Blind Source Separation and Human Speech Extraction for Three Sound Sources Using Silent Interval

Kana Higuchi^{a,*}, Chiharu Okuma^a, Takaaki Ishibashi^a

^aNational Institute of Technology, Kumamoto College
2659-2 Suya, Koshi, Kumamoto 861-1102, Japan

*Corresponding Author: ae16higuchi@g.kumamoto-nct.ac.jp

Abstract

In this paper, we propose a blind source separation method for multiple sound source signals. We have proposed an estimation method for source signals without noise from observed mixture signals. The method estimates the rotation angle of a distribution of observed mixture signals. The distribution is rotated based on the estimated angle. Our method can estimate two source signals from two mixture signals. However, in case of three source signals, the source signals cannot be estimated by one-time processing. Therefore, we apply processing multiple times to observed mixture signals. In this way, we propose an estimation method for the original source.

Keywords: blind source separation, joint distribution, silent interval, human speech extraction

1. Introduction

We are living in an 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 sounds in such noisy places. Recently, a lot of methods of separating noises and target sounds has been proposed by researchers. Blind source separation (BSS) is a method for estimating the source signals from observed mixture signals without the information about the sources and the transfer functions. For BSS, independent component analysis (ICA) is well known^(1,2). In case sources are statistically independent each other, ICA can separate unknown sources from mixtures. However, ICA takes a long time to separate the source signals from mixture signals. Therefore, ICA is not good at a real-time processing. Furthermore, since ICA is separation algorithm, a target extraction processing is needed.

We have already proposed a rotation BSS method for two sound sources signals^(3,4). In the case of two sound sources,

we can estimate the source signals. In this paper, we propose a new BSS method for three source signals. Our method can be applied even when the number of original source signals is three. The processes of the proposed method are as follows. First, we create a joint distribution and a histogram using the mixture signals observed by a pair of microphones. Secondly, a rotation angle of the distribution is estimated. Noise-removed mixture signal is estimated by rotating the distribution using the rotation angle. In the same way, we estimate another noise-removed mixture signal observed by a different pair of microphones. We can estimate the mixture signals of two sources without one source. Additionally, we estimate the source signal using two noise-removed mixture signals again. As a consequence, the source signal can be estimated. The proposed method outputs the human speech. It means that the method can separate and extract the human speech.

In order to verify our proposals, several simulations were carried out when the source signals is two speaker speech and one stationary noise. From the simulation results, the proposed method can remove the noise from the mixture signals with the first processing, because the two speaker speech include a silent interval. In addition, we can obtain the source signals by reprocessing the signals without noise. These results lead to the conclusion that the proposed method can extract the human speech.

2. Rotation BSS

BSS estimates source signals using only observed signals. When source signals $s_n(t)$ ($n = 1, 2, \dots, N$) are observed by microphones, the observed mixture signals $x_m(t)$ ($m = 1, 2, \dots, M$) are expressed as

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

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

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

where w_{nm} denotes an estimated separating parameter. In order to estimate w_{nm} , many methods have been proposed^(1,2).

In order to separate and extract target human speech, we have already proposed a rotation BSS^(3,4). The rotation BSS is based on the rotation of the distribution of observed signals. Consider that a human utterance and noise as shown in Fig. 1 (a) are exist, the mixture signals as shown in (b) are observed by two microphones.

Fig. 1: Source signals and mixture signals

Using the mixture signals, their joint distribution is shown in Fig. 2 (a), where the horizontal and the vertical axis 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 clarity the linear components, directions $\phi(t)$ of distribution are calculated as

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

Using all $\phi(t)$, its histogram as shown in Fig. 2 (b) means that the components of the joint distribution are concentrated 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

$$\theta = \arg \max_{\phi(t)} \text{hist}(\phi(t)). \quad (4)$$

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

Fig. 2: Joint distribution and its directional histogram

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

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

$$y(t) = x_2(t) \cos \theta - x_1(t) \sin \theta \quad (6)$$

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

3. Rotation BSS for Three Source Signals

In this section, an estimation method for source signals based on the rotation BSS in the case of three sources.

From the Eq (1), the observed mixture signals $x_m(t)$ ($m = 1, 2, 3$) are expressed as

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \quad (7)$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \quad (8)$$

$$x_3(t) = a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t) \quad (9)$$

When the two source signals are speaker speech and these have silent interval, we can estimate noise-reduced signals. The components of source signals without a silent interval appear as peaks in the histogram. In other words, the peak of the histogram represents the noise components.

We define ratios r_1 and r_2 of the observed signals as follows.

$$r_1 = \frac{x_2(t)}{x_1(t)} = \frac{a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t)}{a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t)} \quad (10)$$

$$r_2 = \frac{x_3(t)}{x_1(t)} = \frac{a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t)}{a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t)} \quad (11)$$

Assuming that source signals include a silent interval, the ratios r_1 can be regarded as $\frac{a_{21}}{a_{11}}$, $\frac{a_{22}}{a_{12}}$ or $\frac{a_{23}}{a_{13}}$. In the same way, r_2 can be regarded as $\frac{a_{31}}{a_{11}}$, $\frac{a_{32}}{a_{12}}$ or $\frac{a_{33}}{a_{13}}$.

At first, consider that the case which $s_1(t)$ is less silent interval than $s_2(t)$ and $s_3(t)$, the ratios are estimate $r_1 = \frac{a_{21}}{a_{11}}$

and $r_2 = \frac{a_{31}}{a_{11}}$. In this condition, the estimated signal $y_1(t)$ is generated signals without the source signal $s_1(t)$ as follows.

$$\begin{aligned} y_1(t) &= \cos\left(\tan^{-1} \frac{a_{21}}{a_{11}}\right) x_2(t) - \sin\left(\tan^{-1} \frac{a_{21}}{a_{11}}\right) x_1(t) \\ &= \frac{a_{11}}{l_1} (a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t)) \\ &\quad - \frac{a_{21}}{l_1} (a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t)) \\ &= \frac{a_{11}a_{22}}{l_1} s_2(t) + \frac{a_{11}a_{23}}{l_1} s_3(t) \\ &\quad - \frac{a_{12}a_{21}}{l_1} s_2(t) - \frac{a_{13}a_{21}}{l_1} s_3(t) \\ &= \frac{a_{11}a_{22} - a_{12}a_{21}}{l_1} s_2(t) + \frac{a_{11}a_{23} - a_{13}a_{21}}{l_1} s_3(t) \quad (12) \end{aligned}$$

where $l_1 = \sqrt{a_{11}^2 + a_{21}^2}$. The signal $y_2(t)$ is estimated as follows.

$$\begin{aligned} y_2(t) &= \cos\left(\tan^{-1} \frac{a_{31}}{a_{11}}\right) x_3(t) - \sin\left(\tan^{-1} \frac{a_{31}}{a_{11}}\right) x_1(t) \\ &= \frac{a_{11}a_{32} - a_{12}a_{31}}{l_2} s_2(t) + \frac{a_{11}a_{33} - a_{13}a_{31}}{l_2} s_3(t) \quad (13) \end{aligned}$$

where $l_2 = \sqrt{a_{11}^2 + a_{31}^2}$. From the Eqs.(12) and (13), the mixture signals without one source signal is obtained. These signals are mixture signals of two speaker speech because the components to be removed first are noise components. Eqs.(12) and (13) can be rewritten as follows.

$$y_1(t) = \alpha_{11}s_2(t) + \alpha_{12}s_3(t) \quad (14)$$

$$y_2(t) = \alpha_{21}s_2(t) + \alpha_{22}s_3(t) \quad (15)$$

The separated signal is estimated by Eqs.(3), (4) and (6) using the estimated signals $y_1(t)$ and $y_2(t)$. When $s_3(t)$ has more silent interval than $s_2(t)$, the separated signal is estimates as

$$\begin{aligned} y(t) &= \cos\left(\tan^{-1} \frac{\alpha_{21}}{\alpha_{11}}\right) y_2(t) - \sin\left(\tan^{-1} \frac{\alpha_{21}}{\alpha_{11}}\right) y_1(t) \\ &= \frac{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}}{l_3} s_3(t) \quad (16) \end{aligned}$$

where $l_3 = \sqrt{\alpha_{11}^2 + \alpha_{21}^2}$. When $s_2(t)$ has more silent interval than $s_3(t)$, $s_2(t)$ is estimated by Eq(16).

Next, consider the case which $s_2(t)$ is less silent interval than $s_1(t)$ and $s_3(t)$, r_1 become $\frac{a_{22}}{a_{12}}$ and r_2 become $\frac{a_{32}}{a_{12}}$. In the same way, we can calculate $y_1(t)$ and $y_2(t)$ as follows.

$$\begin{aligned} y_1(t) &= \cos\left(\tan^{-1} \frac{a_{22}}{a_{12}}\right) x_2(t) - \sin\left(\tan^{-1} \frac{a_{22}}{a_{12}}\right) x_1(t) \\ &= \frac{a_{12}a_{21} - a_{11}a_{22}}{l_4} s_1(t) + \frac{a_{12}a_{23} - a_{13}a_{22}}{l_4} s_3(t) \quad (17) \end{aligned}$$

$$\begin{aligned} y_2(t) &= \cos\left(\tan^{-1} \frac{a_{32}}{a_{12}}\right) x_3(t) - \sin\left(\tan^{-1} \frac{a_{32}}{a_{12}}\right) x_1(t) \\ &= \frac{a_{12}a_{31} - a_{11}a_{32}}{l_5} s_1(t) + \frac{a_{12}a_{33} - a_{13}a_{32}}{l_5} s_3(t) \quad (18) \end{aligned}$$

where $l_4 = \sqrt{a_{12}^2 + a_{22}^2}$ and $l_5 = \sqrt{a_{12}^2 + a_{32}^2}$. In the same way, the final separated signal is obtained in the case which $s_3(t)$ has more silent interval than $s_1(t)$ as

$$\begin{aligned} y(t) &= \cos\left(\tan^{-1} \frac{\beta_{21}}{\beta_{11}}\right) y_2(t) - \sin\left(\tan^{-1} \frac{\beta_{21}}{\beta_{11}}\right) y_1(t) \\ &= \frac{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}{l_6} s_3(t) \quad (19) \end{aligned}$$

where $l_6 = \sqrt{\beta_{11}^2 + \beta_{21}^2}$. When $s_1(t)$ has more silent interval than $s_3(t)$, $s_1(t)$ is estimated.

Finally, consider the case of $r_1 = \frac{a_{23}}{a_{13}}$ and $r_2 = \frac{a_{33}}{a_{13}}$. Similarly, we can calculate $y_1(t)$ and $y_2(t)$ as follows.

$$\begin{aligned} y_1(t) &= \cos\left(\tan^{-1} \frac{a_{23}}{a_{13}}\right) x_2(t) - \sin\left(\tan^{-1} \frac{a_{23}}{a_{13}}\right) x_1(t) \\ &= \frac{a_{13}a_{21} - a_{11}a_{23}}{l_7} s_1(t) + \frac{a_{13}a_{22} - a_{12}a_{23}}{l_7} s_2(t) \quad (20) \end{aligned}$$

$$\begin{aligned} y_2(t) &= \cos\left(\tan^{-1} \frac{a_{33}}{a_{13}}\right) x_3(t) - \sin\left(\tan^{-1} \frac{a_{33}}{a_{13}}\right) x_1(t) \\ &= \frac{a_{13}a_{31} - a_{11}a_{33}}{l_8} s_1(t) + \frac{a_{13}a_{32} - a_{12}a_{33}}{l_8} s_2(t) \quad (21) \end{aligned}$$

In this case which $s_2(t)$ has more silent interval than $s_1(t)$, the separated signal is follows.

$$\begin{aligned} y(t) &= \cos\left(\tan^{-1} \frac{\gamma_{21}}{\gamma_{11}}\right) y_2(t) - \sin\left(\tan^{-1} \frac{\gamma_{21}}{\gamma_{11}}\right) y_1(t) \\ &= \frac{\gamma_{11}\gamma_{22} - \gamma_{12}\gamma_{21}}{l_9} s_2(t) \quad (22) \end{aligned}$$

where $l_9 = \sqrt{\gamma_{11}^2 + \gamma_{21}^2}$. When $s_1(t)$ has more silent interval than $s_2(t)$, $s_1(t)$ is estimated.

Judging from the above, the estimated signals $y(t)$ are derived as the human source signals with silent interval. Since the proposed method removes the sound sources one by one, the method dose not need to extend to multidimensional space in the case of the sources increase.

4. Simulation

In order to verify our proposals, several simulations were carried out. Two sources were speaker speech data⁽⁵⁾ and one sources was car noise⁽⁶⁾.

The source signals and the mixture signals are shown in Fig. 3(a) and (b), respectively. The separated signal $y_1(t)$ using $x_1(t)$ and $x_2(t)$ by the proposed method is shown in (c). It is found that noise is removed in this waveform. In the same way, the separated signal $y_2(t)$ using $x_1(t)$ and $x_3(t)$ is shown in (d). Also in this case, it is found that noise is removed in this waveform. In addition to these, (e) is the waveform obtained using the estimated signals (c) and (d). This figure shows that

Fig. 3: Simulation results

the estimated signal $y(t)$ is the same waveform as the source signal $s_2(t)$. Consequently, we can extract the target speaker speech from three sound sources. The present result suggested that the source signals can be estimated when the sound source is multiple by using this process repeatedly.

5. Conclusion

We propose a new BSS method for three sound sources. This method based on the rotation BSS for two sound sources. The proposed method can extract the target speaker speech from three sound sources. Furthermore, our method dose not need to extend to multidimensional space when sources increase. Moreover, we can estimate the source signals by repeating simple processing. These results lead to the conclusion that the proposed method can estimate source signal when the number of sound sources increases. In future research, we apply our proposed method when the sound sources becomes multiple sources.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP60455178.

References

- (1) A. Cichocki and S. Amari: “Adaptive blind signal and image processing, learning algorithm and applications”, John Wiley & Sons, Ltd, 2002
- (2) S. Makino, T.-W. Lee and H. Sawada: “Blind speech separation”, Springer, 2007
- (3) 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 (NCSP ’14), pp. 21-24, 2014.
- (4) K. Higuchi, C. Okuma and T. Ishibashi, “Target sound extraction using silent interval of human speech”, International Symposium on Artificial Life and Robotics, AROB 22nd 201, pp. 900-903, 2017.
- (5) Acoustical Society of Japan: “ASJ continuous speech corpus Japanese newspaper article sentences”, JNAS Vols. 1-16, 1997
- (6) NTT Advanced Technology Corporation: “Ambient noise database for telephonometry 1996”, 1996