ICIC Express Letters
Volume **11**, Number **12**, December **2017**
 ISSN 1881-803X
pp. **1715–1721**

FAST BLIND SOURCE SEPARATION AND TARGET HUMAN SPEECH EXTRACTION METHOD FOR ACOUSTIC SIGNALS

Takaaki Ishibashi, Hirohito Shintani and Kazuo Nagata

Department of Information, Communication and Electronic 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. *This paper proposes a fast blind source separation and target human speech extraction method for acoustic signals. The proposed method can estimate a human speech signal using a ratio of observed mixture signals. Since the human speech has silent interval, the ratio of the observed signals depends on transfer functions of noise from the sound source to the microphones. Therefore, the target human speech without noise is extracted by using the ratio of the observed signals. It is found that the proposed method can estimate and extract the target signal from several simulations.*
Keywords: Blind source separation, Noise reduction, Acoustic signal processing, Target speech extraction

1. **Introduction.** BSS (Blind Source Separation) is a method for estimating the sound sources from observed mixture signals without using the information about the sources and the transfer functions. For BSS, ICA (Independent Component Analysis) [1, 2] can estimate original source signals from their mixtures, provided that the sources are statistically independent. For the instantaneous mixtures, the original sources can be completely recovered except for indeterminacy of scale and permutation. The indeterminacy of scale is that the amplitude scale of the separated signals is not equal to that of the source signals. The indeterminacy of permutation is that the order of the separated signals is not equal to that of the source signals. Furthermore, ICA algorithms are iteration methods based on a gradient method or Newton method. This fact means that these algorithms are not good at a real-time processing.

For a real-time separating process, several methods have been proposed. SS (Spectral Subtraction) [3] and SAFIA (sound source Segregation based on estimating incident Angle of each Frequency component of Input signals Acquired by multiple microphones) [4] can estimate the original source signals. In these methods, the musical-noise is generated depending on the parameter. In order to reduce the musical-noise, a method based on the high-order statistics has been proposed [5]. However, multivariate data are necessary for the method. Although a method of forming directivity has been proposed [6], there is a problem that the residual noise remains. To the best of our knowledge, there are no real-time BSS method and its application.

In order to separate in the real-time process, we have already proposed a separation method based on a distribution of observed mixture signals [7, 8]. The previous method needs the trigonometric functions because the separated signals are rotated by estimating rotation angle. In this paper, we propose a fast blind source separation and target human speech extraction method using the ratio of the observed signals. The amount of calculation of the proposed algorithm is reduced and the method is possible to separate more quickly than the previous methods.

2. **Blind Source Separation.** Consider that some source signals are observed by some microphones. In this situation, 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, $s_n(t)$ $(n = 1, 2, \ldots, N)$ denote source signals, and N and M denote the number of the sources and the microphones, respectively. Using matrix and vectors, $\boldsymbol{x}(t) = [x_1(t), \ldots, x_m(t), \ldots, x_M(t)]^T$ by M microphones are expressed as

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) \tag{2}$$

where $\boldsymbol{s}(t) = [s_1(t), \ldots, s_n(t), \ldots, s_N(t)]^T$ denote unknown source signals and A denotes an unknown mixing matrix.

The estimated signals $y_n(t)$ for the sources are expressed as

$$y_n(t) = \sum_{m=1}^{M} w_{nm} x_m(t) \tag{3}$$

where w_{nm} denote estimated separating parameters. Using matrix and vectors in the same way, the estimated signals $\boldsymbol{y}(t) = [y_1(t), \ldots, y_n(t), \ldots, y_N(t)]^T$ are expressed as

$$\boldsymbol{y}(t) = W\boldsymbol{x}(t) \tag{4}$$

where W denotes an estimated demixing matrix.

The natural gradient [1], which is a representative algorithm of ICA, is a gradient method based on finding a minimum of the Kullback-Leibler divergence $I(\boldsymbol{y}(t))$. Using entropy $H(\boldsymbol{y}(t))$ of the separated signals $\boldsymbol{y}(t)$ with density $p(\cdot)$, the Kullback-Leibler divergence $I(\boldsymbol{y}(t))$ is defined as follows.

$$I(\boldsymbol{y}(t)) = \int p(\boldsymbol{y}(t)) \log \frac{p(\boldsymbol{y}(t))}{\prod_{n=1}^{N} p(y_n(t))} d\boldsymbol{y}(t) = \sum_{n=1}^{N} H(y_n(t)) - H(\boldsymbol{y}(t)) \tag{5}$$

$$H(\boldsymbol{y}(t)) = -\int p(\boldsymbol{y}(t)) \log p(\boldsymbol{y}(t)) d\boldsymbol{y}(t) \tag{6}$$

The natural gradient algorithm is formulated as

$$W + \Delta W = W - \eta \frac{\partial I(\boldsymbol{y}(t))}{\partial W} W^T W = W + \eta E\left[I - \boldsymbol{\varphi}\left(\boldsymbol{y}(t)\right) \boldsymbol{y}^T(t)\right] W \tag{7}$$

where $\boldsymbol{\varphi}(\cdot)$ denotes a nonlinear function, η denotes a learning parameter and I denotes a unit matrix.

The FastICA [2], which is another typical algorithm of ICA, is based on a fixed-point iteration scheme for finding a maximum of the non-Gaussianity of the separated signals. A measure of non-Gaussianity is given by negentropy $J(\boldsymbol{y}(t))$ as

$$J(\boldsymbol{y}(t)) = H\left(\boldsymbol{y}_{gauss}(t)\right) - H(\boldsymbol{y}(t)) \propto \{E[G(\boldsymbol{y}(t))] - E[G(\boldsymbol{\nu})]\}^2 \tag{8}$$

where $\boldsymbol{y}_{gauss}(t)$ is a Gaussian random variable of the same covariance matrix as $\boldsymbol{y}(t)$, $G(\cdot)$ denotes non-quadratic function and $\boldsymbol{\nu}$ denotes a Gaussian variable of zero mean and unit variance. Under the assumption that all the whitened mixtures are zero-mean and unit variances, the FastICA algorithm based on the maxima of the approximation of the negentropy for one-unit is formulated as

$$\boldsymbol{w}_n^+ \leftarrow E\left[\boldsymbol{x}(t) g\left(\boldsymbol{w}_n^T \boldsymbol{x}(t)\right)\right] - E\left[g'\left(\boldsymbol{w}_n^T \boldsymbol{x}(t)\right)\right] \boldsymbol{w}_n \tag{9}$$

$$\boldsymbol{w}_n \leftarrow \frac{\boldsymbol{w}_n^+}{\| \boldsymbol{w}_n^+ \|} \tag{10}$$

where $\boldsymbol{w}_n$ is a demixing weight vector, $g(\cdot)$ is a nonlinear function and $g'(\cdot)$ is its differential function. To estimate several components, we run any one-unit algorithm, and $\boldsymbol{w}_n$ is orthogonalized with $\boldsymbol{w}_j$, $(j = 1, \ldots, n-1)$ such as

$$\boldsymbol{w}_n \leftarrow \boldsymbol{w}_n - \sum_{j=1}^{n-1} \left(\boldsymbol{w}_n^T \boldsymbol{w}_j\right) \boldsymbol{w}_j \tag{11}$$

and $\boldsymbol{w}_n$ is again regularized by Equation (10).

ICA algorithms are based on statistically independent of the sources and these algorithms are the iterative method. It means that ICAs are not good at real-time processing.

3. **BSS Based on Rotation of Distribution.** In order to estimate the source signals and to extract the target source signal, we have already proposed a BSS method based on a rotation of a joint distribution of the observed signals. Consider two speakers have uttered in front of two microphones. A joint distribution of the source signals is plotted where the horizontal and the vertical axes are denoted by the amplitude of $s_1(t)$ and $s_2(t)$, respectively. The joint distribution of source signals is orthogonal. Using the observed mixture signals, the joint distribution is oblique. From these facts, the essence of BSS is to transform from the oblique distribution of the mixtures to the orthogonal distribution of the sources.

Our basic rotation BSS method [7] has 3 steps: whitening, rotation and scaling adjustment. In order to orthogonalize a crossed distribution of mixture signals, we calculate as

$$\widetilde{\boldsymbol{x}}(t) = \Lambda^{-\frac{1}{2}} \Phi^T \boldsymbol{x}(t) = Q\boldsymbol{x}(t) \tag{12}$$

where Φ is the orthogonal matrix of eigenvectors of $E\left[\boldsymbol{x}(t)\boldsymbol{x}^T(t)\right]$, Λ is the diagonal matrix of its eigenvalues and Q denotes a whitening matrix. The joint distribution is recovered except for indeterminacy of rotation and scaling.

To solve the indeterminacy of rotation, we calculate the angle for the points of the joint distribution of $\widetilde{\boldsymbol{x}}(t)$ as

$$\phi(t) = \tan^{-1} \frac{\widetilde{x}_2(t)}{\widetilde{x}_1(t)} \tag{13}$$

and obtain a direction histogram of $\phi(t)$. The rotation angle θ is estimated as

$$\theta = \arg\max_{\phi(t)} \mathrm{hist}(\phi(t)) \tag{14}$$

and we estimate the rotation matrix R as follows.

$$R = \begin{bmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{15}$$

In the case that the number of the source signals is three or more, a histogram is calculated from the observed signals by multi microphones. And a joint distribution with a multi-dimensional space is orthogonalized based on the rotation angle from the histogram.

Therefore, the rotation BSS method is formulated as follows.

$$\boldsymbol{y}(t) = W\boldsymbol{x}(t) = RQ\boldsymbol{x}(t) \tag{16}$$

For the scale indeterminacy, we introduce a scale adjustment method [9] as follows.

$$\boldsymbol{v}_n(t) = W^{-1}[0, \ldots, 0, y_n(t), 0, \ldots, 0]^T = (RQ)^{-1}[0, \ldots, 0, y_n(t), 0, \ldots, 0]^T \tag{17}$$

The basic rotation BSS can estimate the source signals. However, when the dimension of the distribution is changed by increasing or decreasing the number of sound sources, the algorithm is complicated because it requires whitening. Therefore, a rotation BSS algorithm without whitening has been proposed [8].

In the same way as Equation (13), we calculate the angle without whitening by using observed mixture signals $\boldsymbol{x}(t)$ as follows.

$$\phi(t) = \tan^{-1} \frac{x_2(t)}{x_1(t)} \tag{18}$$

The rotation angle θ_n of the nth peak of the histogram is estimated as Equation (14). Therefore, the rotation BSS without whitening has been proposed as follows.

$$u_n(t) = x_2(t) \cos\theta_n - x_1(t) \sin\theta_n \tag{19}$$

By detecting the peaks of the histogram, the method can remove the noise corresponding to the peak. The method does not need whitening and orthogonalization.

4. **BSS and Target Extraction.** From the above discussions, the estimated signals recover the source signals. However, the separated signals have a permutation problem. This means that the target source signal cannot be extracted. Therefore, we propose an extraction method for target speech signal under a noisy environment. And the previous method needs the trigonometric functions since the separated signals are rotated by estimating the rotation angle. In this paper, a new method without the trigonometric functions is proposed using a ratio of observed signals.

In the case that the source signals are a human speech and a stationary noise, the distribution and the histogram using their mixtures have only one peak. The peak is the noise component because human speech has a silent interval. In order to separate and extract the human speech, we calculate the ratio $r(t)$ for each point of the joint distribution of $\boldsymbol{x}(t)$ as

$$r(t) = \frac{x_2(t)}{x_1(t)} \tag{20}$$

and obtain the histogram of $r(t)$. Then, we define r calculated as the mode value (the most frequent value) of $r(t)$ as follows.

$$r = \arg\max_{r(t)} \mathrm{hist}(r(t)) \tag{21}$$

The estimated value r means the ratio of the transfer functions from the noise to two microphones.

Therefore, a new blind source separation and target speech extraction method without the trigonometric functions is proposed as follows (see Appendix).

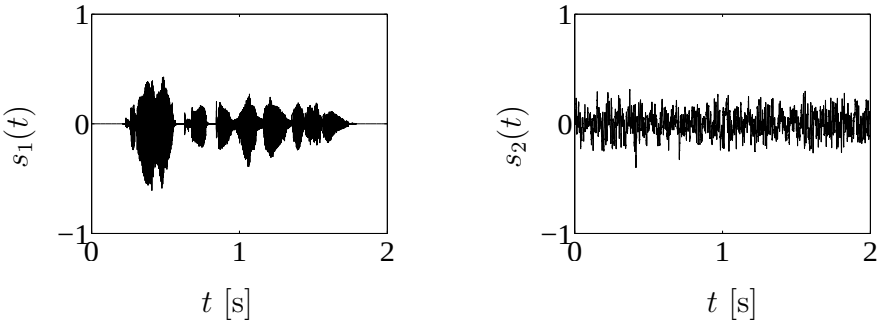
$$y(t) = x_2(t) - r x_1(t) \tag{22}$$

The proposed method can separate and extract the human speech signal at the same time.

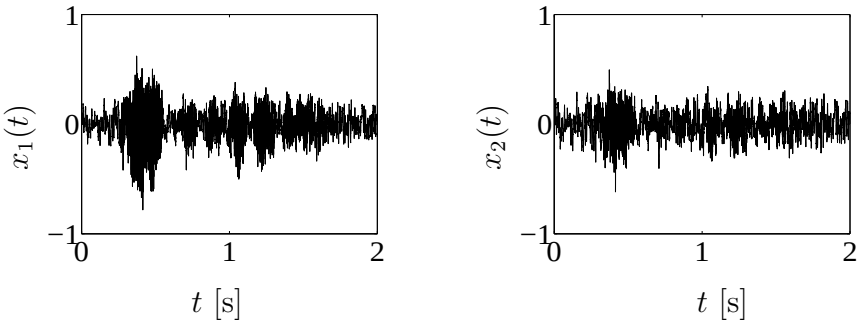
5. **Simulation.** In order to verify our proposals, several simulations were carried out. Target human speeches were 6 speaker's (3 females and 3 males) signals [10] in 2 seconds, and the noise signals were 5 patterns of car noises [11]. The mixture signals were sampled at a rate of 8000Hz with 16bit resolution. The mixture signals were calculated by Equation (2) which the diagonal components have $0.9 \pm \eta$ and non-diagonal components have $0.6 \pm \eta$, and η is a random value from 0 to 0.1. The simulations were carried out using 30 mixture signals.

The conditions of a computer were Windows 7 Professional, Intel(R) Core(TM) i7-3770 CPU @3.40GHz, 8.00GB memory, and MATLAB Version 7.11.0.584(R2010b). The average processing time of 30 mixture signals using the proposed method was 0.02127 seconds. In the NG algorithm, the nonlinear function was chosen as $\varphi_n(y_n(t)) = \tanh(y_n(t))$ and the matrix was initialized by random numbers from -0.5 to 0.5. The average processing time of 30 signals using NG was 0.14421 seconds.

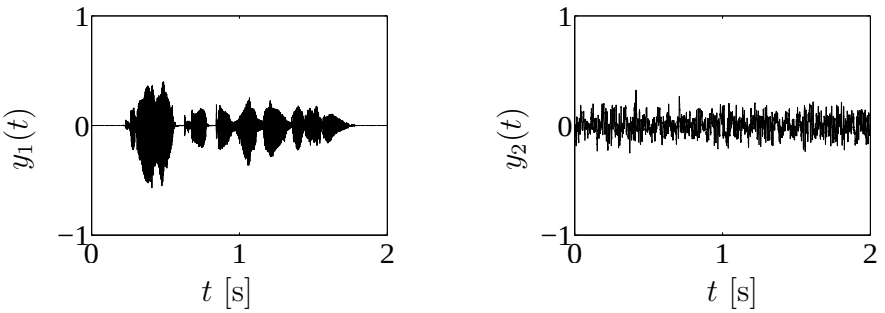
Figure 1 shows the simulation results when a female speaker's utterance is under the car noise. (a) are two sources. One is the female speaker's uttered voice. The other one is car engine sound. The waveforms of the observed mixture signals by two microphones are shown in (b). The separated signals $y_1(t)$ and $y_2(t)$ by NG method are shown in (c). The waveforms of (c) is similar to the waveform of the sources. The extracted signal by the proposed method is shown in (d). The waveform is similar to the waveform of the original female speaker's utterance. From the waveforms, it is found that the estimated



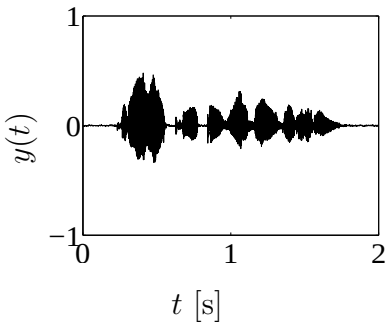
(a) Source signals



(b) Mixture signals



(c) Separated signals by NG algorithm



(d) Extracted signal by the proposed method

FIGURE 1. Simulation results when the female speaker uttered under the car noise: (a) source signals, (b) mixture signals, (c) separated signals by NG algorithm and (d) extracted signal by the proposed method

signals can restore the source signals and the proposed method can extract the human speech. Similar simulation results were obtained for all 30 patterns.

After the scaling indeterminacy of the separated signals is recovered, the average value of RMSE (Root Mean Squared Error) of the 30 patterns of the separated signals by the proposed method was 9.05×10^{-5}. For comparison, the mean value of RMSE was 8.32×10^{-5} when the separated signals are estimated by the NG algorithm of ICA.

From these results, it is found that our proposed method has the same separation performance at ICA and the proposed method can separate the source signals faster than the ICA algorithms.

6. **Conclusions.** This paper proposes a blind source separation method for human speech signal in a noisy environment. The proposed method can estimate the separating parameters based on the amplitude ratio of the joint distribution of the observed mixture signals. The algorithm of the proposed method is very simple using the silent interval of human speech. By using the method, we can separate and extract the human speech signal at the same time. The amount of calculation of the algorithm is reduced. And the method is possible to separate more quickly than the previous methods.

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] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, Ltd, 2001.
[3] S. F. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.ASSP-27, no.2, pp.113-120, 1979.
[4] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda, Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones, *Acoustical Science and Technology*, vol.22, no.2, pp.149-157, 2001.
[5] R. Miyazaki, H. Saruwatari, S. Nakamura, K. Shikano, K. Kondo, J. Blanchette and M. Bouchard, Musical-noise-free blind speech extraction integrating microphone array and iterative spectral subtraction, *Signal Processing (Elsevier)*, vol.102, pp.226-239, 2014.
[6] K. Hayama, T. Ishibashi, C. Okuma and H. Gotanda, Implementation of directional characteristics by real-time processing of sounds observed by two microphones, *ICIC Express Letters*, vol.10, no.1, pp.251-254, 2016.
[7] T. Ishibashi, K. Fujimori, K. Inoue and H. Gotanda, Target speech extraction based on orthogonalization of joint distribution of observed signals, *Proc. of the 44th ISCIE International Symposium on Stochastic Systems Theory and Its Applications*, pp.289-294, 2012.
[8] 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, *2014 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing*, pp.21-24, 2014.
[9] N. Murata, S. Ikeda and A. Ziehe, An approach to blind source separation based on temporal structure of speech signals, *Neurocomputing*, vol.41, nos.1-4, pp.1-24, 2001.
[10] Acoustical Society of Japan, ASJ continuous speech corpus Japanese newspaper article sentences, *JNAS*, vols.1-16, 1997.
[11] NTT Advanced Technology Corporation, *Ambient Noise Database for Telephonometry 1996*, 1996.

Appendix A. **Derivation of Equation (22).** In the case of two-sources and two-microphones, the mixture signals are observed as follows.

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \tag{23}$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \tag{24}$$

First, consider the case where the $s_1(t)$ is human speech and the $s_2(t)$ is stationary noise. Since the human speech has silent interval, the observed mixture signals are presented frequently only noise as follows.

$$x_1(t) = a_{12}s_2(t) \tag{25}$$

$$x_2(t) = a_{22}s_2(t) \tag{26}$$

The ratio r in Equation (21) of the mixture signals is estimated as

$$r = \arg\max_{r(t)} \text{hist}(r(t)) = \arg\max_{r(t)} \text{hist}\left(\frac{x_2(t)}{x_1(t)}\right) = \frac{a_{22}}{a_{12}} \tag{27}$$

Therefore, the estimated signal $y(t)$ is generated by Equation (22) as

$$y(t) = x_2(t) - rx_1(t) \tag{28}$$

$$= \{a_{21}s_1(t) + a_{22}s_2(t)\} - \frac{a_{22}}{a_{12}}\{a_{11}s_1(t) + a_{12}s_2(t)\} \tag{29}$$

$$= \{a_{21}s_1(t) + a_{22}s_2(t)\} - \left\{\frac{a_{11}a_{22}}{a_{12}}s_1(t) + a_{22}s_2(t)\right\} \tag{30}$$

$$= a_{21}s_1(t) - \frac{a_{11}a_{22}}{a_{12}}s_1(t) \tag{31}$$

$$= \left\{a_{21} - \frac{a_{11}a_{22}}{a_{12}}\right\}s_1(t) \tag{32}$$

$$= \frac{a_{12}a_{21} - a_{11}a_{22}}{a_{12}}s_1(t) \tag{33}$$

Equation (33) means that the signal $y(t)$ estimates the original human speech signal except the scaling indeterminacy. And Equation (22) can extract the target human speech.

Next, consider the case where the $s_1(t)$ is stationary noise and the $s_2(t)$ is human speech. The observed mixture signals are represented as follows.

$$x_1(t) = a_{11}s_1(t) \tag{34}$$

$$x_2(t) = a_{21}s_1(t) \tag{35}$$

The ratio r in Equation (21) is estimated as follows.

$$r = \arg\max_{r(t)} \text{hist}(r(t)) = \arg\max_{r(t)} \text{hist}\left(\frac{x_2(t)}{x_1(t)}\right) = \frac{a_{21}}{a_{11}} \tag{36}$$

The estimated signal $y(t)$ is generated by Equation (22) as

$$y(t) = x_2(t) - rx_1(t) \tag{37}$$

$$= \{a_{21}s_1(t) + a_{22}s_2(t)\} - \frac{a_{21}}{a_{11}}\{a_{11}s_1(t) + a_{12}s_2(t)\} \tag{38}$$

$$= \{a_{21}s_1(t) + a_{22}s_2(t)\} - \left\{a_{21}s_1(t) + \frac{a_{12}a_{21}}{a_{11}}s_2(t)\right\} \tag{39}$$

$$= a_{22}s_2(t) - \frac{a_{12}a_{21}}{a_{11}}s_2(t) \tag{40}$$

$$= \left\{a_{22} - \frac{a_{12}a_{21}}{a_{11}}\right\}s_2(t) \tag{41}$$

$$= \frac{a_{11}a_{22} - a_{12}a_{21}}{a_{11}}s_2(t) \tag{42}$$

In this case, the $y(t)$ can estimate and extract the target human speech signal.