PERFORMANCE IMPROVEMENT OF HIGHER-ORDER ICA USING LEARNING PERIOD DETECTION BASED ON CLOSED-FORM SECOND-ORDER ICA AND KURTOSIS

Yuuki Fujihara, Yu Takahashi, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano[†] *, Akira Tanaka*[‡]

[†]Nara Institute of Science and Technology, Ikoma, Nara, 630-0192, JAPAN
[‡]Hokkaido University, Kita-14, Nishi-9, Kita-ku, Sapporo, 060-0814, JAPAN

ABSTRACT

A hands-free system with conventional independent component analysis (ICA) should update the separation filter constantly to follow the hourly environment change. However, when the separation-filter is updated in the period where the user absent, ICA often yields an undesired separation filter. In this paper, we propose a novel algorithm of kurtosis-based voice activity detection (VAD) for an appropriate ICA optimization. In this algorithm, first, closed-form 2nd-order ICA (SO-ICA) is performed for providing the roughly separated signals, and based on the kurtosis of their signals, VAD is processed. Next, when the current time period is voiced block, the higher-order ICA's re-optimization is applied using the previously obtained SO-ICA's solution as the initial filter; this results in a fast and high convergence. The effectiveness of the proposed method is shown in a simulation experiment for blind spatial subtraction array with the proposed method.

Index Terms— Independent component analysis, blind source separation, voice activity detection, closed-form solution

1. INTRODUCTION

Recently, many studies of blind source separation (BSS) based on independent component analysis (ICA) have been conducted [1, 2, 3]. BSS is the unsupervised filtering approach taken to estimate original source signal using only information of mixed signals observed in each input channel. Owing to the attractive features of ICA-based BSS, this technique is applicable to creation of a noise-robust hands-free speech recognition and speech communication system.

The conventional ICA should update the separation filter constantly to follow the hourly environment change. In practical use, however, almost all the time periods do not contain a user's utterance. Therefore, the continuous separation-filter updating is a burden to the hands-free system. In addition, if we conduct the separation-filter updating in the time period in which a user does not exist, ICA often causes unstable behavior.

In this paper, we propose a method in which separation-filter updating is conducted only when a user exists. First, closed-form 2nd-order ICA (SO-ICA) [3] is performed for providing the roughly separated signals periodically. Next, based on the kurtosis of the roughly separated signals, voice activity detection (VAD) is processed. In the voice active time period, the higher-order ICA's re-optimization is applied using the previously obtained closed-form SO-ICA's solution as the initial filter; this results in a fast and high convergence. The effectiveness of the proposed method is revealed via performing a simulation experiment of blind spatial subtraction array (BSSA) [4] with the proposed method under noisy conditions.

This work was partly supported by MIC Strategic Information and Communications R&D Promotion Programme in Japan.

2. MIXING PROCESS

In this study, the number of microphones is K and the number of multiple sound sources is L. By applying the short-time discrete-time Fourier transform frame-wisely, we can express the observed signals, in which multiple source signals are linearly mixed, as follows in the time-frequency domain:

$$\boldsymbol{X}(f,t) = \boldsymbol{A}(f)\boldsymbol{S}(f,t), \tag{1}$$

where $\boldsymbol{X}(f,t) = [X_1(f,t),\ldots,X_K(f,t)]^{\mathrm{T}}$ is the observed signal vector, $\boldsymbol{S}(f,t) = [S_1(f,t),\ldots,S_L(f,t)]^{\mathrm{T}}$ is the source signal vector. Also, $\boldsymbol{A}(f)$ is the mixing matrix, which is complex-valued because we introduce a model to deal with the relative time delays among the microphones and room reverberations.

3. CONVENTIONAL METHODS

3.1. Frequency-domain ICA-based BSS

3.1.1. Separation process

In frequency-domain ICA-based BSS, we perform signal separation using the complex-valued unmixing matrix $\boldsymbol{W}(f)$, so that the L time-series outputs $\boldsymbol{Y}(f,t) = [Y_1(f,t),\ldots,Y_L(f,t)]^{\mathrm{T}}$ become mutually independent; this procedure can be given as

$$\boldsymbol{Y}(f,t) = \boldsymbol{W}(f)\boldsymbol{X}(f,t). \tag{2}$$

Various ICA methods for optimizing $\boldsymbol{W}(f)$ have been proposed, and these are classified into nonclosed-form and closed-form ICAs.

3.1.2. Nonclosed-form higher-order ICA (HO-ICA)

In the nonclosed-form HO-ICA, the optimal $\boldsymbol{W}(f)$ is obtained by the following iterative equation:

$$\boldsymbol{W}_{\mathrm{HO}}^{[m+1]}(f) = \mu\left[\boldsymbol{I} - \left\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\right\rangle_t\right]\cdot \boldsymbol{W}_{\mathrm{HO}}^{[m]}(f) + \boldsymbol{W}_{\mathrm{HO}}^{[m]}(f), \tag{3}$$

where $\boldsymbol{X}^{\mathrm{H}}$ denotes hermitian transpose of matrix $\boldsymbol{X}$, μ is the step-size parameter, $\boldsymbol{I}$ is an identity matrix, $[m]$ is used to express the value of the m-th step in the iteration, $\langle\cdot\rangle_t$ denotes a time-averaging operator, and $\boldsymbol{\Phi}(\boldsymbol{Y}(f,t))$ is the appropriate nonlinear vector function [1].

3.1.3. Closed-form 2nd-order ICA

This subsection briefly describes the overview of signal processing in the closed-form SO-ICA. First, we obtain the correlation matrices with different time periods as

$$\boldsymbol{R}_{t_i}(f) = \left\langle \boldsymbol{X}(f,t)\boldsymbol{X}^{\mathrm{T}}(f,t)\right\rangle_{t\in t_i}, \tag{4}$$

where $\langle\cdot\rangle_{t\in t_i}$ denotes the time-averaging operator over specific time duration t_i, and i $(= 1,2,\ldots)$ represents an index of time-averaging block. Next, we apply singular value decomposition (SVD) to a superposition of $\boldsymbol{R}_{t_i}(f)$, which is represented as

$$\sum_i \boldsymbol{R}_{t_i}(f) = \boldsymbol{U}(f)\mathrm{diag}(\lambda_1,\ldots,\lambda_K)\boldsymbol{U}^{\mathrm{H}}(f), \tag{5}$$

where λ_k $(k = 1, \ldots, K)$ are the eigenvalues, $\mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ denotes the diagonal matrix that includes the eigenvalues, and $\boldsymbol{U}(f)$ is the matrix consisting of the eigenvectors. Next, we obtain a full-rank decomposition for pseudo-inverse of $\sum_i \boldsymbol{R}_{t_i}(f)$ as follows

$$\left[\sum_i \boldsymbol{R}_{t_i}(f)\right]^{+} = \boldsymbol{L}(f)\boldsymbol{L}^{\mathrm{H}}(f), \tag{6}$$

$$\boldsymbol{L}(f) = \boldsymbol{U}(f)\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_K}}\right). \tag{7}$$

If the covariance of the sources $\boldsymbol{S}(f,t)$ in t_i is negligible, it can be proved that every $\boldsymbol{L}^{\mathrm{H}}(f)\boldsymbol{R}_{t_i}(f)\boldsymbol{L}(f)$ for any i shares the same eigenvectors, and this is given via SVD form as

$$\boldsymbol{L}^{\mathrm{H}}(f)\boldsymbol{R}_{t_i}(f)\boldsymbol{L}(f) = \boldsymbol{T}(f)\mathrm{diag}(\sigma_1(t_i), \sigma_2(t_i), \ldots)\boldsymbol{T}^{\mathrm{H}}(f), \tag{8}$$

where $\sigma_k(t_i)$ are the eigenvalues for a specific time block t_i, and $\boldsymbol{T}(f)$ denotes the matrix consisting of shared eigenvectors that are *independent* of time-block index i. Therefore, for any i, the simultaneous diagonalization of $\boldsymbol{R}_{t_i}(f)$ can be achieved as follows:

$$\boldsymbol{T}^{\mathrm{H}}(f)\boldsymbol{L}^{\mathrm{H}}(f)\boldsymbol{R}_{t_i}(f)\boldsymbol{L}(f)\boldsymbol{T}(f) = \mathrm{diag}(\sigma_1(t_i), \sigma_2(t_i), \ldots), \tag{9}$$

and this means that the optimal separation filter matrix in the 2nd-order sense is given by

$$\boldsymbol{W}_{\mathrm{SO}}(f) = (\boldsymbol{L}(f)\boldsymbol{T}(f))^{\mathrm{H}}. \tag{10}$$

Computational cost in the closed-form SO-ICA is very small. In fact, the whole computational cost in the closed-form solution is almost the same as those for 1 or 2 iterations in the nonclosed-form HO-ICA [5]. However, the separation performance of the closed-form SO-ICA is inferior to that of the nonclosed-form HO-ICA. Also, the closed-form SO-ICA provides a good initial filter, and the nonclosed-form HO-ICA can update the separation filters from the advantageous status. This enable us to reduce the computational complexities without deteriorating the separation performance [5].

3.2. Blind spatial subtraction array [4]

In a hands-free system in a real environment, it is required to extract a target speech and reduce noises that cannot be regarded as point sources. Although the conventional ICA-based BSS could work especially in point source mixing, it is difficult to apply ICA to non-point source noise reduction. BSSA has been proposed to extract a target speech in such a case. In BSSA, ICA is partly utilized as a noise estimator because of the fact that ICA is proficient in noise estimation rather than target estimation [4]. BSSA consists of two paths: a delay-and-sum array based primary path as the target speech enhancing part, and an ICA-based reference path as the noise estimation part (see Fig. 1). Based on the spectral subtraction method, the BSSA's output $Y_{\mathrm{BSSA}}(f,t)$ can be given by

$$Y_{\mathrm{BSSA}}(f,t) = \begin{cases} \left\{|Y_{\mathrm{DS}}(f,t)|^2 - \alpha \cdot |Z(f,t)|^2\right\}^{\frac{1}{2}} \\ \qquad (\text{if } |Y_{\mathrm{DS}}(f,t)|^2 - \alpha \cdot |Z(f,t)|^2 \geq 0), \\ \beta \cdot |Y_{\mathrm{DS}}(f,t)| \quad (\text{otherwise}), \end{cases} \tag{11}$$

where $Y_{\mathrm{DS}}(f,t)$ is the output signal from the primary path, $Z(f,t)$ is the output signal from the reference path, α represents over-subtraction parameter, and β denotes the flooring parameter.

4. PROPOSED METHOD

4.1. Motivation and strategy

In an actual environment, a hands-free system with ICA should update the separation filter constantly to follow the rapid environmental change. This also holds for a hands-free system with BSSA utilizing ICA as a noise estimator. However, speech application, e.g., hands-free spoken-oriented guidance, is confronted with almost all the time periods where the user is absent and only noise exists. In such a time period, the desired separation filter cannot be optimized and, in fact, an unstable separation filter appears. For these reasons, the separation filter should be optimized only when the user exists. Therefore, it is indispensable to detect the time periods where a user exists, i.e., voice active time periods. Here, VAD should work correctly under noisy conditions. In addition, when the current time period becomes a voice active period just after a noise-only period, we are requested to rebuild the separation filter as fast as possible.

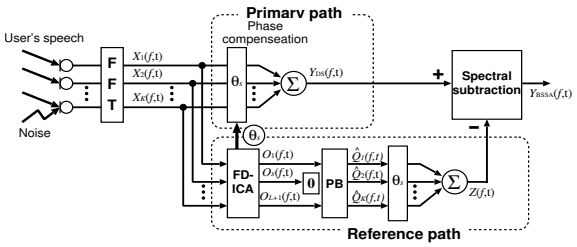

Fig. 1. Block diagram of BSSA.

For these two requirements, we newly propose the following method. In this method, the closed-form SO-ICA is performed for providing the roughly separated signals and enhances the speech signal. Thus VAD is processed correctly under noisy conditions. Also, when the current time period becomes a voice active period just after a noise-only period, nonclosed-form HO-ICA's re-optimization is performed with using the closed-form SO-ICA's solution as the initial filter. As a result, we rebuild the separation filter quickly.

4.2. Kurtosis-based voice activity detection

Kurtosis is an indicator that represents non-Gaussianity of a random variable. As probability density function (PDF) of a signal diverges from Gaussian distribution, this kurtosis is apart from zero. The kurtosis of a signal $y(t)$ is defined by

$$\mathrm{kurt}(y(t)) = \left\langle y^4(t)\right\rangle_t \left[\left\langle y^2(t)\right\rangle_t\right]^{-2} - 3. \tag{12}$$

In our application, the noise signal's kurtosis is assumed to be nearly zero. This is due to the fact that PDF of mixed signal of various kinds of noise becomes Gaussian distribution based on the central limit theorem. In contrast, the speech signal's PDF is expressed by Laplace distribution, and this kurtosis tends to be higher than the noise signal's kurtosis [6].

In the proposed method, first, the closed-form SO-ICA is performed for providing the roughly separated signals periodically. Next, based on kurtosis of roughly separated signals, VAD is processed. In the time period where both target speech and noise exist, the roughly separated signals provided by the closed-form SO-ICA become a speech-enhanced signal and noise-estimated signal. In general, the speech signal's kurtosis is different from the noise signal's kurtosis owing to difference of distribution shapes of their PDFs. Therefore, a difference among kurtosis of output signals is caused. Otherwise, when only noise exists, both roughly separated signals provided by the closed-form SO-ICA are noise-estimated signals, and their kurtosis values are almost the same. Based on this difference, the proposed method determines whether the user exists or not. In the proposed method, the evaluation score that shows the difference among kurtosis of roughly separated signals is defined. In the case of $L = 2$, the evaluation score is written as

$$K_s(b) = \left\langle \left|\mathrm{kurt}(\mathrm{real}(Y_1^{(\mathrm{SO})}(f,t,b))) - \mathrm{kurt}(\mathrm{real}(Y_2^{(\mathrm{SO})}(f,t,b)))\right|\right\rangle_{f(25)}, \tag{13}$$

where b $(= 1, 2, \ldots)$ denotes the index of time durations, $Y_l^{(\mathrm{SO})}(f,t,b)$ denotes the roughly separated signal provided by the closed-form SO-ICA in time duration b, l is the channel index, $\langle\cdot\rangle_{f(25)}$ denotes a frequency-averaging operator of the top 25 kurtosis difference, and $\mathrm{real}(\cdot)$ expresses the real part of $\cdot$. This evaluation score can avoid

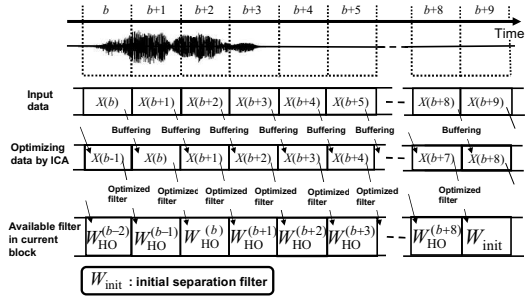

Fig. 2. Algorithm of updating separation filter in conventional method.

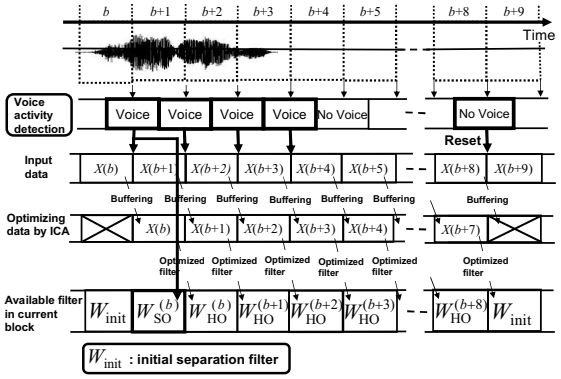

Fig. 3. Algorithm of updating separation filter in proposed method.

the permutation effect in VAD because this score is based on the difference among kurtosis of roughly separated signals, which is calculated in each frequency bin (we call kurtosis difference). Also, since speech signals are often sparse, some frequency sub-bands of speech signal have few powers. In this score, however, we average the top 25 of kurtosis difference. Thus we can extract only frequency sub-bands where speech signals have a lot of power.

4.3. Real-time processing

The proposed method should be worked in real-time for practical use. Thus, we construct real-time algorithms of the conventional ICA and the proposed method (see Figs. 2 and 3).

In the conventional real-time algorithm, nonclosed-form HO-ICA's separation filter is updated constantly without VAD. In this paper, the length of the one time duration is set to be 3 seconds. First, the input data of a time duration b is buffered. Secondly, the nonclosed-form HO-ICA's separation filter is optimized using the buffering data in the next time duration $b+1$. The optimized separation filter is applied to the data in the next time duration $b+2$. This is due to the fact that the filter update in the nonclosed-form HO-ICA requires substantial computational complexities and cannot provide the optimal separation filter for current time period data.

In the proposed method, the nonclosed-form HO-ICA optimizes the separation filter while a user exists. The optimization process is the same as the conventional method.

4.4. Process flow of proposed method

Figure 4 shows process flow of proposed method. The details of the proposed method is described below.

[Step 1: Observed signal segmentation]
Split up the observed signal into fixed length segments. $\boldsymbol{X}(f,t,b)$ is the split signals.

[Step 2: Source separation by closed-form SO-ICA]
Separate the observed signal via closed-form SO-ICA in the each

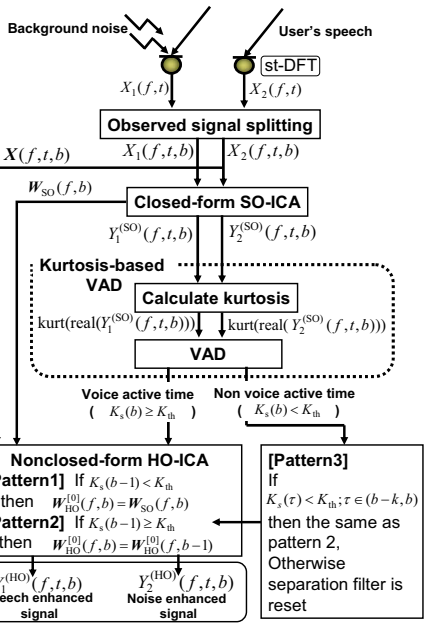

Fig. 4. Process flow of proposed method (case of $K = L = 2$).

time duration by **Step 1**. This can be represented by

$$\boldsymbol{Y}_{\mathrm{SO}}(f,t,b) = \boldsymbol{W}_{\mathrm{SO}}(f,b)\boldsymbol{X}(f,t,b), \tag{14}$$

where $\boldsymbol{W}_{\mathrm{SO}}(f,b)$ is the closed-form SO-ICA's separation filter optimized for time duration b, and $\boldsymbol{Y}_{\mathrm{SO}}(f,t,b)$ denotes the roughly separated signals provided by the closed-form SO-ICA in time duration b.

[Step 3: Calculation of kurtosis]
Calculate the evaluation score $K_s(b)$ of the roughly separated signals obtained by **Step 2**.

[Step 4: To determine whether voice is active or not]
Determine whether voice is active or not based on $K_s(b)$ calculated in **Step 3** as follows:

$$\begin{cases} \text{Voice active time} & (\text{if } K_s(b) \geq K_{\mathrm{th}}), \\ \text{Non voice active time} & (\text{otherwise}). \end{cases} \tag{15}$$

[Step 5: Optimization of the nonclosed-form HO-ICA's separation filter]
Based on the result of **Step 4**, the nonclosed-form HO-ICA's optimization is processed only when current time period is voiced block. This optimization depends on K_{th} and $K_s(b)$.

[Pattern 1: $K_s(b) \geq K_{\mathbf{th}}, K_s(b-1) < K_{\mathbf{th}}$] In this case, block b is the time duration where the user start to talk. Thus, in the time duration $b+1$, the nonclosed-form HO-ICA's re-optimization is performed with using the obtained closed-form SO-ICA's solution in **Step 2** as the initial filter: this result in a fast convergence [5]. This can be given by

$$\boldsymbol{W}^{[0]}_{\mathrm{HO}}(f,b) = \boldsymbol{W}_{\mathrm{SO}}(f,b), \tag{16}$$

$$\boldsymbol{W}^{[m+1]}_{\mathrm{HO}}(f,b) = \mu\left[\boldsymbol{I} - \left\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t,b))\boldsymbol{Y}^{\mathrm{H}}(f,t,b)\right\rangle_t\right]\cdot \boldsymbol{W}^{[m]}_{\mathrm{HO}}(f,b) + \boldsymbol{W}^{[m]}_{\mathrm{HO}}(f,b). \tag{17}$$

The optimized separation filter $\boldsymbol{W}_{\mathrm{HO}}(f,b)$ is applied to the data in the time duration $b+2$.

$$\boldsymbol{Y}(f,t,b+2) = \boldsymbol{W}_{\mathrm{HO}}(f,b)\boldsymbol{X}(f,t,b+2). \tag{18}$$

In the time duration b, the optimal separation filter is not provided. Thus, the separation is performed by the initial filter given in ad-

vance.

$$\boldsymbol{Y}(f,t,b) = \boldsymbol{W}_{\text{init}}(f)\boldsymbol{X}(f,t,b). \tag{19}$$

[Pattern 2: $K_s(b) \geq K_{\text{th}}, K_s(b-1) \geq K_{\text{th}}$**]** This is speech continued case. In this case, the nonclosed-form HO-ICA's optimization is continued. Thus, the initial filter for HO-ICA are given by

$$\boldsymbol{W}^{[0]}_{\text{HO}}(f,b) = \boldsymbol{W}_{\text{HO}}(f,b-1). \tag{20}$$

In the time duration $b+1$, the nonclosed-form HO-ICA's re-optimization is performed the same as Eq.(17). The optimized separation filter $\boldsymbol{W}_{\text{HO}}(f,b)$ is applied to the data in the time duration $b+2$ the same as Eq.(18). If the case of $K_s(b-2) < K_{\text{th}}$, the optimal separation filter is not provided in the current duration b. Thus, the separation is performed by the obtained closed-from SO-ICA's solution in the time duration $b-1$.

$$\begin{cases} \boldsymbol{Y}(f,t,b) = \boldsymbol{W}_{\text{SO}}(f,b-1)\boldsymbol{X}(f,t,b) & (\text{if } K_s(b-2) < K_{\text{th}}), \\ \boldsymbol{Y}(f,t,b) = \boldsymbol{W}_{\text{HO}}(f,b-2)\boldsymbol{X}(f,t,b) & (\text{otherwise}). \end{cases} \tag{21}$$

[Pattern 3: $K_s(b) < K_{\text{th}}$**]** In **Pattern 3**, time duration b is not voiced block. It can be consider the following case: $K_s(\tau) \geq K_{\text{th}}; \tau \in (b-k, b)$ or not. In the first case, it remain possible that the same speaker use this application in the moment time durations after that. Thus, the nonclosed-form HO-ICA's optimization is continued the same as **Pattern 2**. In the latter case, the nonclosed-from HO-ICA's optimization is not conducted, and the current separation filter is reset.

5. EXPERIMENTS AND RESULT

5.1. Experimental setup

We assess the effectiveness of the proposed method by performing a simulation experiment for BSSA with the proposed method. We carry out experiments in a real reverberant room illustrated in Fig. 5. In this experiment, we use the following 8 kHz sampled signals: speech signals, which are assumed to arrive from different directions, (θ_1, θ_2), are outputted from loudspeakers in the different time period, and the noise is actually-recorded railway-station noise from 36 loudspeakers. We use speech signals by 2 males and 2 females. Thus, 12 combinations of speakers are used in the experiment. The source signals are mixtures with 210-second noise signal and 15-second two target signals, which simulate the user existence time period and non-existence time period. The input SNR of test data is set to 6 dB. The DFT size is 1024, and frame shift length is 256. The block size for calculation of each $\boldsymbol{R}_{t_i}(f)$ in the closed-form SO-ICA is set to 1 second. The initial filter is the null beamformer [2]. In this experiment, real-time algorithms in Sect. 4.3 are used. In each time period, the number of iterations in the nonclosed-form HO-ICA part is 30. Also, $\alpha = 2.0$, $\beta = 0.03$ was used for BSSA. We use noise reduction rate (NRR), which is defined as the output SNR in dB minus the input SNR in dB, for evaluation [2].

5.2. Experiments result

Figure 6 shows the NRR that corresponds to noise estimation by the nonclosed-form HO-ICA in BSSA with the proposed method or the conventional method in each time period. We can see that NRR of the proposed method overtakes that of the conventional method, especially for early time period of the speech. Therefore, in the time period where a user does not exist, the proposed method prevents from degrading the separation filter by VAD. When the current time period becomes voiced block, the nonclosed-form HO-ICA's re-optimization is performed using the previously obtained SO-ICA's solution as the initial filter. Also, we can confirm the same pattern in Fig. 7, which shows the NRR of BSSA with the proposed method or the conventional method in each time period. Improvements of more than 6 dB are achieved in the early parts of voiced periods, which is an obvious and significant improvement for human hearing. These results reveal the effectiveness of the proposed method in an actual environment.

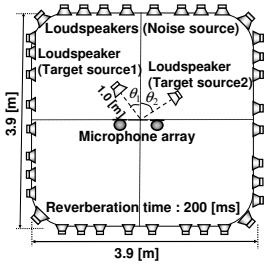

Fig. 5. A layout of reverberant room used in our experiment.

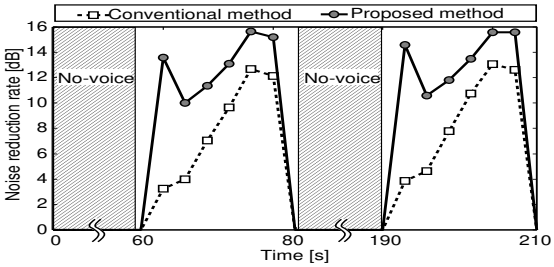

Fig. 6. NRR by ICA (reference path) in each time period.

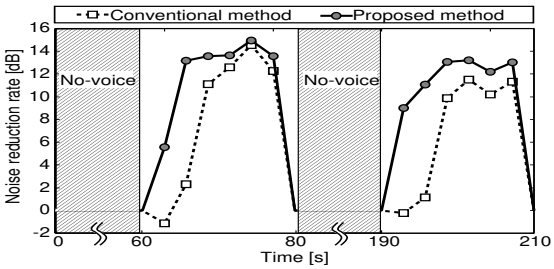

Fig. 7. NRR by BSSA in each time period.

6. CONCLUSIONS

In this paper, we propose a novel algorithm of VAD based on kurtosis and closed-form SO-ICA. This is a method in which separation-filter updating is conducted only when a user exists. Experimental results reveal the effectiveness of the proposed method by performing a simulation experiment for BSSA with the proposed method.

7. REFERENCES

[1] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol.22, pp.21–34, 1998.

[2] H. Saruwatari, et al., "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Speech & Audio Process.*, vol.14, pp.666–678, 2006.

[3] A. Tanaka, et al., "Theoretical foundations of second-order-statistics-based blind source separation for non-stationary sources," *Proc. ICASSP*, pp.III-600–III-603, 2006.

[4] Y. Takahashi, et al., "Blind spatial subtraction array with independent component analysis for hands-free speech recognition," *Proc. IWAENC*, 2006.

[5] K. Tachibana, et al., "Efficient blind source separation combining closed-form second-order ICA and nonclosed-form higher-order ICA," *Proc. ICASSP*, pp.I-45–I-48, 2007.

[6] T-W. Lee, *Independent Component Analysis,* Norwell, MA: Kluwer Academic, 1998.