ICIC Express Letters
Part B: Applications
Volume **12**, Number **5**, May **2021**

ICIC International ©2021 ISSN 2185-2766
pp. **443–451**

BLIND SOURCE SEPARATION FOR HUMAN SPEECHES BASED ON ORTHOGONALIZATION OF JOINT DISTRIBUTION OF OBSERVED MIXTURE SIGNALS

Takaaki Ishibashi[1] and Kei Eguchi[2]

[1]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

[2]Department of Information Electronics
Fukuoka Institute of Technology
3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka 811-0295, Japan

Received October 2020; accepted January 2021

Abstract. *In this paper, we propose a real-time BSS (Blind Source Separation) method for human speech signals. The proposed method can estimate unknown source signals based on orthogonalization of the joint distribution of observed mixture signals by microphones. For a real-time separation, orthogonalization and scaling adjustment are applied to a short-time frame. In this case, indeterminacy of output channel occurs at every short-time frame. Therefore, an output channel selection is also proposed using our estimation method of ratio based on transfer functions. From several simulation results, the proposed method can estimate the original source signals for a human speech noise, a stationary noise and a convoluted human speech noise.*

Keywords: Blind source separation, Independent component analysis, Orthogonalization of joint distribution, Output channel selection, Real-time BSS

1. **Introduction.** Speech recognition technology [1] was improved to provide a speech recognition engine with extremely high recognition capabilities in ideal environments, i.e., no surrounding noise or in a well-known noisy environment. However, it is still difficult to achieve a desirable recognition rate in a household or office with sounds of daily activities. Therefore, some preprocessing prior to recognition is necessary to reduce noise and to select the target speech signal.

Several methods of noise reduction using the ICA (Independent Component Analysis) have been proposed [2, 3, 4]. The ICA may separate unknown sources from their mixtures without information on transfer functions, provided that the sources are statistically independent. For instantaneous mixtures, the original sources can be fully recovered, except for indeterminacy of scaling and permutation problems.

ICA-based applications are expected in many areas, including speech recognition technology, EEG (Electroencephalogram) data analysis, MEG (Magnetoencephalography) data analysis and image processing [5, 6, 7]. However, the device using ICA rarely exists since ICAs are not good at real-time processing.

For real-time separation, many methods have been proposed. SS (Spectral Subtraction) [8], SAFIA (sound source Segregation based on estimating incident Angle of each Frequency component of Input signals Acquired by multiple microphones) [9] and a microphone system with variable arbitrary directional pattern [10] can estimate the original source signals. In these methods, the musical-noise was generated depending on the parameter. To reduce the musical-noise, a method based on high-order statistics has been

DOI: 10.24507/icicelb.12.05.443

proposed [11]. However, real-time processing is difficult with the method since multivariate information is needed for the method.

This paper proposes a BSS method for human speech signals. The authors have previously reported method for extracting target human speech [12]. However, the existing method could not solve the scaling indeterminacy. The proposed BSS is based on orthogonalization and scale adjustment. In order for a real-time processing, the proposed method can be expanded using short-time frames. In a real-time processing, a channel selection problem occurs at each time frame. The proposed method can also be corrected with our method of estimating the ratio of transfer functions. From the simulation results, it has been confirmed that the proposed separation and selection method is valid.

The manuscript contains the following sections. Section 1 is the introduction of this work. Section 2 describes the BSS method and the scaling adjustment for the separated signals. Section 3 proposes a new BSS method for speech signals. In addition, we propose a real-time BSS that introduces the short-time frame processing. Section 4 shows the experimental results of sound source separation for instantaneous mixing and convolution mixing. Section 5 briefly summarizes the results of this work.

2. Blind Source Separation.

Under the situation that some sound sources are observed by microphones, a BSS is a method to estimate the sound sources without using the information about the sources and the transfer functions. For the BSS, ICA can separate the sources from their mixtures when they are statistically independent.

Consider the case where sound sources are observed by microphones. We assume that the observed mixture signals $\boldsymbol{x} = [x_1, \dots, x_m, \dots, x_M]^T$ are generated as a linear mixture of the sources as

$$\boldsymbol{x} = A\boldsymbol{s} \tag{1}$$

where $\boldsymbol{s} = [s_1, \dots, s_n, \dots, s_N]^T$ denotes unknown source signals and A denotes an unknown mixing matrix whose elements are a_{mn}.

The separated signals $\boldsymbol{u} = [u_1, \dots, u_n, \dots, u_N]^T$, the estimate of the source signals $\boldsymbol{s}$, are expressed as

$$\boldsymbol{u} = W\boldsymbol{x} \tag{2}$$

where $W = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_n, \dots, \boldsymbol{w}_N]^T$ denotes a demixing matrix. The matrix W is estimated by ICA algorithms such as natural gradient [3] and FastICA [4].

ICA can estimate the sources $\boldsymbol{s}$ except for indeterminacy of scaling and permutation under the assumption that each component of $\boldsymbol{s}$ is statistically independent. The separated matrix using ICA algorithms has scaling indeterminacy and permutation problem as

$$WA = PD \tag{3}$$

where P is a permutation matrix, in which all elements of each column and row are 0 except for one element with value 1, and $D = \mathrm{diag}[d_1, \dots, d_n, \dots, d_N]$ a diagonal matrix, of which elements d_n denote the scaling factors.

The indeterminacy of scaling is that the 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.

In order to solve the scaling indeterminacy, a method using the inverse of the demixing matrix W^{-1} has been proposed as follows [13].

$$\boldsymbol{v}_n = W^{-1}[0, \dots, 0, u_n, 0, \dots, 0]^T \tag{4}$$

We have proved that the $\boldsymbol{v}_n = [v_{n1}, \dots, v_{nm}, \dots, v_{nM}]^T$ is uniquely expressed as a product of the n-th source s_n and the transfer function a_{mn} from the n-th source to the m-th microphone as follows [14].

$$\boldsymbol{v}_n = [a_{1n}s_n, \ldots, a_{mn}s_n, \ldots, a_{Mn}s_n]^T \tag{5}$$

This means that v_{nm} is the observation of the n-th source s_n through the m-th microphone. These output signals by this approach are the same as [15]. It is also clarified that every v_{nm} has no ambiguity of scale in that the scaling factor is a transfer function itself, while the scale factor d_n for the separated signal u_n varies arbitrarily. The estimated signals $\boldsymbol{y} = [y_1, \ldots, y_n, \ldots, y_N]^T$ are selected as follows.

$$y_n = \max_m v_{nm} = \max_m a_{mn}s_n \tag{6}$$

3. **A New Blind Source Separation.** The original source signals can be recovered using ICA. However, ICAs are based on statistical independence of the sources and these algorithms are the iterative method. It means that ICAs are not good at real-time processing. Therefore, we propose a new BSS method for real-time process. The authors have already proposed a fast BSS method [16]. The proposed method is applied to short-time frame processing for real-time operation.

3.1. **Principle of our BSS.** In order to orthogonalize for the crossed distribution of mixture signals, we calculate as

$$\widetilde{\boldsymbol{x}} = \Lambda^{-\frac{1}{2}}\Phi^T\boldsymbol{x} = Q\boldsymbol{x} \tag{7}$$

where Φ is the orthogonal matrix of eigenvectors of $E[\boldsymbol{x}\boldsymbol{x}^T]$, Λ is the diagonal matrix of its eigenvalues and Q denotes a whitening matrix. From whitening processing, the joint distribution of the sources 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 $\widetilde{\boldsymbol{x}}$ as

$$\phi = \tan^{-1}\frac{\widetilde{x}_2}{\widetilde{x}_1} \tag{8}$$

and obtain the direction histogram of ϕ.

The rotation angle θ is estimated as

$$\theta = \arg\max_\phi \mathrm{hist}(\phi) \tag{9}$$

and we estimate the rotation matrix R as follows.

$$R = \begin{bmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{10}$$

In the case which 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, our BSS method is as follows.

$$\boldsymbol{u}(t) = W\boldsymbol{x}(t) = RQ\boldsymbol{x}(t) \tag{11}$$

Using the proposed method, the rotated signals are recovered the original sources except for the scale indeterminacy.

For the scale indeterminacy, we introduce Equation (4) as follows.

$$\boldsymbol{v}_n = W^{-1}[0, \ldots, 0, u_n, 0, \ldots, 0]^T = (RQ)^{-1}[0, \ldots, 0, u_n, 0, \ldots, 0]^T \tag{12}$$

3.2. **A real-time BSS.** For a real-time BSS, we introduce a short-time frame separation as

$$\boldsymbol{x}_k = [\boldsymbol{x}_k(0), \ldots, \boldsymbol{x}_k(l), \ldots, \boldsymbol{x}_k(L-1)] \tag{13}$$

$$= [\boldsymbol{x}((k-1)L), \ldots, \boldsymbol{x}((k-1)L+l), \ldots, \boldsymbol{x}(kL-1)] \tag{14}$$

where l $(= 0, \dots, L-1)$ denotes a data point in the frame, L denotes a frame length, k $(= 1, \dots, K)$ denotes an index of the frames and K denotes a number of the frames.

Using the short-time $\boldsymbol{x}_k$, the mixing model Equation (1) and the separating model Equation (2) are represented as

$$\boldsymbol{x}_k = A_k \boldsymbol{s}_k \tag{15}$$

$$\boldsymbol{u}_k = W_k \boldsymbol{x}_k \tag{16}$$

Then we estimate the source signals based on whitening, rotation and scale adjustment as

$$\boldsymbol{u}_k = R_k \Lambda_k^{-\frac{1}{2}} \Phi_k{}^T \boldsymbol{x}_k \tag{17}$$

$$\boldsymbol{v}_{kn} = W_k^{-1} [0, \dots, 0, u_{kn}, 0, \dots, 0]^T \tag{18}$$

However, the channel selection problem occurs. This fact is also called the permutation problem. This means that the order n of the separated signal y_n is not necessarily consistent with n of the source signal s_n. Therefore, the indeterminacy of permutation must be settled to get a meaningful signal y_n, before y_n is output.

In order to solve for the channel selection problem, a ratio r_{km} of $\boldsymbol{v}_{kn}$ is simply used because of Equation (5) as

$$r_{km} = \frac{v_{km2}}{v_{km1}} = \frac{a_{k2i} s_{ki}}{a_{k1i} s_{ki}} = \frac{a_{k2i}}{a_{k1i}} \tag{19}$$

The ratio r_{km} of $\boldsymbol{v}_{kn}$ is expressed by the ratio of the transfer functions. Based on this fact, indeterminacy of the output channel can be solved as follows.

$$\boldsymbol{y}_{k1} = \begin{cases} \boldsymbol{v}_{k1} & \text{if } |r_{k1}| - |r_{k-1,1}| \leq |r_{k2}| - |r_{k-1,1}| \\ \boldsymbol{v}_{k2} & \text{if } |r_{k1}| - |r_{k-1,1}| > |r_{k2}| - |r_{k-1,1}| \end{cases} \tag{20}$$

$$\boldsymbol{y}_{k2} = \begin{cases} \boldsymbol{v}_{k1} & \text{if } |r_{k2}| - |r_{k-1,2}| \leq |r_{k1}| - |r_{k-1,2}| \\ \boldsymbol{v}_{k2} & \text{if } |r_{k2}| - |r_{k-1,2}| > |r_{k1}| - |r_{k-1,2}| \end{cases} \tag{21}$$

4. Simulations.

In order to verify our proposals, several simulations were carried out.

4.1. Simultaneous utterance of speakers.

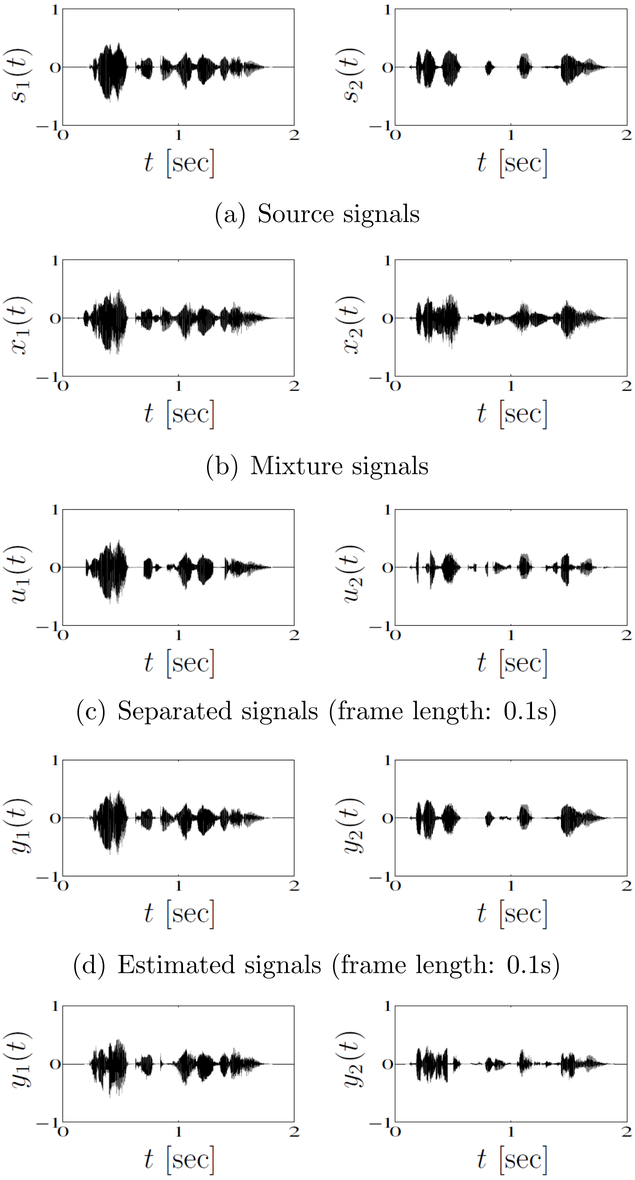
Sources were 6 speaker's (3 females and 3 males) speech signals [17] in 2 seconds. The mixed signals were sampled at a rate of 8000Hz with 16bit resolution. The mixture signals were calculated by Equation (1) in which the diagonal components have $0.9 \pm \eta$ and non-diagonal components have $0.6 \pm \eta$, η is a random value from 0 to 0.1. The simulations were carried out using 30 mixture signals.

In Figure 1, Figure 1(a) show a male and a female speaker uttered source signals, respectively, Figure 1(b) show mixture signals using Figure 1(a), and Figure 1(c) show separated signals. It is found that the separated signals without the channel selection cannot recover the source signals. Figure 1(d) are estimated signals of the male and the female utterance, respectively, by the proposed method including the channel selection. From the waveforms, it is found that the estimated signals can restore the source signals in the case which the frame length is 0.1 seconds. In the case which the frame length is shorter than 0.05 seconds as shown in Figure 1(e), the method cannot work well. It is because that calculation of whitening and rotation was not completed in the silent interval.

Table 1 shows NRR (Noise Reduction Rate) and processing time. The NRR is defined as follows [18].

$$\mathrm{NRR} = \frac{1}{2} \sum_{n=1}^{2} (\mathrm{SNR_O}n - \mathrm{SNR_I}n), \quad \mathrm{SNR_O}n = 10 \log_{10} \frac{|h_{nn} s_n|^2}{|h_{ni} s_i|^2}, \quad \mathrm{SNR_I}n = 10 \log_{10} \frac{|a_{nn} s_n|^2}{|a_{ni} s_i|^2} \tag{22}$$

where $\mathrm{SNR}_{\mathrm{O}}n$ and $\mathrm{SNR}_{\mathrm{I}}n$ are the output SNR and the input SNR, respectively, and $n \neq i$. h_{ji} is the element in the jth row and the ith column of the matrix $H = WA$. The SNRs are calculated under the assumption that the speech signal of the undesired speaker is regarded as noise. The processing time is the calculated time for whitening, rotation, and channel selection. From these results, it is found that the proposed method functions well



(a) Source signals

(b) Mixture signals

(c) Separated signals (frame length: 0.1s)

(d) Estimated signals (frame length: 0.1s)

(e) Estimated signals (frame length: 0.05s)

FIGURE 1. Waveforms of source signals to estimated signals under a speech noise

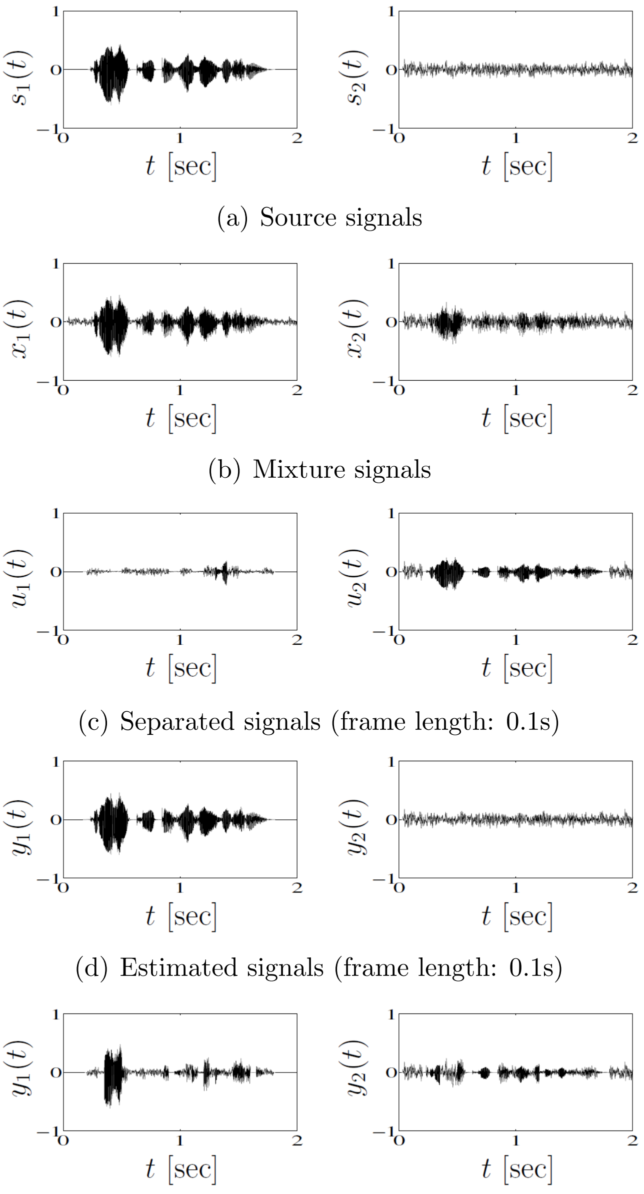
TABLE 1. NRR and processing time under a speech noise

Frame length (point)	NRR	Processing time
2s (16000)	53.2651dB	0.1003s
1s (8000)	31.2439dB	0.0542s
0.5s (4000)	31.6139dB	0.0278s
0.2s (1600)	15.8473dB	0.0113s
0.1s (800)	14.7130dB	0.0059s
0.05s (400)	3.9322dB	0.0032s

in the case which the frame length is longer than 0.1 seconds. The reason for this result is that whitening and rotation processing work accurately when the frame length is long.

4.2. **Utterance under stationary noise.** The simulations were carried out using 30 mixture signals generated by 6 speaker's (3 females and 3 males) speeches [17] and 5 pattern of car noises [19].

In Figure 2, Figure 2(a) show a male speaker uttered source signal and a stationary noise of a car, Figure 2(b) show mixture signals, Figure 2(c) show separated signals without our channel selection, and Figure 2(d) show estimated signals of the male's speech and the noise, respectively, by the proposed method. From these waveforms, it is found that the proposed separation and selection method can estimate the original sources when the frame length is 0.1 seconds. When the frame length was 0.05 seconds, the estimated signals as shown in Figure 2(e) could not be restored. This is due to whitening failure.



(a) Source signals

(b) Mixture signals

(c) Separated signals (frame length: 0.1s)

(d) Estimated signals (frame length: 0.1s)

(e) Estimated signals (frame length: 0.05s)

FIGURE 2. Waveforms of source signals to estimated signals under a stationary noise

Table 2 shows NRR and processing time. From the results, it is also found that the proposed method is functional well in the case that the frame length is longer than 0.1 seconds.

TABLE 2. NRR and processing time under a stationary noise

Frame length (point)	NRR	Processing time
2s (16000)	40.9086dB	0.0986s
1s (8000)	38.6070dB	0.0555s
0.5s (4000)	30.2176dB	0.0277s
0.2s (1600)	17.2991dB	0.0113s
0.1s (800)	16.2506dB	0.0059s
0.05s (400)	3.1983dB	0.0031s

4.3. **Virtual room with reflections and reverberation.** In order to verify that the proposed method operates in a real environment, we carried out an experiment to estimate the original source signals using convolutional mixed signals.

A method for generating impulse functions is provided in a virtual room [20]. We set a reverberation time 0.2 seconds in the virtual room with a length of 4 m, a width of 5 m, and a height of 3 m. In the virtual room, the source signals and the microphones were located as shown in Figure 3. The height of all sources and microphones are set 1.5 meters from the human mouth and ears are a height of about 1.5 meters. Using the virtual room, mixed convolutional signals were created using the impulse functions and the source signals.

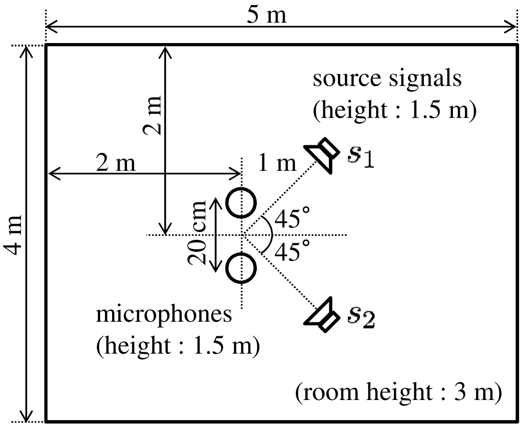
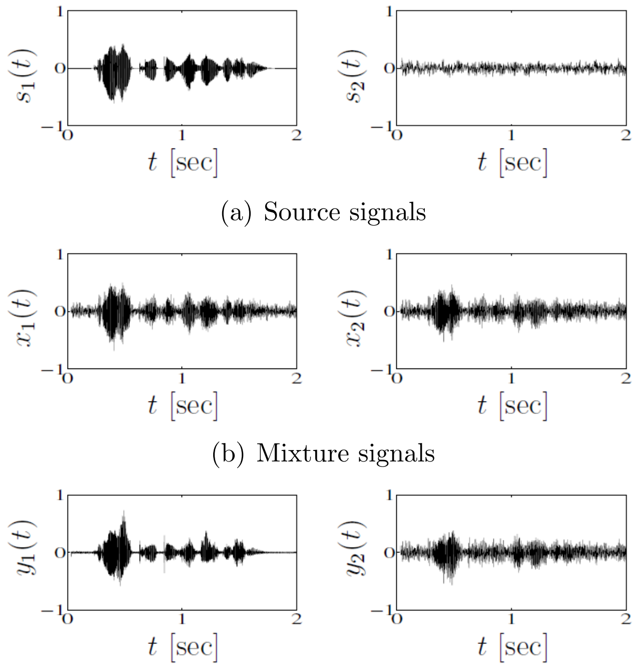


FIGURE 3. Placement of source signals and microphones in the virtual room

The waveforms of the estimated signals are shown in Figure 4. In the convolutional mixing, the arrival time of the sound wave was deviated, so that the estimation accuracy was lowered. However, it is found that the waveforms are very similar to the signals of the speaker's speech, and when the speech was heard, it was confirmed that the noise was removed well. Furthermore, the problem, that the existing method could not be solved by the scaling indeterminacy, could be solved by the proposed method.

5. **Conclusions.** This paper proposes a real-time blind source separation method for acoustic signals. For a real-time separation, we calculate orthogonalization for a short-time frame. In order to solve the problem of the channel selection, we also propose the correction method based on the estimated ratio of the transfer functions. From the simulation results, the proposed method can estimate the original source signals not only for human speech noise but also for stationary noise. In the convolutional mixture, it is



(a) Source signals

(b) Mixture signals

(c) Estimated signals (frame length: 0.5s)

FIGURE 4. Waveforms of source signals to estimated signals under reverberation

found that the waveforms are very similar to the speaker's speech and the noise was well removed.

The proposed method processes in a short frame. Therefore, under the condition that the moving sound source can be regarded as a fixed sound in a short time, it is expected that it can be separated by this algorithm. As in the frequency domain ICA, it can be extended to processing using spectrograms. By using the spectrogram processing, it is expected that the separation performance will be further improved for the convolution mixture signals.

Acknowledgment. This work was supported by JSPS KAKENHI Grant Number JP 20K12763. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers.

REFERENCES

[1] H. M. S. Naing, R. Hidayat, B. Winduratna and Y. Miyanaga, Psychoacoustical masking effect-based feature extraction for robust speech recognition, *International Journal of Innovative Computing, Information and Control*, vol.15, no.5, pp.1641-1654, 2019.

[2] T. W. Lee, *Independent Component Analysis, Theory and Applications*, Kluwer Academic Publishers, 1998.

[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing, Learning Algorithm and Applications*, John Wiley & Sons, Ltd., 2002.

[4] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, Inc., 2001.

[5] S. Ikeda and K. Toyama, Independent component analysis for noisy data: MEG data analysis, *Neural Networks*, vol.13, no.10, pp.1063-1074, 2000.

[6] J. Cao, N. Murata, S. Amari, A. Cichocki and T. Takeda, Independent component analysis for unaveraged single-trial MEG data decomposition and single-dipole source localization, *Neurocomputing*, vol.49, pp.255-277, 2002.

[7] A. Hyvärinen, J. Hurri and P. O. Hoyer, *Natural Image Statistics, A Probabilistic Approach to Early Computational Vision*, Springer, 2009.

[8] 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.

[9] M. Aoki, K. Furuya and A. Kataoka, Improvement of "SAFIA" source separation method under reverberant conditions, *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol.89, no.3, pp.22-37, 2006.

[10] C. Okuma, K. Hayama and T. Ishibashi, Two-channel microphone system with variable arbitrary directional pattern, *ICIC Express Letters*, vol.12, no.3, pp.229-236, 2018.

[11] Y. Takahashi, H. Saruwatari, K. Shikano and K. Kondo, Musical-noise analysis in methods of integrating microphone array and spectral subtraction based on higher-order statistics, *EURASIP Journal on Advances in Signal Processing*, DOI: 10.1155/2010/431347, 2010.

[12] T. Ishibashi, K. Higuchi and C. Okuma, Target human speech extraction method based on silent interval detection, *ICIC Express Letters, Part B: Applications*, vol.8, no.12, pp.1603-1610, 2017.

[13] 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.

[14] T. Ishibashi, K. Inoue, H. Gotanda and K. Kumamaru, Frequency domain independent component analysis without permutation and scale indeterminacy, *Proc. of the 41st ISCIE International Symposium on Stochastic Systems Theory and Its Applications*, pp.190-195, 2009.

[15] K. Matsuoka, Elimination of filtering indeterminacy in blind source separation, *Neurocomputing*, vol.71, nos.10-12, pp.2113-2126, 2008.

[16] T. Ishibashi H. Shintani and K. Nagata, Fast blind source separation and target human speech extraction method for acoustic signals, *ICIC Express Letters*, vol.11, no.12, pp.1715-1721, 2017.

[17] Acoustical Society of Japan, *ASJ Continuous Speech Corpus Japanese Newspaper Article Sentences (JNAS)*, 1997.

[18] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa and K. Shikano, Blind source separation combining independent component analysis and beamforming, *EURASIP Journal on Applied Signal Processing*, vol.2003, no.11, pp.1135-1146, 2003.

[19] NTT Advanced Technology Corporation, *Ambient Noise Database for Telephonometry 1996*, 1996.

[20] E. A. P. Habets, *Room Impulse Response Generator*, Tech. Rep, Technische Universiteit Eindhoven, 2.2.4, 2006.