ミュージカルノイズフリー音声抽出における音声カートシス比に基づく反復回数の制御*

☆平野佑佳，宮崎亮一，猿渡洋，中村哲 (奈良先端大)

1 Introduction

To achieve high-quality speech enhancement, noise reduction using a microphone array has been widely studied. In recent years, we previously proposed a BSSA [1] that consists of accurate noise estimation by independent component analysis (ICA) [2] and the following speech extraction procedure based on nonlinear noise reduction such as spectral subtraction (SS). However, BSSA always suffers from artificial distortion, so-called musical noise, owing to nonlinear signal processing.

To solve this problem, we have proposed iterative BSSA [3] consisting of dynamic noise estimation by ICA and musical-noise-free iterative SS [4]. This method can perform noise reduction with perfectly no musical noise even with increasing the number of iterations in SS, but instead always suffers from speech distortion. Since speech distortion cannot be measured without a clean *reference* speech signal, we should have to decide the number of iterations manually based on our auditory perception.

Recently, we have proposed a speech kurtosis ratio (4th-order statistics) as a new unsupervised measurement of speech distortion [5]. Therefore, in this paper, first, we propose a new speech kurtosis estimation method using the noise signal estimated by ICA. Next, we propose an automatic control of the number of iterations based on the speech kurtosis ratio estimated using ICA in iterative BSSA.

2 Related Works

2.1 Musical-Noise-Free Iterative BSSA

In this section, we describe iterative BSSA that can perform noise reduction with perfectly no musical noise in nonstationary noise (see Fig. 1). This method consists of iterative blind dynamic noise estimation by ICA and musical-noise-free speech extraction by modified iterative SS, where multiple iterative SS are applied to each channel while maintaining the multi-channel property reused for ICA.

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

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

$$\boldsymbol{x}^{[0]}(f,\tau) = \boldsymbol{h}(f)s(f,\tau) + \boldsymbol{n}(f,\tau), \tag{1}$$

where f denotes the frequency subband, τ is the frame index, $\boldsymbol{h}(f) = [h_1(f), h_2(f) \ldots, h_K(f)]^{\mathrm{T}}$ 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.

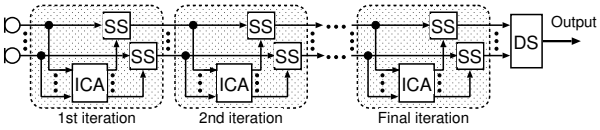

Fig. 1 Block diagram of iterative BSSA.

Step(II) We perform signal separation using ICA

$$\boldsymbol{o}^{[i]}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f)\boldsymbol{x}^{[i]}(f,\tau), \tag{2}$$

$$\boldsymbol{W}_{\mathrm{ICA}}^{[i][p+1]}(f) = \mu[\boldsymbol{I} - \langle \boldsymbol{\varphi}(\boldsymbol{o}^{[i]}(f,\tau))(\boldsymbol{o}^{[i]}(f,\tau))^{\mathrm{H}}\rangle_\tau] \cdot \boldsymbol{W}_{\mathrm{ICA}}^{[i][p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[i][p]}(f), \tag{3}$$

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*,

$$\boldsymbol{o}_{\mathrm{noise}}^{[i]}(f,\tau) = [o_1^{[i]}(f,\tau), \ldots, o_{U-1}^{[i]}, 0, o_{U+1}^{[i]}(f,\tau), \ldots, o_K^{[i]}(f,\tau)]^{\mathrm{T}}, \tag{4}$$

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

$$\boldsymbol{z}^{[i]}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f)^{-1}\boldsymbol{o}_{\mathrm{noise}}^{[i]}(f,\tau). \tag{5}$$

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

$$x_k^{[i+1]}(f,\tau) = \begin{cases} \sqrt{|x_k^{[i]}(f,\tau)|^2 - \beta|z_k^{[i]}(f,\tau)|^2}\exp(j\arg(x_k^{[i]}(f,\tau))) \\ (\text{if } \ |x_k^{[i]}(f,\tau)|^2 > \beta|z_k^{[i]}(f,\tau)|^2) \\ \eta x_k^{[i]}(f,\tau) \quad (\text{otherwise}) \end{cases}, \tag{6}$$

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

Step(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 is expressed by

$$y(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{x}^{[*]}(f,\tau), \tag{7}$$

$$\boldsymbol{w}_{\mathrm{DS}}(f) = [w_1^{(\mathrm{DS})}(f), \ldots, w_K^{(\mathrm{DS})}(f)], \tag{8}$$

$$w_k^{(\mathrm{DS})}(f) = \frac{1}{K}\exp(-2j(f/N)f_{\mathrm{s}}d_k\sin\theta_U/c), \tag{9}$$

* "Unsupervised control of speech quality based on higher-order statistics in musical-noise-free blind speech extraction," by Yuka Hirano, Ryoichi Miyazaki, Hiroshi Saruwatari, and Satoshi Nakamura (Nara Institute of Science and Technology)

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 [1]. Moreover, $[\boldsymbol{A}]_{lj}$ represents the entry of $\boldsymbol{A}$ in the lth row and jth column.

This method can generate almost no musical noise even with increasing noise reduction, but instead always suffers from large speech distortion because of no justification of applying ICA to such signals nonlinearly distorted by SS. Since speech distortion cannot be measured without a clean *reference* speech signal, we should have to decide the number of iterations manually based on our auditory perception.

2.2 Unsupervised Measurement of Speech Distortion

2.2.1 Speech Kurtosis Ratio

As an evaluation of speech distortion, cepstral distortion is widely used. However, cepstral distortion cannot be measured without a clean reference speech signal. To solve this problem, speech kurtosis ratio was proposed as an unsupervised measurement of speech distortion. The speech kurtosis ratio is obtained as [5]

$$\text{kurtsis ratio}^{[\mathrm{s}]} = \mathrm{kurt}^{[\mathrm{s}]}_{\mathrm{proc}}/\mathrm{kurt}^{[\mathrm{s}]}_{\mathrm{org}}, \tag{10}$$

where $\mathrm{kurt}^{[\mathrm{s}]}_{\mathrm{proc}}$ is the speech kurtosis after signal processing and $\mathrm{kurt}^{[\mathrm{s}]}_{\mathrm{org}}$ is the speech kurtosis before signal processing. It is proved that the speech kurtosis ratio is strongly correlated to cepstral distortion [5].

2.2.2 Speech Kurtosis Estimation Method

The observed signal in the time-frequency domain, $x(f,\tau)$, is given by $x(f,\tau) = s(f,\tau) + n(f,\tau)$. Since the speech component is always contaminated with noise at every time-frequency grid, it is difficult to estimate the speech kurtosis via theoretical analysis. Therefore, we inversely calculate the kurtosis of the speech power spectrum in a data-driven manner, utilizing two observable statistics of the noisy speech signal $x(f,\tau)$ and noise signal $n(f,\tau)$ [5]. Note that the proposed speech kurtosis estimation is an unsupervised method because it requires no reference (clean) speech signals, unlike cepstral distortion.

To cope with the mathematical problem that the mixing of speech and noise is additive but generally their higher-order moments are not additive, we introduce the *cumulant*, which retains the additivity for additive variables. Meanwhile, in the transformation from a waveform to its power spectrum, the exponentiation operation is conducted. However, the cumulant does not have a straightforward relationship. In this case, we use the moment instead of the cumulant. Thus, we previously proposed the use of a *moment-cumulant transformation* [5]. In moment-cumulant transformation, The mth-order moment $\mu_m(x)$ can be written as

$$\mu_m(x) = \sum_{\pi(m)} \prod_{B\in\pi(m)} \kappa_{|B|}(x), \tag{11}$$

where $\pi(m)$ runs through the list of all partitions of a set of size m, $B \in \pi(m)$ means that B is one of the blocks into which the set is partitioned, and $|B|$ is the size of set B. In the same manner, the mth-order cumulant $\kappa_m(x)$ is given by

$$\kappa_m(x) = \sum_{\pi(m)} (-1)^{|\pi(m)|-1}(|\pi(m)|-1)! \prod_{B\in\pi(m)} \mu_{|B|}(x). \tag{12}$$

Hereafter, when we define complex-valued variables of the observed (noisy speech) signal, the original speech signal, and the noise signal as $(x_{\mathrm{R}} + ix_{\mathrm{I}})$, $(s_{\mathrm{R}} + is_{\mathrm{I}})$, and $(n_{\mathrm{R}} + in_{\mathrm{I}})$, in [5], the kurtosis of the speech power spectrum is estimated from moment-cumulant transformation, and the additivity of cumulants, as

$$\mathrm{kurt}_{\mathrm{speech}} = \frac{\mu_4(s_{\mathrm{R}}^2 + s_{\mathrm{I}}^2)}{\mu_2^2(s_{\mathrm{R}}^2 + s_{\mathrm{I}}^2)} = \frac{\mathcal{N}\left(\mu_m(x_{\mathrm{R}}), \mu_m(n_{\mathrm{R}})\right)}{\mathcal{D}\left(\mu_m(x_{\mathrm{R}}), \mu_m(n_{\mathrm{R}})\right)}, \tag{13}$$

where

$$\begin{aligned}
&\mathcal{D}\left(\mu_m(x_{\mathrm{r}}), \mu_m(x_{\mathrm{i}}), \mu_m(n_{\mathrm{r}}), \mu_m(n_{\mathrm{i}})\right) \\
&= \Big[\mu_4(x_{\mathrm{r}}) + \mu_4(x_{\mathrm{i}}) - \mu_4(n_{\mathrm{r}}) - \mu_4(n_{\mathrm{i}}) \\
&\quad + \{2\mu_2(x_{\mathrm{i}}) - 6\mu_2(n_{\mathrm{r}}) - 2\mu_2(n_{\mathrm{i}})\}\,\mu_2(x_{\mathrm{r}}) \\
&\quad + \{-2\mu_2(n_{\mathrm{r}}) - 6\mu_2(n_{\mathrm{i}})\}\,\mu_2(x_{\mathrm{i}}) \\
&\quad + 6\mu_2(n_{\mathrm{r}})^2 + 2\mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}}) + 6\mu_2(n_{\mathrm{i}})^2\Big]^2,
\end{aligned} \tag{14}$$

$$\begin{aligned}
&\mathcal{N}\left(\mu_m(x_{\mathrm{r}}), \mu_m(x_{\mathrm{i}}), \mu_m(n_{\mathrm{r}}), \mu_m(n_{\mathrm{i}})\right) \\
&= \mu_8(x_{\mathrm{r}}) + \mu_8(x_{\mathrm{i}}) - \mu_8(n_{\mathrm{r}}) - \mu_8(n_{\mathrm{i}}) \\
&\quad + \{4\mu_2(x_{\mathrm{i}}) - 28\mu_2(n_{\mathrm{r}}) - 4\mu_2(n_{\mathrm{i}})\}\,\mu_6(x_{\mathrm{r}}) \\
&\quad + \{4\mu_2(x_{\mathrm{r}}) - 4\mu_2(n_{\mathrm{r}}) - 28\mu_2(n_{\mathrm{i}})\}\,\mu_6(x_{\mathrm{i}}) \\
&\quad + \{-28\mu_2(x_{\mathrm{r}}) - 4\mu_2(x_{\mathrm{i}}) \\
&\qquad + 56\mu_2(n_{\mathrm{r}}) + 4\mu_2(n_{\mathrm{i}})\}\,\mu_6(n_{\mathrm{r}}) \\
&\quad + \{-4\mu_2(x_{\mathrm{r}}) - 28\mu_2(x_{\mathrm{i}}) \\
&\qquad + 4\mu_2(n_{\mathrm{r}}) + 56\mu_2(n_{\mathrm{i}})\}\,\mu_6(n_{\mathrm{i}}) \\
&\quad + \Big[6\mu_4(x_{\mathrm{i}}) - 70\mu_4(n_{\mathrm{r}}) - 6\mu_4(n_{\mathrm{i}}) + 420\mu_2(n_{\mathrm{r}})^2 \\
&\qquad + \{-60\mu_2(n_{\mathrm{r}}) - 36\mu_2(n_{\mathrm{i}})\}\,\mu_2(x_{\mathrm{i}}) \\
&\qquad + 60\mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}}) + 36\mu_2(n_{\mathrm{i}})^2\Big]\mu_4(x_{\mathrm{r}}) \\
&\quad + \Big[-6\mu_4(n_{\mathrm{r}}) - 70\mu_4(n_{\mathrm{i}}) + 36\mu_2(n_{\mathrm{r}})^2 \\
&\qquad - \{36\mu_2(n_{\mathrm{r}}) + 60\mu_2(n_{\mathrm{i}})\}\,\mu_2(x_{\mathrm{r}}) \\
&\qquad + 60\mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}}) + 420\mu_2(n_{\mathrm{i}})^2\Big]\mu_4(x_{\mathrm{i}}) \\
&\quad + 70\mu_4(n_{\mathrm{r}})^2 + \mu_4(n_{\mathrm{r}}) + 70\mu_4(n_{\mathrm{i}})^2 + \mu_4(n_{\mathrm{i}}) + \mu_2(x_{\mathrm{r}}) \\
&\quad - \Big\{360\mu_2(n_{\mathrm{r}})^3 + 216\mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}})^2 \\
&\qquad + 360\mu_2(n_{\mathrm{i}})^2\mu_2(n_{\mathrm{r}}) + 2520\mu_2(n_{\mathrm{i}})^3\Big\}\,\mu_2(x_{\mathrm{i}}) \\
&\quad + 2520\mu_2(n_{\mathrm{r}})^4 + 360\mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}})^3 \\
&\quad + 216\mu_2(n_{\mathrm{i}})^2\mu_2(n_{\mathrm{r}})^2 \\
&\quad + 360\mu_2(n_{\mathrm{i}})^3\mu_2(n_{\mathrm{r}}) + 2520\mu_2(n_{\mathrm{i}})^4.
\end{aligned} \tag{15}$$

This method can estimate speech kurtosis with high precision. However, if the noise signal estimated by the speech absence part is short, this method cannot work stably.

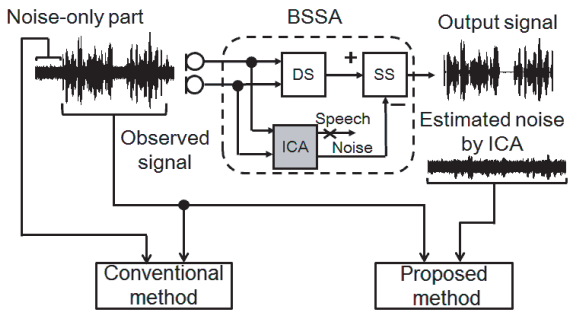

Fig. 2 Block diagram of ICA-based kurtosis estimation.

3 Proposed Method

3.1 Speech Kurtosis Estimation by ICA

The main problem in the conventional method is the low robustness in the estimation of higher-order statistics. In this method, it is necessary to calculate up to eighth-order statistics. Since such higher-order statistics are very sensitive to outliers, we cannot estimate them stably from limited few samples in speech absence part, causing considerable degradation of estimated speech kurtosis. To solve this problem, we propose a new speech kurtosis estimation method using the noise signal estimated by ICA instead of that in speech absence period (see Fig. 2). ICA can dynamically estimate nonstationary noise signal with great accuracy. By using the noise signal estimated by ICA, we can obtain enough the noise signal of length, and we can estimate higher-order statistics stably from a sufficient number of samples.

3.2 Speech-Kurtosis-Based Quality Control

Since iterative BSSA can generate no musical noise by keeping its higher-order statistics, it can be assumed that the statistical quantity of the residual noise signal in iterative BSSA does not change before/after processing. Therefore, we propose the control method of the number of iterations in the SS part using the speech kurtosis estimation method described in Sect. 3.1.

In this method, first, we calculate speech kurtosis before processing using the ICA-based noise estimation. Next, as for the estimate of speech kurtosis after processing, we can efficiently obtain it thanks to the assumption of fixed noise statistics in musical-noise-free properties.

The mth-order moment of noise signal n before signal processing given by

$$\mu_m(n) = \int_0^\infty n^m P(n) dn, \tag{16}$$

where $P(n)$ is the probability density function of a power-spectral-domain signal n. Since we consider that the noise statistics after signal processing are the same as that before signal processing, we calculated $\mu_m(n')$ after signal processing as

$$\begin{aligned}\mu_m(n') &= \int_0^\infty \alpha^m n^m P(n) dn \\ &= \alpha^m \mu_m(n),\end{aligned} \tag{17}$$

where $*'$ is a signal after signal processing, and α is a noise reduction rate, $\alpha = n'/n$, which can be estimated by speech absence part. We can estimate speech kurtosis after signal processing using $\mu_m(n')$ and (13) as,

$$\begin{aligned}&\mathcal{D}\left(\mu_m(x'_{\mathrm{r}}), \mu_m(x'_{\mathrm{i}}), \mu_m(n'_{\mathrm{r}}), \mu_m(n'_{\mathrm{i}})\right) \\ &= \Big[\mu_4(x'_{\mathrm{r}}) + \mu_4(x'_{\mathrm{i}}) - \alpha^4\{\mu_4(n_{\mathrm{r}}) + \mu_4(n_{\mathrm{i}})\} \\ &\quad + 2\left\{\mu_2(x'_{\mathrm{i}}) - 3\alpha^2\mu_2(n_{\mathrm{r}}) - \alpha^2\mu_2(n_{\mathrm{i}})\right\}\mu_2(x'_{\mathrm{r}}) \\ &\quad - 2\alpha^2\{\mu_2(n_{\mathrm{r}}) + 3\mu_2(n_{\mathrm{i}})\}\mu_2(x'_{\mathrm{i}}) \\ &\quad + 2\alpha^4\Big\{6\mu_2(n_{\mathrm{r}})^2 + \mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}}) + 3\mu_2(n_{\mathrm{i}})^2\Big\}\Big]^2,\end{aligned} \tag{18}$$

$$\begin{aligned}&\mathcal{N}\left(\mu_m(x'_{\mathrm{r}}), \mu_m(x'_{\mathrm{i}}), \mu_m(n'_{\mathrm{r}}), \mu_m(n'_{\mathrm{i}})\right) \\ &= \mu_8(x'_{\mathrm{r}}) + \mu_8(x'_{\mathrm{i}}) - \alpha^8\mu_8(n_{\mathrm{r}}) - \alpha^8\mu_8(n_{\mathrm{i}}) \\ &\quad + 4\left\{\mu_2(x'_{\mathrm{i}}) - 7\alpha^2\mu_2(n_{\mathrm{r}}) - \alpha^2\mu_2(n_{\mathrm{i}})\right\}\mu_6(x'_{\mathrm{r}}) \\ &\quad + 4\left\{\mu_2(x'_{\mathrm{r}}) - \alpha^2\mu_2(n_{\mathrm{r}}) - 7\alpha^2\mu_2(n_{\mathrm{i}})\right\}\mu_6(x'_{\mathrm{i}}) \\ &\quad + 4\alpha^6\left\{-7\mu_2(x'_{\mathrm{r}}) - \mu_2(x'_{\mathrm{i}}) + 14\alpha^2\mu_2(n_{\mathrm{r}}) + \alpha^2\mu_2(n_{\mathrm{i}})\right\}\mu_6(n_{\mathrm{r}}) \\ &\quad + 4\alpha^6\left\{-\mu_2(x'_{\mathrm{r}}) - 7\mu_2(x'_{\mathrm{i}}) + \alpha^2\mu_2(n_{\mathrm{r}}) + 14\alpha^2\mu_2(n_{\mathrm{i}})\right\}\mu_6(n_{\mathrm{i}}) \\ &\quad + \Big[6\mu_4(x'_{\mathrm{i}}) - 70\alpha^4\mu_4(n_{\mathrm{r}}) - 6\alpha^4\mu_4(n_{\mathrm{i}}) \\ &\qquad - 12\alpha^2\{5\mu_2(n_{\mathrm{r}}) + 3\mu_2(n_{\mathrm{i}})\}\mu_2(x'_{\mathrm{i}}) \\ &\qquad + 12\alpha^4\Big\{35\mu_2(n_{\mathrm{r}})^2 + 5\mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}}) + 3\mu_2(n_{\mathrm{i}})^2\Big\}\Big]\mu_4(x'_{\mathrm{r}}) \\ &\quad + \Big[-\alpha^4\{6\mu_4(n_{\mathrm{r}}) + 70\mu_4(n_{\mathrm{i}})\} \\ &\qquad - 12\alpha^2\{3\mu_2(n_{\mathrm{r}}) + 5\mu_2(n_{\mathrm{i}})\}\mu_2(x'_{\mathrm{r}}) \\ &\qquad + 12\alpha^4\Big\{3\mu_2(n_{\mathrm{r}})^2 + 35\mu_2(n_{\mathrm{i}})^2 \\ &\qquad\qquad + 5\mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}})\Big\}\Big]\mu_4(x'_{\mathrm{i}}) \\ &\quad + 70\alpha^8\Big\{\mu_4(n_{\mathrm{r}})^2 + \mu_4(n_{\mathrm{i}})^2\Big\} \\ &\quad + \alpha^4\left\{\mu_4(n_{\mathrm{r}}) + \mu_4(n_{\mathrm{i}})\right\} + \mu_2(x'_{\mathrm{r}}) \\ &\quad + 72\alpha^6\Big[-5\mu_2(n_{\mathrm{r}})^3 - 3\mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}})^2 \\ &\qquad - 5\mu_2(n_{\mathrm{i}})^2\mu_2(n_{\mathrm{r}}) - 35\mu_2(n_{\mathrm{i}})^3\Big]\mu_2(x'_{\mathrm{i}}) \\ &\quad + 72\alpha^8\Big[35\mu_2(n_{\mathrm{r}})^4 + 5\mu_2(n_{\mathrm{i}})\mu_2(n_{\mathrm{r}})^3 + 3\mu_2(n_{\mathrm{i}})^2\mu_2(n_{\mathrm{r}})^2 \\ &\qquad + 5\mu_2(n_{\mathrm{i}})^3\mu_2(n_{\mathrm{r}}) + 35\mu_2(n_{\mathrm{i}})^4\Big].\end{aligned} \tag{19}$$

This processing is advantageous because we can omit the re-estimation process of noise kurtosis that is difficult to estimate after nonlinear signal processing like SS. Finally, based on the above-mentioned estimates of speech kurtosis before/after processing, we predict speech distortion and control the maximum number of iterations in iterative BSSA.

4 Evaluation Experiments

4.1 Speech Kurtosis Estimation

To confirm the effectiveness of the proposed the speech kurtosis estimation method, we conducted objective evaluation experiments. In this experiment, the speech kurtosis estimated by the conventional method and the proposed method were compared. We used a two-element microphone array with an interelement spacing of 2.15 cm, and the direction of the target speech was set to be

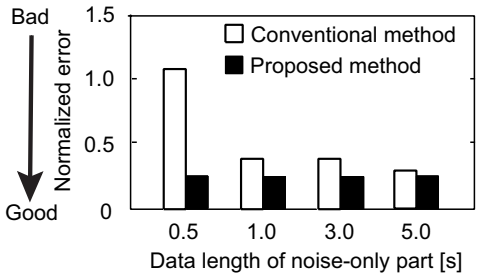

Fig. 3 Experimental result of speech kurtosis estimation.

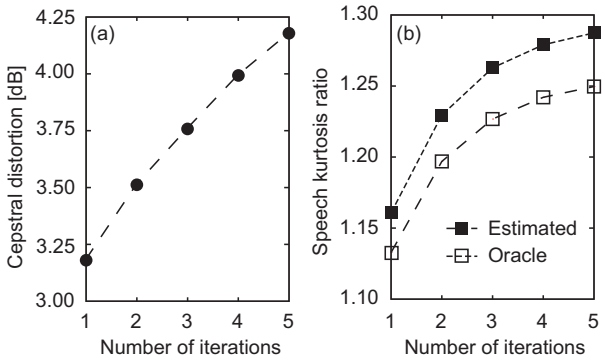

Fig. 4 (a) Relation between number of iterations and cepstral distortion, (b) Relation between number of iterations and speech kurtosis ratio in oracle or estimated.

normal to the array. The size of the experimental room was $4.2 \times 3.5 \times 3.0$ m^3 and the reverberation time was approximately 200 ms. All the signals used in this experiment were sampled at 16 kHz with 16-bit accuracy. The observed signal consisted of the target signal of one speaker (female) and one real-recorded diffuse noise (railway station noise) emitted from eight surrounding loudspeakers. The input SNR was set to 0 dB. The FFT size was 1024, and the frame shift length was 256. The length of speech absent part was set to 0.5, 1.0, 3.0 or 5.0 s. We calculated the normalized error of the estimates in the conventional and proposed methods and compared the accuracy of speech kurtosis estimation. The normalized error is defined as $e_{\mathrm{norm}} = |\mathrm{kurt}_{\mathrm{oracle}} - \mathrm{kurt}_{\mathrm{speech}}|/\mathrm{kurt}_{\mathrm{oracle}}$, where $\mathrm{kurt}_{\mathrm{oracle}}$ is the power spectral kurtosis of the clean speech signal and $\mathrm{kurt}_{\mathrm{speech}}$ is the estimate of the speech power spectral kurtosis.

The result is shown in Fig. 3. In the conventional method, the normalized error increases as the data length of the noise-only part decreases. In contrast, since the proposed method can use sufficient data length, the normalized error is lower in each case, meaning that the proposed method can stably estimate speech kurtosis with high accuracy compared with the conventional method. Therefore, we can confirm the validity of the proposed method.

4.2 Control in Iterative BSSA

We conducted objective evaluation experiments to confirm the validity of the proposed method of control of the number of iterations in iterative BSSA. In this experiment, we calculated cepstral distortion and estimated speech kurtosis of the output signal of iterative BSSA. The number of iterations in iterative BSSA was set to 1, 2, 3, 4 and 5, and the results of each case were compared. Thus, we evaluated whether or not estimated speech kurtosis was used as measure of speech distortion instead of cepstral distortion. We used a four-element microphone array with an interelement spacing of 2.15 cm. The input SNR was set to 5 dB. The noise reduction rate was set to 10 dB. We used the noise signal estimated by ICA as the noise signal before/after processing. Other experimental conditions are the same as those in the previous subsection.

The result is shown in Fig. 4. Figure 4 (a) shows a relation between the number of iterations and cepstral distortion for the extracted speech in iterative BSSA, and (b) shows a relation between the number of iterations and speech kurtosis ratio in oracle or estimated by the proposed method. In Fig. 4, both cepstral distortion and the estimated speech kurtosis ratio increase as the number of iterations increases. Thus, the speech kurtosis ratio is valid for an unsupervised measurement of speech distortion.

From this results, the number of iterations of iterative BSSA is controllable by limiting the value of the speech kurtosis ratio within an allowable value in human perception. Therefore, we can control the number of iterations with a constraint on sound quality degradation.

5 Conclusions

In this paper, we propose a new speech kurtosis estimation method using the noise signal estimated by ICA, and an automatic control of the number of iterations in iterative BSSA. Experimental evaluations confirmed the efficacy of the proposed methods.

Acknowledgements

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

References

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

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

[3] R. Miyazaki, H. Saruwatari, S. Nakamura, K. Shikano, " Toward musical-noise-free blind speech extraction: concept and its applications," *Proc. APSIPA*, 2013.

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

[5] R. Miyazaki, H. Saruwatari, R. Wakisaka, K. Shikano, T. Takatani, " Theoritical analysis of parametric blind spatial subtraction array and its application to speech recognition performance prediction, " *Proc. HSCMA*, pp.19–24, 2011.