Separation and Dereverberation Performance of Frequency Domain Blind Source Separation for Speech in a Reverberant Environment

Ryo Mukai, Shoko Araki, Shoji Makino

NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{ryo,shoko,maki}@cslab.kecl.ntt.co.jp

Abstract

In this paper, we investigate the performance of an unmixing system obtained by frequency domain Blind Source Separation (BSS) based on Independent Component Analysis (ICA). Since ICA is based on statistics, *i.e.*, it only attempts to make outputs independent, it is not easy to predict what is going on in a BSS system. We therefore investigate the detailed components in the processed signals of a whole BSS system by measuring four impulse responses of the system. In particular, we focuse on the direct sound and reverberation in the target and jammer signals. As a result, we reveal that the direct sound and reverberation of the jammer can be reduced compared to a null beamformer (NBF), while the reverberation of the target cannot be reduced.

1. Introduction

Blind Source Separation (BSS) is a technique that separates and extracts target signals only from mixture signals observed without using information on the characteristics of the source signals and the acoustic system [1, 2]. There are several BSS algorithms such as Independent Component Analysis (ICA), which assumes the independency of source signals. Most BSS algorithms are considerably effective for instantaneous (non-convolutive) mixtures of signals, and some attempts have been made to apply BSS to signals mixed in convolutive environments [3, 4]. However, it has also been pointed out that a sufficient performance cannot be obtained in environments with a lot of reverberation [5, 6].

In this paper, we examine the performance of a separation system obtained by frequency domain BSS based on ICA. We focus our attention on the power of (1) the direct sound of the target signal, (2) the reverberation of the target signal, (3) the direct sound of the jammer signal, and (4) the reverberation of the jammer signal, and evaluate each power separately. As a result, it is shown that frequency domain BSS based on ICA can reduce the direct sound and reverberation of the jammer signal, while it does not remove the reverberation of the target signal.

2. Frequency domain BSS of convolutive mixtures

When the source signals are $s_i(1 \leq i \leq N)$, the signals observed by microphone j are $x_j(1 \leq j \leq M)$, and the unmixed signals are $y_i(1 \leq i \leq N)$, the model can be described by the following equations:

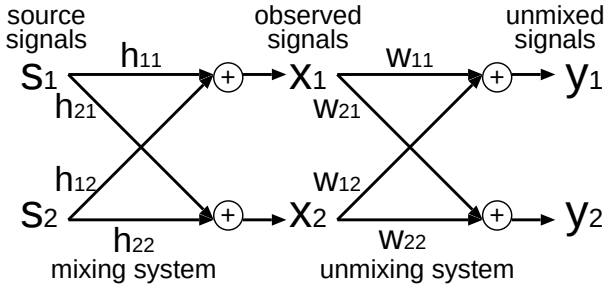

Figure 1: *Model of the BSS system*

$$x_j(t) = \sum_{i=1}^{N} \boldsymbol{h}_{ji} * s_i(t) \tag{1}$$

$$y_i(t) = \sum_{j=1}^{M} \boldsymbol{w}_{ij} * x_j(t), \tag{2}$$

where $\boldsymbol{h}_{ji}$ is the impulse response from source i to microphone j, $\boldsymbol{w}_{ij}$ is the coefficient when the unmixing system $\boldsymbol{W}$ is assumed as an FIR filter, and the $*$ is the convolution operator.

In this paper, we consider a two-input, two-output convolutive BSS problem, *i.e.*, $N = M = 2$ (Fig. 1). In addition, it is assumed that s_1 is separated to y_1, and s_2 is separated to y_2.

Because it is possible to convert a convolutive mixture in the time domain into an instantaneous mixture in the frequency domain, frequency domain BSS is effective for separating signals mixed in a reverberant environment.

Using a T-point short time Fourier transform for (1), we obtain

$$\boldsymbol{X}(\omega, m) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega, m). \tag{3}$$

We assume that the following separation has been completed in a frequency bin ω:

$$\boldsymbol{Y}(\omega, m) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, m), \tag{4}$$

where $\boldsymbol{X}(\omega) = [X_1(\omega), X_2(\omega)]^T$ is the observed signal in frequency bin ω, $\boldsymbol{Y}(\omega) = [Y_1(\omega), Y_2(\omega)]^T$ is the

estimated source signal, and $\boldsymbol{W}(\omega)$ represents the unmixing matrix. $\boldsymbol{W}(\omega)$ is determined so that $Y_1(\omega, m)$ and $Y_2(\omega, m)$ become mutually independent. The above calculations are carried out for each frequency independently.

For the calculation of unmixing matrix $\boldsymbol{W}$, we use an optimization algorithm based on the minimization of the Kullback-Leibler divergence [7, 8]. The optimal $\boldsymbol{W}$ is obtained by using the following iterative equation:

$$\boldsymbol{W}_{i+1} = \boldsymbol{W}_i + \eta[\mathrm{diag}(\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^H\rangle) - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^H\rangle]\boldsymbol{W}_i \quad (5)$$

where $\langle\cdot\rangle$ denotes the averaging operator, i is used to express the value of the i-th step in the iterations, and η is the step size parameter. In addition, we define the nonlinear function $\Phi(\cdot)$ as

$$\Phi(\boldsymbol{Y}) = \frac{1}{1+e^{-\mathrm{Re}(\boldsymbol{Y})}} + j\frac{1}{1+e^{-\mathrm{Im}(\boldsymbol{Y})}} \quad (6)$$

where $\mathrm{Re}(\boldsymbol{Y})$ and $\mathrm{Im}(\boldsymbol{Y})$ are the real and imaginary parts of $\boldsymbol{Y}$, respectively.

In general, it is necessary to solve the permutation problem and scaling problem when ICA is used. In our experiment, the effect of the permutation problem was negligible and so we did not coordinate the permutation. The problem of scaling was solved by adjusting the power of the target signal in the output signal to 0 dB.

3. Evaluation method

The performance of BSS is usually evaluated by the ratio of a target-originated signal to a jammer-originated signal. This measure is reasonable for evaluating the separation performance, but is unsuitable for evaluating the dereverberation performance because of its inability to distinguish the direct sound and reverberation. Since we want to know the detailed components in separated signals, *i.e.*, the direct sound and reverberation of the target and jammer, we take the following procedure,

(1) estimate unmixing matrix $\boldsymbol{W}(\omega)$ for each frequency.
(2) by using IFFT, transform frequency domain unmixing matrix $\boldsymbol{W}(\omega)$ to time domain unmixing filter $\boldsymbol{w}_{ij}(t)$.
(3) driving with the impulse as a source signal, measure four impulse responses, from s_1 to y_1, s_1 to y_2, s_2 to y_1, and s_2 to y_2.
(4) investigate the four impulse responses in detail and compare them to the responses of a null beamformer (NBF).

3.1. Definitions of performance measurement factors

We consider a separated signal y_1, target signal s_1, and jammer signal s_2. When the target s_1 is an impulse $\delta(t)$ and the jammer $s_2 = 0$, we call the observed signal x_1 as x_{1s1} [Fig. 2(a)], and y_1 as y_{1s1} [Fig. 2(b)]. Similarly, when $s_1 = 0$ and $s_2 = \delta(t)$, we call x_1 as x_{1s2}, and y_1 as y_{1s2} [Fig. 2(c)]. x_{1s1} is an impulse response from s_1 to x_1 by the mixing system $\boldsymbol{H}$, and y_{1s1} is an impulse response from s_1 to y_1 by the whole system $\boldsymbol{W}\cdot\boldsymbol{H}$. These are calculated by using $\boldsymbol{h}_{ij}$ and $\boldsymbol{w}_{ij}$ as follows.

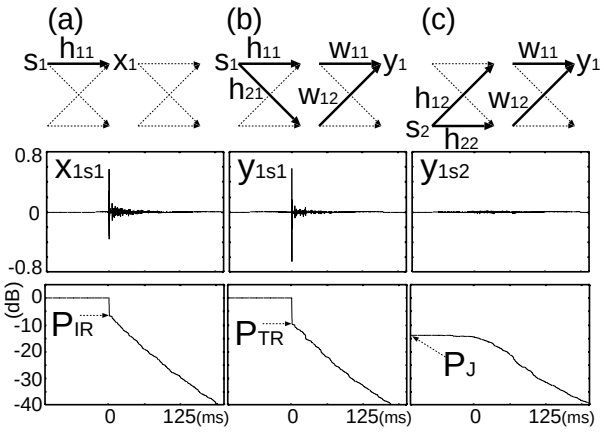

Figure 2: *Definitions of factors.*

$$x_{1s1} = \boldsymbol{h}_{11} \quad (7)$$
$$x_{1s2} = \boldsymbol{h}_{12} \quad (8)$$
$$y_{1s1} = \boldsymbol{w}_{11} * \boldsymbol{h}_{11} + \boldsymbol{w}_{12} * \boldsymbol{h}_{21} \quad (9)$$
$$y_{1s2} = \boldsymbol{w}_{11} * \boldsymbol{h}_{12} + \boldsymbol{w}_{12} * \boldsymbol{h}_{22} \quad (10)$$

From the viewpoint of source separation, we can consider y_{1s1} as the direct and reverberant sound of target s_1, and y_{1s2} as the remaining sound of jammer s_2.

To simplify the evaluation, we normalize $\boldsymbol{h}_{ji}$ so that the power of the observed signals x_{1s1} and x_{1s2} is equal to 0 dB, and make the following definitions (Fig. 2).

- P_{IR}: the power of the reverberant sound in x_{1s1},
- P_{TR}: the power of the reverberant sound in y_{1s1},
- P_J: the power of y_{1s2}.

We also define the reduction of the reverberation of target signal R_T and the reduction of jammer signal R_J as follows

$$R_T = -(P_{TR} - P_{IR}) \quad (11)$$
$$R_J = -P_J. \quad (12)$$

4. Experiments

In order to examine what is separated by an unmixing system based on ICA, and what remains as noise, we investigated impulse responses of the system. In frequency domain BSS, it has been confirmed that the separation performance changes according to the length of the frame [6], so we chose the frame length and the frame shift as parameters.

4.1. Conditions for the experiments

The layout of the room we used to measure the impulse responses of the mixing system $\boldsymbol{H}$ is shown in Fig. 3. The reverberation time of the room was 300 ms. We used a two-element array with inter-element spacing of 4 cm.

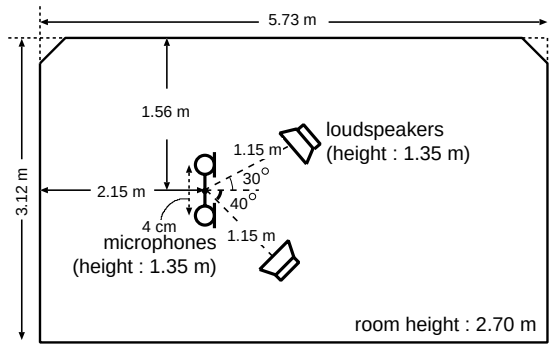

Figure 3: *Layout of the room used in experiments. Reverberation time = 300 ms.*

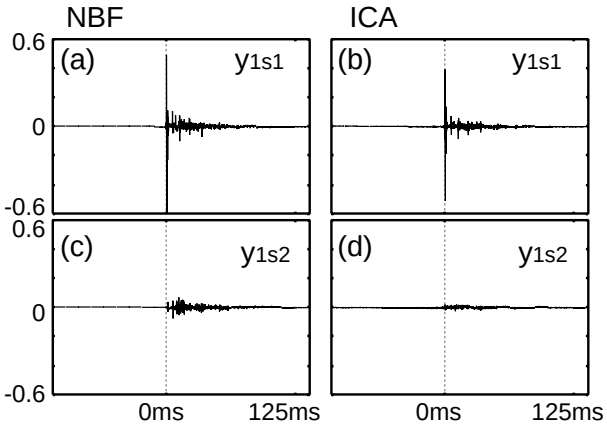

Figure 4: *Target and jammer impulse responses of NBF and ICA*

The speech signals arrived from two directions, *i.e.*, -30° and 40°. The contribution of the direct sound of $\boldsymbol{h}_{11}$ and $\boldsymbol{h}_{21}$ was 6.6 dB, that of $\boldsymbol{h}_{12}$ and $\boldsymbol{h}_{22}$ was 5.7 dB.

Two sentences spoken by two male speakers selected from the ASJ continuous speech corpus for research were used as the source signals. The lengths of these mixed speech signals were about eight seconds each. We used the entire eight seconds of the mixed data for learning according to (5).

In these experiments, we changed the frame length T from 32 to 4096 and investigated the performance for each condition. The sampling rate was 8 kHz, and the analysis window was a Hamming window. The frame shift S was $T/2$ and $T/32$, which correspond to double and 32 times oversampling.

The number of iterations for (5) was 100, except when $S = T/2$ and $T = 1024$, 2048, and 4096, where the iteration was stopped at 70, 30, and 20, respectively, because deterioration of the performance was observed.

4.2. Experimental results

Figures 4(a) and (c) show examples of impulse responses y_{1s1} and y_{1s2} of the unmixing system obtained by a null beamformer (NBF) that forms a steep null directivity pattern towards a jammer under the assumption of the jammer's direction being known. Figures 4(b) and (d) are results obtained by ICA.

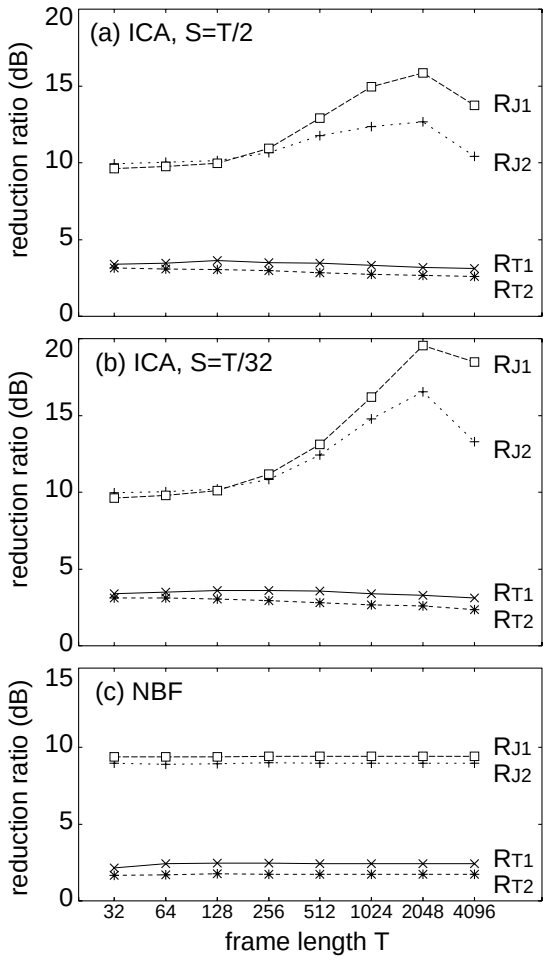

Figure 5: *Relationship between frame length and reduction ratio.*

For the target signal, we can see that the reverberation passes the system in both cases (NBF and ICA) in Figs. 4(a) and (b). Figure 4(c) shows that the direct sound of the jammer is removed, but the reverberation is not removed by NBF. This was expected. On the other hand, Fig. 4(d) indicates that ICA not only removes the direct sound, but also reduces the reverberation of the jammer.

Figure 5 shows the relationship between the frame length T and the reduction ratios R_T and R_J defined by (11) and (12). R_{T1} and R_{J1} are R_T and R_J when the target signal is s_1. R_{T2} and R_{J2} are results when the target signal is s_2. Figures 5(a) and (b) show results by ICA when $S = T/2$ and $S = T/32$, respectively. For the sake of comparison, the performance of NBF is shown in Fig. 5(c).

Note that these results are measured by the power of impulse responses, and different from the noise reduction rate (NRR) [6] measured by using a speech signal having a highly colored spectrum. Our results indicate seemingly better values than the NRR of the speech signal.

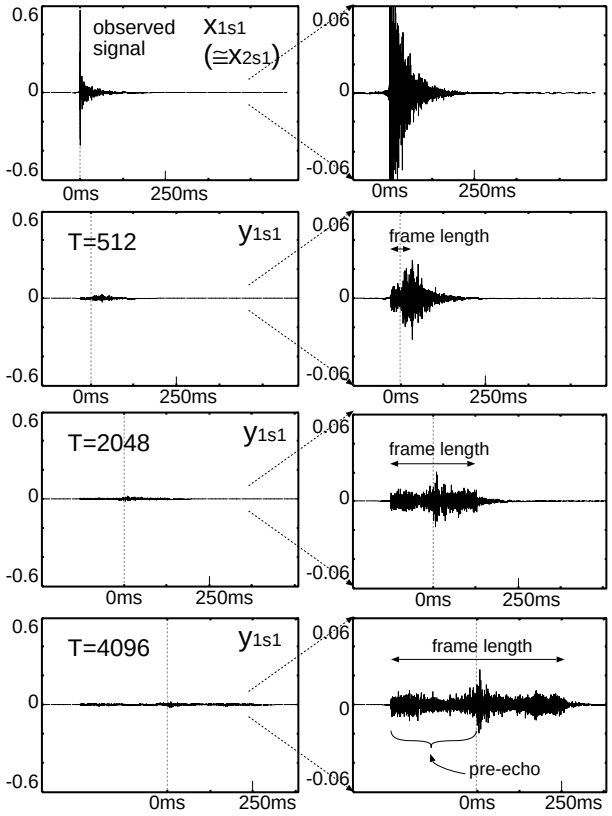

Figure 6: *Jammer impulse response of a BSS system*

5. Discussion

First, we discuss the jammer reduction ratio R_J. When $T \leq 128$, the reduction performance of BSS is as poor as that of NBF, and when $256 \leq T \leq 2048$, the reduction ratio increases. In the case of $T = 2048$, $S = T/32$, R_{J1}=19.5 dB, R_{J2}=16.6 dB. This is far greater than the contribution of the direct sound, *i.e.*, 6.6 dB and 5.7 dB. This means that the system of ICA can reduce not only the direct sound of the jammer but also the reverberant sound of the jammer. In addition, comparing the results of S=$T/2$ and S=$T/32$ [Figs. 5(a) and (b)], we can see that oversampling improves the jammer reduction ratio.

On the other hand, the reduction ratio of the reverberation of target R_T is low, and does not vary through the entire frame length T. This means that dereverberation was not achieved for the target signal.

From these results, it should be concluded that $\boldsymbol{W}$ is not the approximation of the inverse system of $\boldsymbol{H}$, but a filter that can eliminate the jammer signal [9].

Finally, we show the reason why the reduction ratio of jammer signal R_J declines when T is too long. Figure 6 shows the jammer signal's impulse response y_{1s2}, when T = 512, 2048, and 4096. The best performance is obtained when T = 2048. In the case of T = 512, the length of the unmixing system is much shorter than the length of the reverberation, accordingly, the reverberation longer than the frame cannot be reduced at all. On the other hand, when T = 4096, which is longer than the reverberant time, the unmixing system can wholly cover the reverberation, but because each tap of the filter has errors that derive from the statistical method of ICA, the amount of errors increases. Moreover, the pre-echo noise also increases, and this causes the poor performance.

6. Conclusion

We investigated the performance of an unmixing system obtained by frequency domain BSS based on ICA using the impulse responses of target and jammer signals.

As a result, we revealed that ICA not only removes the direct sound of the jammer signal, but also reduces the reverberation, while the reverberation of the target is not reduced.

The jammer reduction performance increases as the frame length becomes longer. However, an overly long frame length decreases the performance due to accumulating errors. The performance of the target reverberation reduction does not depend on the frame length and is as poor as that of NBF.

7. Acknowledgements

We would like to thank Dr. Hiroshi Saruwatari for his valuable discussions. We also thank Dr. Shigeru Katagiri for his continuous encouragement.

8. References

[1] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, Vol. 7, No. 6, pp. 1129–1159, 1995.

[2] S. Haykin, ed., *Unsupervised adaptive filtering*, John Wiley & Sons, 2000.

[3] T. W. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real world signals," *Neural Networks*, Vol. 4, pp. 2129–2134, 1997.

[4] J. Xi and J. P. Reilly, "Blind separation and restoration of signals mixed in convolutive environment," in *Proc. ICASSP97*, pp. 1327–1330, 1997.

[5] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proc. ICASSP2000*, pp. 1041–1044, 2000.

[6] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," in *Proc. ICASSP2001*, 2001.

[7] S. Ikeda and N. Murata, "An information-maximization approach to blind separation and blind deconvolution," in *Proc. ICA99*, pp. 365–370, 1999.

[8] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. ICASSP2000*, pp. 3140–3143, 2000.

[9] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," in *Proc. Eurospeech2001*, 2001, (submitted).