ICIC Express Letters
Part B: Applications
Volume **8**, Number **12**, December **2017**
 ISSN 2185-2766
pp. **1603–1610**

TARGET HUMAN SPEECH EXTRACTION METHOD BASED ON SILENT INTERVAL DETECTION

Takaaki Ishibashi[1], Kana Higuchi[2] and Chiharu Okuma[3]

[1]Department of Information, Communication and Electronic Engineering
[2]Advanced Course of Electronics and Information Systems Engineering
[3]Department of Human-Oriented Information Systems Engineering
National Institute of Technology, Kumamoto College
2659-2, Suya, Koshi, Kumamoto 861-1102, Japan
ishibashi@kumamoto-nct.ac.jp

Received June 2017; accepted September 2017

Abstract. *In this paper, we propose a target human speech extraction method for acoustic signals. The proposed method can estimate the target speech signals without noise from observed mixture signals. The method estimates a rotation angle of a distribution of observed mixture signals. Based on the estimated angle, our extraction method estimates the speech by rotating the distribution. The estimated angle is dependent on the silent interval included in the human speech. In a case of short-time processing, some rotation angle cannot be estimated because some observed mixture signals do not have a silent interval in a short-time frame. Therefore, a silent interval detection method is proposed for estimating the presence or absence of the silent interval by calculating the correlation or the kurtosis of the distribution. Using the proposed method, we can estimate the human speech in all short-time frames.*
Keywords: Blind source separation, Noise reduction, Silent interval, Speech signal processing, Target speech extraction

1. **Introduction.** We are living in the environment where there are a variety of sounds. And we can recognize the target speech from among a lot of sounds. However, a machine is hard to recognize their target sound in such noisy places. Blind source separation (BSS) is a method for estimating the source signals from observed mixture signals without the information about the source and the transfer functions [1]. Recently, a lot of methods of separating noises and target sounds have been proposed by researchers [2, 3]. Many noise reduction typical methods using independent component analysis (ICA) have been proposed [4, 5]. ICA is one of the methods based on the BSS. ICA can separate unknown sources from their mixtures in the condition which the sources are statistically independent of each other. However, ICA takes a long time to separate the source signals from mixture signals. Therefore, ICA is not good at a real-time processing. Since ICA is a separation algorithm, another target extraction algorithm is needed.

In this paper, we propose a new BSS method for speech signals. Our method is divided into a short-time frame speech signal in order to perform a short-time processing, and the method can extract the target speaker speech. The processes of the proposed method are as follows. Firstly, we create a joint distribution and a histogram using the mixture signals observed by two microphones. Secondly, a rotation angle of the distribution is estimated. Our method needs the silent interval included in a human speech in order to estimate the angle. In the case of a short-time frame, some rotation angle cannot be estimated because some observed mixture signals in the frame do not have a silent interval. When the sound in the short-time frame does not have a silent interval, the value of the correlation of the distribution and that of kurtosis of the histogram is low. Therefore, we estimate the presence or absence of the silent interval by calculating the

correlation or the kurtosis. Using this fact, we propose an estimation method for the human speech. The proposed method outputs only the target speech. It means that the method can separate and extract the target human speech in all short-time frames.

In order to verify our proposals, several simulations were carried out. From the simulation results, the proposed method can extract the target speech in all frames. The method works even in the case of the frame length in 0.01 seconds. Therefore, it can be expected that the target speech is estimated when the sound source will be moved or the sound environment will be changed.

The rest of this paper is organized as follows. In Section 2, we describe the outline of BSS and our basic separation principle. In Section 3, we propose our BSS method using silent interval detection. In Section 4, we evaluate the estimation performance of the proposed method. Finally, the paper is concluded in Section 5.

2. Blind Source Separation.

When source signals $s_n(t)$ $(n = 1, 2, \ldots, N)$ are observed by sensors, the observed mixture signals $x_m(t)$ $(m = 1, 2, \ldots, M)$ are expressed as

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t) \tag{1}$$

where a_{mn} denote unknown mixing parameters, and N and M denote the number of the sources and the sensors, respectively. The estimated signals $y_n(t)$ for the sources are expressed as

$$y_n(t) = \sum_{m=1}^{M} w_{nm} x_m(t) \tag{2}$$

where w_{nm} denote estimated separating parameters. In order to estimate w_{nm}, many methods have been proposed, e.g., Makino et al. [1], Hyvärinen et al. [4], Cichocki and Amari [5].

In order to extract the target human speech, we have already proposed a rotation BSS [6]. The rotation BSS is based on the rotation of a distribution of observed signals. Consider that a human utterance and noise exist as shown in Figure 1(a), and the mixture signals are observed by two microphones as shown in Figure 1(b).

Using the mixture signals, their joint distribution is shown in Figure 2(a), where the horizontal and the vertical axes are denoted by the amplitude of $x_1(t)$ and $x_2(t)$, respectively. From the figure, we recognize a straight line in the distribution. For clarifying the linear components, directions $\phi(t)$ of distribution are calculated as

$$\phi(t) = \tan^{-1} \frac{x_2(t)}{x_1(t)} \tag{3}$$

Using all $\phi(t)$, its histogram as shown in Figure 2(b) means that the components of the joint distribution are concentrated on one direction. The reason is that since the human speech has a silent interval, it becomes only noise in the silent interval.

From the above discussion, we define θ calculated as the mode value (the most frequent value) of $\phi(t)$ as

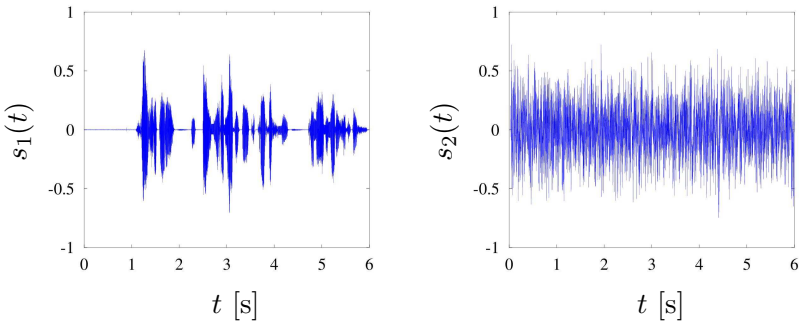
$$\theta = \arg\max_{\phi(t)} \mathrm{hist}(\phi(t)) \tag{4}$$

Based on the θ, the separated signals are rotated by the rotation matrix as follows.

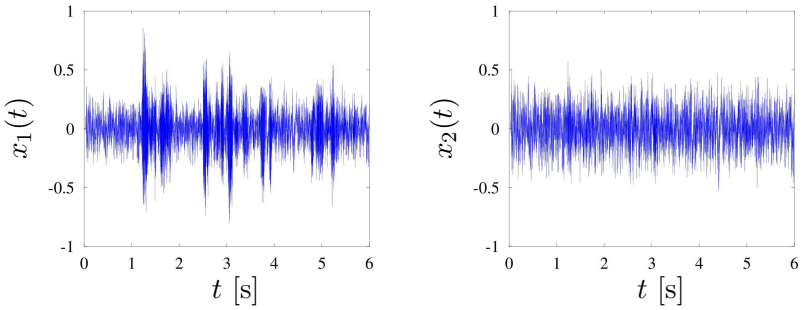
$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{5}$$

In order to extract the target human speech $y(t)$, we calculate as

$$y(t) = x_2(t)\cos\theta - x_1(t)\sin\theta \tag{6}$$

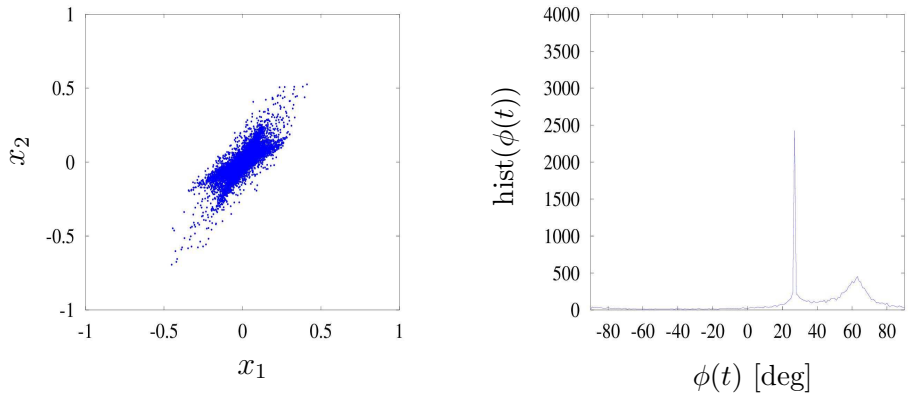


(a) Human speech and car noise



(b) Observed signals by two microphones

FIGURE 1. Source signals and mixture signals



(a) Joint distribution (b) Histogram

FIGURE 2. Joint distribution and its directional histogram of mixture signals

because $y_1(t)$ and $y_2(t)$ in Equation (5) are estimated of the noise and the target signal, respectively.

For a real-time processing, we introduce a short-time frame $x_{m,l}(t)$ as

$$x_{m,l}(t) = [x_m((l-1)T), x_m((l-1)T+1), \ldots, x_m(lT-1)] \tag{7}$$

where l denotes an index of the frames and T denotes the number of data points in the frame.

In the case of the 6 seconds mixture signals as shown in Figure 1(b), distributions in the 1 second frame are shown in Figure 3. From the distributions, we calculate the directions $\theta_l(t)$ of l frames as follows.

$$\phi_l(t) = \tan^{-1} \frac{x_{2,l}(t)}{x_{1,l}(t)} \tag{8}$$

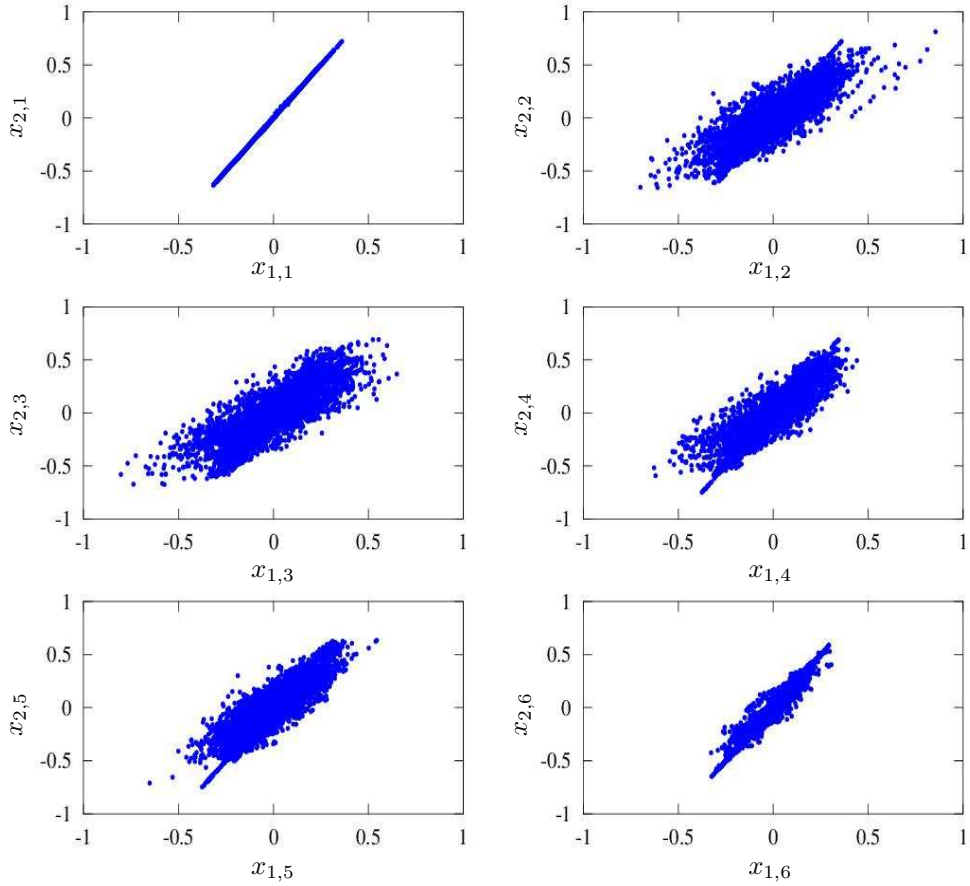


FIGURE 3. Distributions in each frame

Based on the $\phi_l(t)$, the histograms in each frame are shown in Figure 4. The mode value in each frame θ_l is calculated as follows.

$$\theta_l = \arg\max_{\phi_l(t)} \mathrm{hist}(\phi_l(t)) \tag{9}$$

In these ways, the rotation BSS in short-time frames are proposed as follows [7].

$$y_l(t) = \cos\theta_l x_{2,l}(t) - \sin\theta_l x_{1,l}(t) \tag{10}$$

The method can separate and extract the human speech. Therefore, the permutation problem does not occur. However, when the frame length is shortened, there is a problem that the method cannot separate because there is no silent interval in the frame. This problem is solved in Section 3.

3. **Silent Interval Detection.** The rotation BSS can separate and extract the target human speech in the condition which the human speech has a silent interval. However, when the sound data in the short-time frame does not have a silent interval, the linear component of the distribution or the peak of the histogram is not clear. Therefore, we propose a new BSS using a silent interval detection method.

Figure 3 shows the distributions in each frame. From these figures, it is found that the frames with silent interval have clearly straight line. Therefore, we propose a silent interval detection method using a correlation coefficient r_l as follows.

$$r_l = \frac{\sum_{t=0}^{T-1}\left(x_{1,l}(t) - \overline{x_{1,l}(t)}\right)\left(x_{2,l}(t) - \overline{x_{2,l}(t)}\right)}{\sqrt{\sum_{t=0}^{T-1}\left(x_{1,l}(t) - \overline{x_{1,l}(t)}\right)^2}\sqrt{\sum_{t=0}^{T-1}\left(x_{2,l}(t) - \overline{x_{2,l}(t)}\right)^2}} \tag{11}$$

In the case of $r_l \simeq 1$, the rotation BSS functions well. In the case of otherwise, the BSS selects a rotation direction in the near past frame.

Using the histograms as shown in Figure 4, we also propose another detection method. When the sound data has a silent interval, a peak is recognized in the histogram. Therefore, a silent interval detection method using a kurtosis k_l in the l frame is proposed as follows.

$$k_l = \frac{\frac{1}{T-1}\sum_{t=0}^{T-1}\left(\phi_l(t) - \overline{\phi_l(t)}\right)^4}{\sqrt{\sum_{t=0}^{T-1}\left(\phi_l(t) - \overline{\phi_l(t)}\right)^2}^4} \tag{12}$$

Similar to the method based on the correlation, the rotation BSS selects a rotation direction in the near past in the case that the k_l is low value.

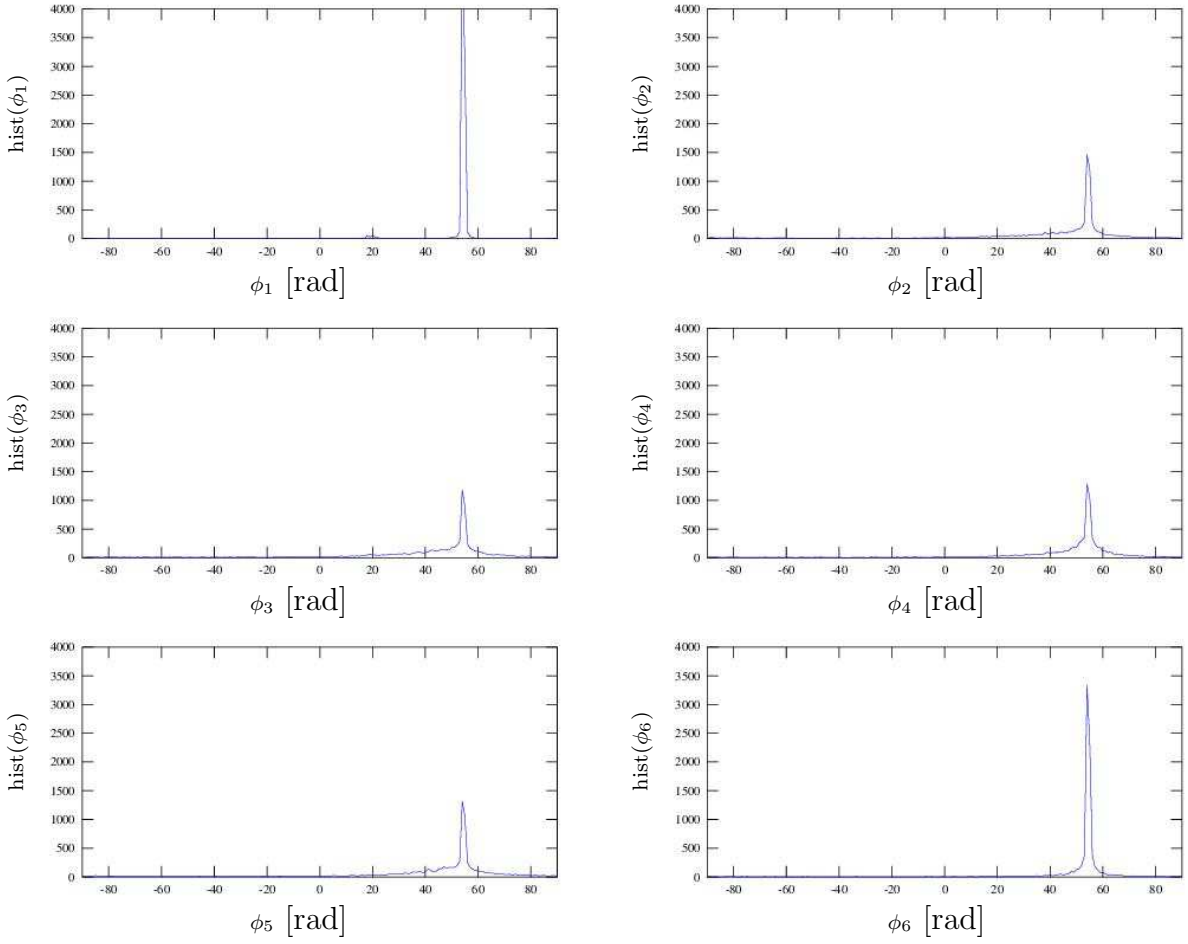
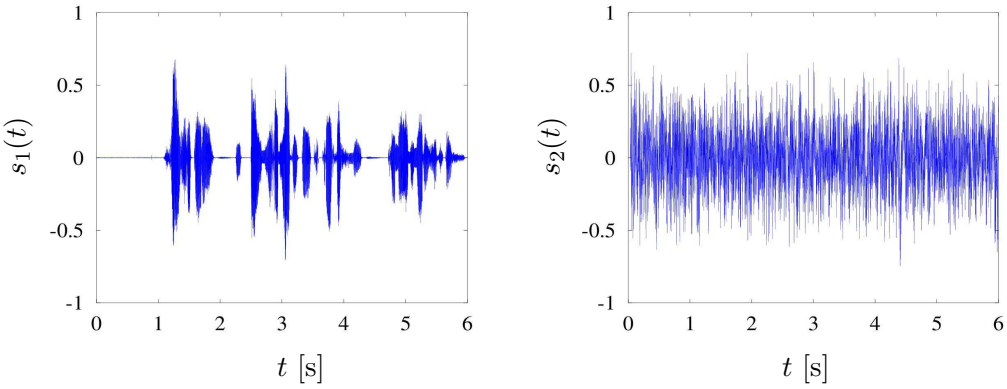


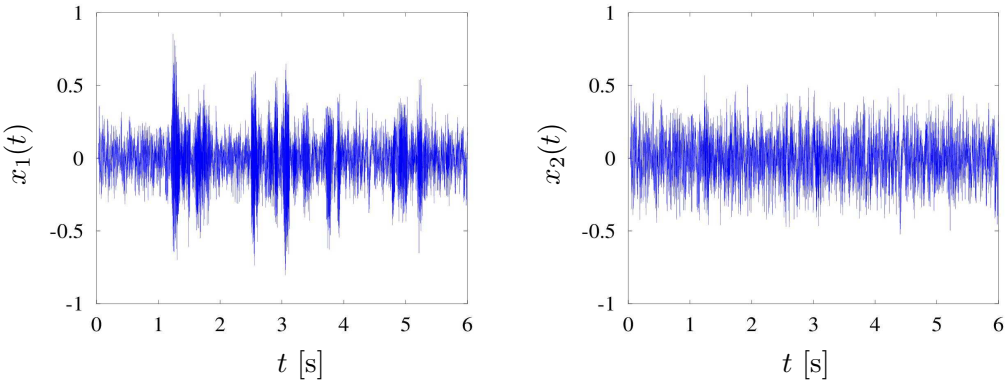
FIGURE 4. Histograms in each frame

4. **Simulations and Results.** In order to verify our proposals, several simulations were carried out. Target signals were 6 (3 females and 3 males) speaker's speech data [8] and noise was car noise [9]. The source signals were sampled at 8000Hz with 16bit resolution. The 30 mixture signals were calculated by Equation (1).

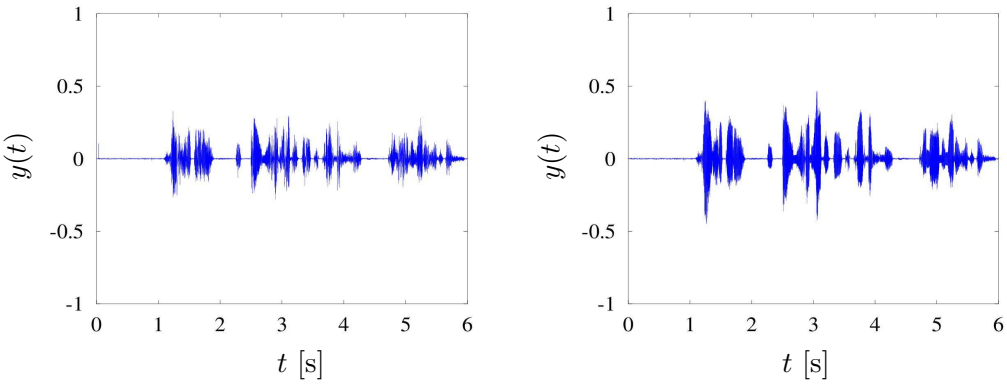
When the frame length was 1 second, the source signals and mixture signals are shown in Figure 5(a) and Figure 5(b), respectively. The separated signal using only the rotation BSS is shown in Figure 5(c). From this figure, it is found that some components



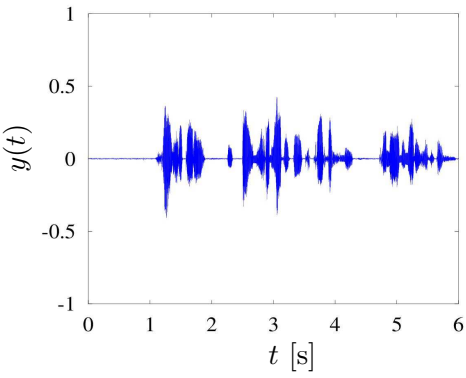
(a) Human speech and noise



(b) Mixture signals



(c) Estimated signal using only rotation BSS (d) Estimated signal using correlation



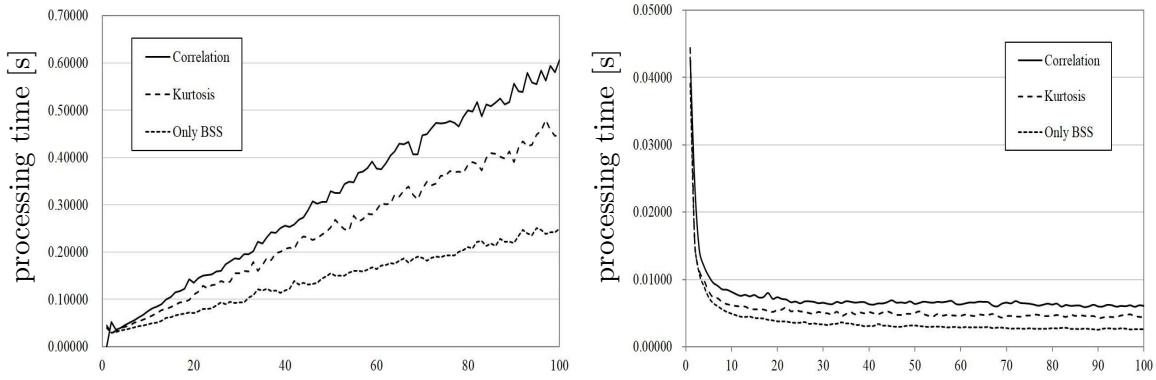
(e) Estimated signal using kurtosis

FIGURE 5. Simulation results

of target speech signal are lost. Figure 5(d) and Figure 5(e) are the estimated signals using correlation and kurtosis, respectively. From both these waveforms, the proposed method

can estimate the target signal except a scaling indeterminacy. From the simulation results, the proposed method can extract the target speech in all frames.

Figure 6 shows the processing time. The conditions of a computer were Windows 7 Professional, Intel(R)Celeron(R) CPU G460@1.8GHz, 4.00GB memory, and Octave-Ver3.6.4. The average processing time of all frames of 30 mixture signals is shown in Figure 6(a). The horizontal axis denotes the number of frames and the vertical axis denotes the processing time. From this result, it is found that the BSS with kurtosis is faster than that with correlation. Additionally, Figure 6(b) shows the processing time in each frame. From this figure, it is clear that our proposed methods can extract the target speech quickly in less than 0.01 second.



(a) Processing time of all frames (b) Processing time in each frame

FIGURE 6. Processing time

After the scaling indeterminacy of the separated signals are recovered, the average value of RMSE (root mean squared error) of the 30 patterns of separated signals by the proposed method was 5.15×10^{-5} in the case of the number of frames $L = 100$ and using kurtosis. From these results, it is found that our proposed method can extract the target human speech in the short-time frame.

5. **Conclusions.** For the rotation BSS in a short-time frame, the silent interval of the human speech detection methods based on the correlation and the kurtosis are proposed. When there is a silent interval in the frame, the rotation BSS updates the rotation angle. In a case of the short-time frame without the silent interval, the rotation BSS selects a rotation angle in the near past frame. From the simulations, it is found that the proposed method works even in the case of the frame length in 0.01 seconds. Therefore, it can be expected that the target speech is estimated when the sound source will be moved or the sound environment will be changed.

Acknowledgment. This work was supported by JSPS KAKENHI Grant Number JP 60455178. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] S. Makino, T.-W. Lee and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[2] J. Hwang, J. Seo, J.-W. Cho and H.-M. Park, A speech enhancement algorithm based on blind signal cancelation in diffuse noise environments, *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol.E99-A, no.1, pp.407-411, 2016.

[3] K. Kwon, J. W. Shin and N. S. Kim, Target source separation based on discriminative nonnegative matrix factorization incorporating cross-reconstruction error, *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol.E98-D, no.11, pp.2017-2020, 2015.

[4] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, Ltd., 2001.
[5] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing, Learning Algorithm and Applications*, John Wiley & Sons, Ltd., 2002.
[6] T. Ishibashi, Y. Tajiri, K. Inoue and H. Gotanda, An approach to blind source separation based on rotation of joint distribution of observed mixture signals, *RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing*, pp.21-24, 2014.
[7] K. Higuchi, C. Okuma and T. Ishibashi, Target sound extraction using silent interval of human speech, *The 22nd International Symposium on Artificial Life and Robotics*, pp.900-903, 2017.
[8] Acoustical Society of Japan, ASJ continuous speech corpus Japanese newspaper article sentences, *JNAS*, vols.1-16, 1997.
[9] NTT Advanced Technology Corporation, *Ambient Noise Database for Telephonometry 1996*, 1996.