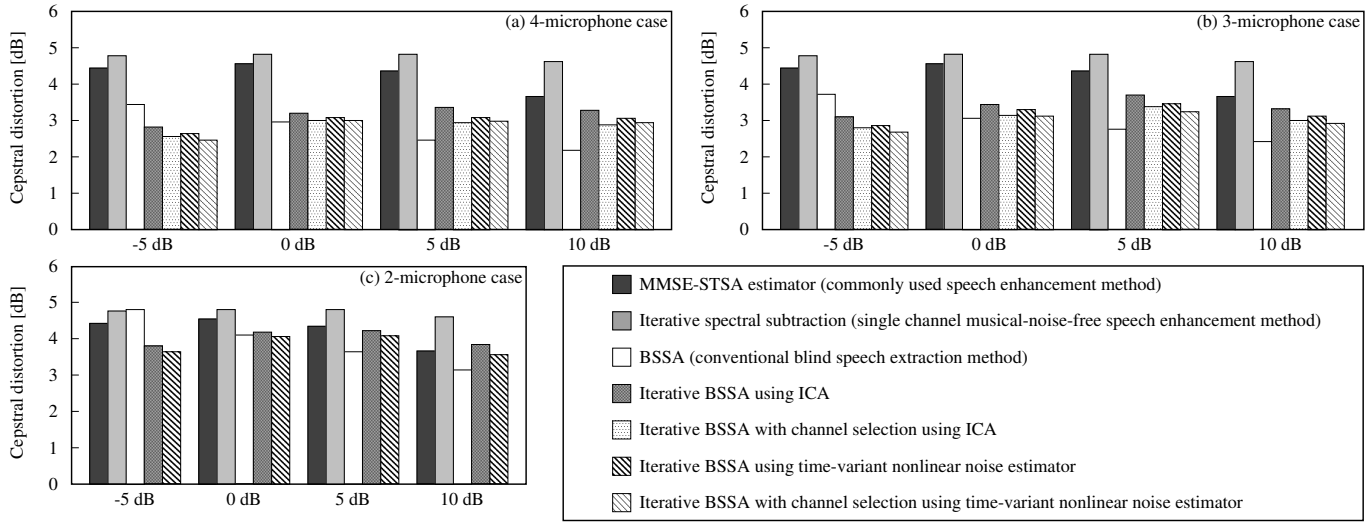

Fig. 10. Cepstral distortion obtained from experiment for railway station noise under 10-dB NRR condition.

this estimator to (19), we can perform the speech extraction. The procedure of this noise PSD estimator is described in Appendix.

V. Experiment in Real World

A. Experimental conditions

We conducted objective and subjective evaluation experiments to confirm the validity of the proposed methods under the diffuse and nonstationary noise condition. The size of the experimental room was $4.2\times3.5\times3.0\ \mathrm{m}^3$ and the reverberation time was approximately 200 ms. We used a two-, three- or four-element microphone array with an interelement spacing of 2.15 cm, and the direction of the target speech was set to be normal to the array. All the signals used in this experiment were sampled at 16 kHz with 16-bit accuracy. The FFT size was 1024, and the frame shift length was 256. We use 10 speakers (5 males and 5 females) as sources of the original target speech signal. The input SNR was -5, 0, 5, and 10 dB.

B. Objective evaluation

We conducted an objective experiment evaluation under the same NRR condition. Figures 7, 8, 9, and 10 show the kurtosis ratio and cepstral distortion obtaind from the experiments with real traffic noise and railway station noise, where we evaluate 10-dB NRR (i.e., output SNRs = 5, 10, 15, and 20 dB) signals processed by three conventional methods, namely, the minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator [21], simple combination

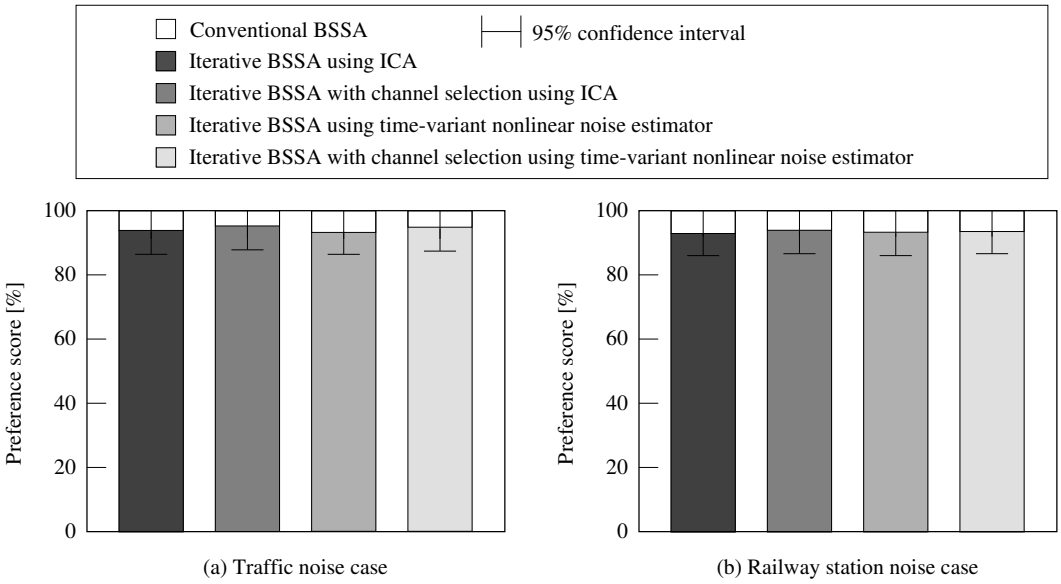

Fig. 11. Subjective evaluation results for (a) traffic noise and (b) railway station noise.

of BSSA and musical-noise-free iterative spectral subtraction, and our proposed methods, iterative BSSA (using ICA or time-variant noise estimator with/without channel selection), respectively. Here, we did not apply the channel selection method to the two-microphone case because ICA or time-variant noise estimation need at least two-channel signals. Also, we applied minimum statistics noise PSD estimator [5] to the MMSE STSA estimator and musical-noise-free iterative spectral subtraction, and we use the decision-directed approach for a priori SNR estimation in the MMSE STSA estimator. From Figs. 7 and 9, we can confirm that iterative BSSA methods outperform the MMSE STSA estimator and the conventional BSSA in the kurtosis ratio. In particular, the kurtosis ratios of the proposed methods are mostly close to 1.0. This means that the proposed iterative methods did not generate any musical noise. However, iterative BSSA methods leads to greater speech distortion compared with the conventional BSSA (see Figs. 8 and 10). Therefore, a trade-off exists between the amount of musical noise generation and speech distortion in the conventional BSSA and iterative BSSA methods.

C. Subjective evaluation

Since we found the above-mentioned trade-off, we next conducted a subjective evaluation for setting the performance competition. In the evaluation, we presented a pair of 10-dB NRR signals processed by the conventional BSSA and four our proposed iterative BSSAs (using ICA or time-variant noise estimator with/without channel selection) in random order to 10 examinees, who selected which signal they preferred from the viewpoint of total sound quality, e.g., less musical noise, less speech distortion, etc.

The result of experiment is shown in Fig. 11 for (a) traffic noise and (b) railway station noise. It is found that the output signals of some iterative BSSAs are preferred to that of the conventional BSSA, indicating the higher sound quality of the proposed method in terms of human perception. This result is plausible because humans are often more sensitive to musical noise than to speech distortion as indicated in past studies, e.g., [22].

VI. Conclusion

In this paper, we first proposed iterative BSSA using a new BSSA structure, which generates almost no musical noise even with increasing noise reduction performance. Our theoretical analysis indicates that the accuracy of noise estimation is degraded along with its speech distortion. Next, we conducted a channel selection strategy in ICA for achieving high accuracy of noise estimation. From the evaluation experiments, it was shown that there is a trade-off between the amount of musical noise generation and speech distortion in both the conventional BSSA and iterative BSSA. However, in a subjective preference test, iterative BSSA obtained a higher preference score than the conventional BSSA. Thus, iterative BSSA is advantageous to the conventional BSSA in terms of sound quality.

Appendix

This appendix provides a brief review of the time-variant nonlinear noise estimator. For more detailed information, Ref. [15] can be available.

Let $x_{\mathrm{L}}(f,\tau)$ and $x_{\mathrm{R}}(f,\tau)$ be noisy signals received at the left and right microphones in the time-frequency domain, defined as

$$x_{\mathrm{L}}(f,\tau) = h_{\mathrm{L}}(f)s(f,\tau) + n_{\mathrm{L}}(f,\tau), \tag{39}$$

$$x_{\mathrm{R}}(f,\tau) = h_{\mathrm{R}}(f)s(f,\tau) + n_{\mathrm{R}}(f,\tau), \tag{40}$$

where $h_{\mathrm{L}}(f)$ and $h_{\mathrm{R}}(f)$ are the left and right transfer functions, respectively. Next, the left and right auto-power spectral

densities, $\Gamma_{\mathrm{LL}}(f)$ and $\Gamma_{\mathrm{RR}}(f)$, can be expressed as follows:

$$\Gamma_{\mathrm{LL}}(f,\tau) = |H_{\mathrm{L}}(f)|^2\Gamma_{\mathrm{SS}}(f,\tau) + \Gamma_{\mathrm{NN}}(f,\tau), \quad (41)$$

$$\Gamma_{\mathrm{RR}}(f,\tau) = |H_{\mathrm{R}}(f)|^2\Gamma_{\mathrm{SS}}(f,\tau) + \Gamma_{\mathrm{NN}}(f,\tau), \quad (42)$$

where $\Gamma_{\mathrm{SS}}(f,\tau)$ is the power spectral density of the target speech signal, and $\Gamma_{\mathrm{NN}}(f,\tau)$ is the power spectral density of the noise signal. In this paper, we assume that the left and right noise power spectral densities are approximately the same, i.e., $\Gamma_{\mathrm{N_L N_L}}(f,\tau) \simeq \Gamma_{\mathrm{N_R N_R}}(f,\tau) \simeq \Gamma_{\mathrm{NN}}(f,\tau)$.

Next, we consider the Wiener solution between the left and right transfer functions, which is defined as

$$H_{\mathrm{W}}(f,\tau) = \frac{\Gamma_{\mathrm{LR}}(f,\tau)}{\Gamma_{\mathrm{RR}}(f,\tau)}, \quad (43)$$

where $\Gamma_{\mathrm{LR}}(f)$ is the cross-power spectral density between the left and the right noisy signals. The cross-power spectral density expression then becomes

$$\Gamma_{\mathrm{LR}}(f,\tau) = \Gamma_{\mathrm{SS}}(f,\tau)H_{\mathrm{L}}(f)H_{\mathrm{R}}^*(f). \quad (44)$$

Therefore, substituting (44) into (43) yields

$$H_{\mathrm{W}}(f,\tau) = \frac{\Gamma_{\mathrm{SS}}(f,\tau)H_{\mathrm{L}}(f)H_{\mathrm{R}}^*(f)}{\Gamma_{\mathrm{RR}}(f,\tau)}. \quad (45)$$

Furthermore, using (41) and (42), the squared magnitude response of the Wiener solution in (45) can be also expressed as

$$|H_{\mathrm{W}}(f,\tau)|^2 = \frac{(\Gamma_{\mathrm{LL}}(f,\tau) - \Gamma_{\mathrm{NN}}(f,\tau))(\Gamma_{\mathrm{RR}}(f,\tau) - \Gamma_{\mathrm{NN}}(f,\tau))}{\Gamma_{\mathrm{RR}}^2(f,\tau)}. \quad (46)$$

Equation (46) is rearranged into a quadratic equation as in the following:

$$\Gamma_{\mathrm{NN}}^2(f,\tau) - \Gamma_{\mathrm{NN}}(f,\tau)\left(\Gamma_{\mathrm{LL}}(f,\tau) + \Gamma_{\mathrm{RR}}(f,\tau)\right) + \Gamma_{\mathrm{EE}}(f,\tau)\Gamma_{\mathrm{RR}}(f,\tau) = 0, \quad (47)$$

where

$$\Gamma_{\mathrm{EE}}(f,\tau) = \Gamma_{\mathrm{LL}}(f,\tau) - \Gamma_{\mathrm{RR}}(f,\tau)|H_{\mathrm{W}}(f)|^2. \quad (48)$$

Consequently, the noise power spectral density $\Gamma_{\mathrm{NN}}(f)$ can be estimated by solving the quadratic equation in (47) as follows:

$$\Gamma_{\mathrm{NN}}(f,\tau) = \frac{1}{2}\left(\Gamma_{\mathrm{LL}}(f,\tau) + \Gamma_{\mathrm{RR}}(f,\tau)\right) - \Gamma_{\mathrm{LRavg}}(f,\tau), \quad (49)$$

$$\Gamma_{\mathrm{LRavg}}(f,\tau) = \frac{1}{2}\{(\Gamma_{\mathrm{LL}}(f,\tau) + \Gamma_{\mathrm{RR}}(f,\tau))^2 - 4\Gamma_{\mathrm{EE}}(f,\tau)\Gamma_{\mathrm{RR}}(f,\tau)\}^{0.5}. \quad (50)$$

ACKNOWLEDGMENT

This work was supported by JST Core Research of Evaluational Science and Technology (CREST) and Grant-in-Aid for JSPS Fellows.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement Theory and Practice* CRC Press, Taylor & Francis Group FL, 2007.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol.27, no.2, pp.113–120, 1979.

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. ICASSP79*, pp.208–211, 1979.

[4] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.28, no.2, pp.137–145, 1980.

[5] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. EUSIPCO94*, pp.1182–1185, 1994.

[6] K. Yamashita, S. Ogata, T. Shimamura, "Spectral subtraction iterated with weighting factors," *Proc. IEEE Speech Coding Workshop*, pp.138–140, 2002.

[7] M. R. Khan, T. Hansen, "Iterative noise power subtraction technique for improved speech quality," *Proc. ICECE2008*, pp.391–394, 2008.

[8] S. Li, J.-Q. Wang, M. Niu, X.-J. Jing, T. Liu, "Iterative spectral subtraction method for millimeter-wave conducted speech enhancement," *Journal of Biomedical Science and Engineering*, vol.2010, no.3, pp.187–192, 2010.

[9] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, K. Kondo, "Theoretical analysis of iterative weak spectral subtraction via higher-order statistics," *Proc. MLSP2010*, pp.220–225, 2010.

[10] R. Miyazaki, H Saruwatari, T Inoue, Y. Takahashi, K. Shikano, K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Transactions on Audio, Speech and Language Processing,* vol.20, no.7, pp.2080–2094, 2012.

[11] R. Miyazaki, H. Saruwatari, K. Shikano, K. Kondo, "Musical-noise-free blind speech extraction using ICA-based noise estimation and iterative spectral subtraction," *Proc. ISSPA2012*, pp.322–327, 2012.

[12] R. Miyazaki, H. Saruwatari, K. Shikano, K. Kondo, "Musical-noise-free blind speech extraction using ICA-based noise estimation and iterative spectral subtraction with channel selection," *Proc. IWAENC2012*, 2012.

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

[14] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Speech and Audio Processing*, vol.14, no.2, pp.666–678, 2006.

[15] A. Homayoun, M. bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.17, no.4, pp.521–533, 2009.

[16] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," *Proc. IWAENC2008*, 2008.

[17] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.17, no.4, pp.650–664, 2009.

[18] Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, "Musical-noise analysis in methods of integrating microphone array and spectral subtraction based on higher-order statistics," *EURASIP Jouranl on Advances in Signal Processing*, vol.2010, Article ID 431347, 25 pages, 2010.

[19] H. Saruwatari, Y. Ishikawa, Y. Takahashi, T. Inoue, K. Shikano, K. Kondo, "Musical noise controllable algorithm of channelwise spectral subtraction and adaptive beamforming based on higher-order statistics," *IEEE Transactions on Audio, Speech and Language Processing*, vol.19, no.6, pp.1457–1466, 2011.

[20] R. Miyazaki, H. Saruwatari, K. Shikano, "Theoretical analysis of amounts of musical noise and speech distortion in structure-generalized parametric spatial subtraction array," *IEICE Transactions Fundamentals*, vol.95-A, no.2, pp.586–590, 2012.

[21] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.32, no.6, pp.1109–1121, 1984.

[22] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, "Musical noise generation analysis for noise reduction methods based on spectral subtraction and MMSE STSA estimation," *Proc. ICASSP*, pp.4433–4436, 2009.