Equivalence between Frequency Domain Blind Source Separation and Frequency Domain Adaptive Beamformers

Shoko Araki † *Shoji Makino* † *Ryo Mukai* † *Hiroshi Saruwatari* ‡

† NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: shoko@cslab.kecl.ntt.co.jp
‡ Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara 630-0101, Japan

Abstract

Frequency domain Blind Source Separation (BSS) is shown to be equivalent to two sets of frequency domain adaptive microphone arrays, *i.e.,* Adaptive Beamformers (ABF). The minimization of the off-diagonal components in the BSS update equation can be viewed as the minimization of the mean square error in the ABF. The unmixing matrix of the BSS and the filter coefficients of the ABF converge to the same solution in the mean square error sense if the two source signals are ideally independent. Therefore, we can conclude that the performance of the BSS is upper bounded by that of the ABF. This understanding clearly explains the poor performance of the BSS in a real room with long reverberation.

1. Introduction

Blind Source Separation (BSS) is an approach to estimate source signals $s_i(t)$ using only the information of mixed signals $x_j(t)$ observed in each input channel. This technique is applicable to the achievement of noise robust speech recognition and high-quality hands-free telecommunication systems. It might also become one of the cues for auditory scene analysis.

To achieve the BSS of convolutive mixtures, several methods have been proposed [1, 2]. In this paper, we consider the BSS of convolutive mixtures of speech in the frequency domain [3, 4], for the sake of mathematical simplicity and the reduction of computational complexity.

Signal separation by using a noise cancellation framework with signal leakage into the noise reference was discussed in [5, 6]. It was shown that the least squares criterion is equivalent to the decorrelation criterion of a noise free signal estimate and a signal free noise estimate. The error minimization was shown to be completely equivalent with a zero search in the crosscorrelation.

Inspired by their discussions, but apart from the noise cancellation framework, we attempt to see the frequency domain BSS problem with a frequency domain adaptive microphone array, *i.e.*, Adaptive Beamformer (ABF) frameworks. The equivalence and difference between the BSS and ABF are discussed theoretically.

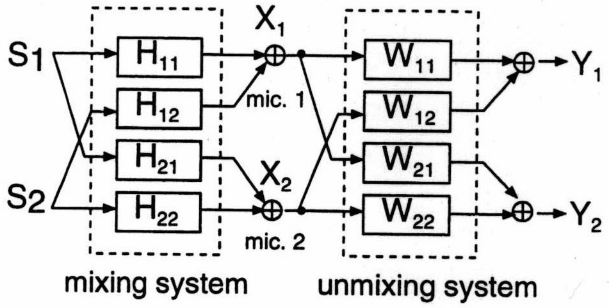


Figure 1: *BSS system configuration.*

2. Frequency domain BSS of convolutive mixtures of speech

The signals recorded by M microphones are given by

$$x_j(n) = \sum_{i=1}^{N}\sum_{p=1}^{P} h_{ji}(p)s_i(n-p+1) \quad (j=1,\cdots,M), \tag{1}$$

where s_i is the source signal from a source i, x_j is the received signal by a microphone j, and h_{ji} is the P-point impulse response from a source i to a microphone j. In this paper, we consider a two-input, two-output convolutive BSS problem, *i.e.*, $N = M = 2$ (Fig. 1).

The frequency domain approach to the convolutive mixture is to transform the problem into an instantaneous BSS problem in the frequency domain [3, 4]. Using a T-point short time Fourier transform for (1), we obtain

$$\boldsymbol{X}(\omega,m) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega,m), \tag{2}$$

where $\boldsymbol{S}(\omega,m) = [S_1(\omega,m), S_2(\omega,m)]^T$. We assume that a (2×2) mixing matrix $\boldsymbol{H}(\omega)$ is invertible, and $H_{ji}(\omega) \neq 0$.

The unmixing process can be formulated in a frequency bin ω:

$$\boldsymbol{Y}(\omega,m) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega,m), \tag{3}$$

where $\boldsymbol{X}(\omega,m) = [X_1(\omega,m), X_2(\omega,m)]^T$ is the observed signal at frequency bin ω, $\boldsymbol{Y}(\omega,m) = [Y_1(\omega,m), Y_2(\omega,m)]^T$ is the estimated source signal, and $\boldsymbol{W}(\omega)$ represents a (2×2) 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 at each frequency independently.

2.1. Frequency domain BSS of convolutive mixtures using Second Order Statistics (SOS)

It is well known that the decorrelation criterion is insufficient to solve the problem. In [6], however, it is pointed out that non-stationary signals provide enough additional information to estimate all W_{ij}. Some authors have utilized the SOS for mixed speech signals [7, 8].

The source signals $S_1(\omega,m)$ and $S_2(\omega,m)$ are assumed to be zero mean and mutually uncorrelated, that is,

$$\begin{aligned}\boldsymbol{R}_S(\omega,k) &= \frac{1}{M}\sum_{m=0}^{M-1}\boldsymbol{S}(\omega,Mk+m)\boldsymbol{S}^*(\omega,Mk+m)\\ &= \boldsymbol{\Lambda}_s(\omega,k),\end{aligned} \tag{4}$$

where $*$ denotes the conjugate transpose, and $\boldsymbol{\Lambda}_s(\omega,k)$ is a different diagonal matrix for each k.

In order to determine $\boldsymbol{W}(\omega)$ so that $Y_1(\omega,m)$ and $Y_2(\omega,m)$ become mutually uncorrelated, we seek a $\boldsymbol{W}(\omega)$ that diagonalizes the covariance matrices $\boldsymbol{R}_Y(\omega,k)$ simultaneously for all k,

$$\begin{aligned}\boldsymbol{R}_Y(\omega,k) &= \boldsymbol{W}(\omega)\boldsymbol{R}_X(\omega,k)\boldsymbol{W}^*(\omega)\\ &= \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)\boldsymbol{\Lambda}_s(\omega,k)\boldsymbol{H}^*(\omega)\boldsymbol{W}^*(\omega)\\ &= \boldsymbol{\Lambda}_c(\omega,k),\end{aligned} \tag{5}$$

where $\boldsymbol{R}_X$ is the covariance matrix of $\boldsymbol{X}(\omega)$ as follows,

$$\boldsymbol{R}_X(\omega,k) = \frac{1}{M}\sum_{m=0}^{M-1}\boldsymbol{X}(\omega,Mk+m)\boldsymbol{X}^*(\omega,Mk+m), \tag{6}$$

and $\boldsymbol{\Lambda}_c(\omega,k)$ is an arbitrary diagonal matrix.

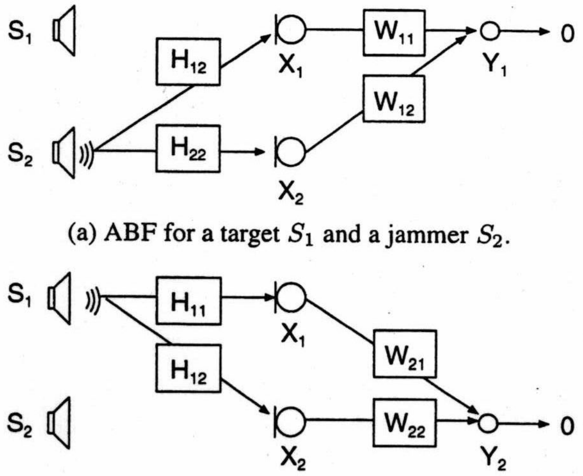
The diagonalization of $\boldsymbol{R}_Y(\omega,k)$ can be written as an overdetermined least-square problem,

$$\begin{aligned}&\arg\min_{\boldsymbol{W}(\omega)}\sum_k ||\text{off-diag}\boldsymbol{W}(\omega)\boldsymbol{R}_X(\omega,k)\boldsymbol{W}^*(\omega)||^2 \quad (7)\\ &s.t., \sum_k \text{diag}||\boldsymbol{W}(\omega)\boldsymbol{R}_X(\omega,k)\boldsymbol{W}^*(\omega)||^2 \neq 0,\end{aligned}$$

where $||\boldsymbol{x}||^2$ is the squared Frobenius norm.

3. Frequency domain adaptive beamformer

Here, we consider the frequency domain adaptive beamformer (ABF), which forms a null directivity pattern towards a jammer. Since our aim is to separate two signals S_1 and S_2 with two microphones, two sets of ABF are used (Fig. 2), that is, an ABF that forms a null directivity pattern towards source S_2 by using filter coefficients W_{11} and W_{12}, and an ABF that forms a null directivity pattern towards source S_1 by using filter coefficients W_{21} and W_{22}. Note that an ABF can be adapted when only a jammer exists but a target does not exist.



(a) ABF for a target S_1 and a jammer S_2.

(b) ABF for a target S_2 and a jammer S_1.

Figure 2: *Two sets of ABF system configurations.*

3.1. ABF null towards S_2

First, we consider the case of target S_1 and jammer S_2 [Fig. 2(a)]. When target $S_1 = 0$, output $Y_1(\omega,m)$ is expressed as

$$\begin{aligned}Y_1(\omega,m) &= W_{11}(\omega)X_1(\omega,m) + W_{12}(\omega)X_2(\omega,m)\\ &= \boldsymbol{W}(\omega)\boldsymbol{X}(\omega,m),\end{aligned} \tag{8}$$

where

$\boldsymbol{W}(\omega) = [W_{11}(\omega), W_{12}(\omega)], \boldsymbol{X}(\omega,m) = [X_1(\omega,m), X_2(\omega,m)]^T$.

To minimize jammer $S_2(\omega,m)$ in output $Y_1(\omega,m)$ when target $S_1 = 0$, mean square error $J(\omega)$ is introduced as

$$\begin{aligned}J(\omega) &= E[Y_1^2(\omega,m)]\\ &= \boldsymbol{W}(\omega)E[\boldsymbol{X}(\omega,m)\boldsymbol{X}^*(\omega,m)]\boldsymbol{W}^*(\omega)\\ &= \boldsymbol{W}(\omega)\mathbf{R}(\omega)\boldsymbol{W}^*(\omega),\end{aligned} \tag{9}$$

where E is the expectation and

$$\mathbf{R}(\omega) = E\begin{bmatrix} X_1(\omega,m)X_1^*(\omega,m) & X_1(\omega,m)X_2^*(\omega,m)\\ X_2(\omega,m)X_1^*(\omega,m) & X_2(\omega,m)X_2^*(\omega,m)\end{bmatrix}. \tag{10}$$

By differentiating cost function $J(\omega)$ with respect to $\boldsymbol{W}$ and setting the gradient equal to zero

$$\frac{\partial J(\omega)}{\partial \boldsymbol{W}} = 2\mathbf{R}\boldsymbol{W}^* = 0, \tag{11}$$

we obtain the equation to solve as follows [(ω,m), etc., are omitted for convenience],

$$E\begin{bmatrix} X_1X_1^* & X_1X_2^*\\ X_2X_1^* & X_2X_2^*\end{bmatrix}\begin{bmatrix} W_{11}^*\\ W_{12}^*\end{bmatrix} = \begin{bmatrix} 0\\ 0\end{bmatrix}, \tag{12}$$

or in a separate formula

$$E[X_1X_1^*]W_{11}^* + E[X_1X_2^*]W_{12}^* = 0 \tag{13}$$

$$E[X_2X_1^*]W_{11}^* + E[X_2X_2^*]W_{12}^* = 0. \tag{14}$$

Using $X_1 = H_{12}S_2$, $X_2 = H_{22}S_2$, we get

$$W_{11}H_{12} + W_{12}H_{22} = 0. \tag{15}$$

With (15) only, we have trivial solution W_{11}=W_{12}=0. Therefore, an additional constraint should be added to ensure target signal S_1 in output Y_1. With this constraint, output Y_1 is expressed as

$$\begin{aligned} Y_1 &= W_{11}X_1 + W_{12}X_2 \\ &= W_{11}H_{11}S_1 + W_{12}H_{21}S_1 = c_1S_1, \end{aligned} \tag{16}$$

which leads to

$$W_{11}H_{11} + W_{12}H_{21} = c_1, \tag{17}$$

where c_1 is an arbitrary complex constant. Since H_{12} and H_{22} are unknown, the minimization of (9) with adaptive filters W_{11} and W_{12} is used to derive (15) with constraint (17). This means that the ABF solution is derived from simultaneous equations (15) and (17).

3.2. ABF null towards S_1

Similarly for target S_2, jammer S_1, and output Y_2 [Fig. 2(b)], we obtain

$$W_{21}H_{11} + W_{22}H_{21} = 0 \tag{18}$$

$$W_{21}H_{12} + W_{22}H_{22} = c_2. \tag{19}$$

3.3. Two sets of ABF

By combining (15), (17), (18), and (19), the simultaneous equations for two sets of ABF are summarized as

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}. \tag{20}$$

4. Equivalence between Blind Source Separation and Adaptive Beamformers

As we showed in (7), the SOS BSS algorithm works to minimize off-diagonal components in

$$E\begin{bmatrix} Y_1Y_1^* & Y_1Y_2^* \\ Y_2Y_1^* & Y_2Y_2^* \end{bmatrix}, \tag{21}$$

[see (5)]. Using $\boldsymbol{H}$ and $\boldsymbol{W}$, outputs Y_1 and Y_2 are expressed in each frequency bin as follows,

$$Y_1 = aS_1 + bS_2 \tag{22}$$

$$Y_2 = cS_1 + dS_2, \tag{23}$$

where

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}. \tag{24}$$

4.1. When $S_1 \neq 0$ and $S_2 \neq 0$

We now analyze what is going on in the BSS framework. After convergence, the expectation of the off-diagonal component $E[Y_1Y_2^*]$ is expressed as

$$\begin{aligned} &E[Y_1Y_2^*] \\ &= ad^*E[S_1S_2^*] + bc^*E[S_2S_1^*] + (ac^*E[S_1^2] + bd^*E[S_2^2]) \\ &= 0. \end{aligned} \tag{25}$$

Since S_1 and S_2 are assumed to be uncorrelated, the first term and the second term become zero. Then, the BSS adaptation should drive the third term of (25) to be zero. By squaring the third term and setting it equal to zero

$$\begin{aligned} &(ac^*E[S_1^2] + bd^*E[S_2^2])^2 \\ &= a^2c^2(E[S_1^2])^2 + 2abc^*d^*E[S_1^2]E[S_2^2] + b^2d^2(E[S_2^2])^2 \\ &= 0 \end{aligned} \tag{26}$$

(26) is equivalent to

$$ac^* = bd^* = 0, \quad abc^*d^* = 0. \tag{27}$$

<u>**CASE 1:** $a = c_1, c = 0, b = 0, d = c_2$</u>

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix} \tag{28}$$

This equation is exactly the same as that of the ABF (20).

<u>**CASE 2:** $a = 0, c = c_1, b = c_2, d = 0$</u>

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} 0 & c_2 \\ c_1 & 0 \end{bmatrix} \tag{29}$$

This equation leads to the permutation solution which is $Y_1 = c_2S_2, Y_2 = c_1S_1$.

<u>**CASE 3:** $a = 0, c = c_1, b = 0, d = c_2$</u>

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ c_1 & c_2 \end{bmatrix} \tag{30}$$

This equation leads to undesirable solution $Y_1 = 0, Y_2 = c_1S_1 + c_2S_2$.

<u>**CASE 4:** $a = c_1, c = 0, b = c_2, d = 0$</u>

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & c_2 \\ 0 & 0 \end{bmatrix} \tag{31}$$

This equation leads to undesirable solution $Y_1 = c_1S_1 + c_2S_2, Y_2 = 0$.

Note that CASE 3 and CASE 4 do not appear in general since we assume that $\boldsymbol{H}(\omega)$ is invertible, and $H_{ji}(\omega) \neq 0$. That is, if $a = 0$ then $b \neq 0$ (CASE 2), and if $c = 0$ then $d \neq 0$ (CASE 1).

If the uncorrelated assumption between $S_1(\omega)$ and $S_2(\omega)$ collapses, the first and second terms of (25) become the bias noise to get the correct coefficients a, b, c, d.

4.2. When $S_1 \neq 0$ and $S_2 = 0$

The BSS can adapt, even if there is only one active source. In this case, only one set of ABF is achieved.

When $S_2 = 0$, we have

$$Y_1 = aS_1 \text{ and } Y_2 = cS_1 \tag{32}$$

then

$$E[Y_1Y_2^*] = E[aS_1c^*S_1^*] = ac^*E[S_1^2] = 0, \tag{33}$$

and therefore, the BSS adaptation should drive

$$ac^* = 0. \tag{34}$$

CASE 5: $c = 0, a = c_1$

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & * \\ 0 & * \end{bmatrix}, \tag{35}$$

where $*$ shows a don't care. Since $S_2 = 0$, the output can be derived correctly $Y_1 = c_1S_1, Y_2 = 0$ as follows.

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} c_1 & * \\ 0 & * \end{bmatrix} \begin{bmatrix} S_1 \\ 0 \end{bmatrix} = \begin{bmatrix} c_1S_1 \\ 0 \end{bmatrix} \tag{36}$$

CASE 6: $c = c_1, a = 0$

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} 0 & * \\ c_1 & * \end{bmatrix} \tag{37}$$

This equation leads to the permutation solution which is $Y_1 = 0, Y_2 = c_1S_1$.

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 0 & * \\ c_1 & * \end{bmatrix} \begin{bmatrix} S_1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ c_1S_1 \end{bmatrix} \tag{38}$$

4.3. Fundamental limitation of frequency domain BSS

Frequency domain BSS and frequency domain ABF are shown to be equivalent [see equations (20) and (28)] if the independent assumption ideally holds [see equation (25)]. Moreover, we have shown in [9], that a long frame size works poorly in frequency domain BSS for speech data of a few seconds, because the assumption of independency between $S_1(\omega)$ and $S_2(\omega)$ does not hold in each frequency. Therefore, the performance of the BSS is upper bounded by that of the ABF.

We can form only one null towards the jammer in the case of two microphones. Although the directivity pattern becomes duller when there is a long reverberation, the BSS and ABF mainly remove the sound from the jammer direction. This understanding clearly explains the poor performance of the BSS in a real room with long reverberation.

The BSS was shown to outperform a null beamformer that forms a steep null directivity pattern towards a jammer under the assumption of the jammer's direction being known [10, 11]. It is well known that an adaptive beamformer outperforms a null beamformer in long reverberation. Our understanding also clearly explains the result.

Our discussion here is essentially also true for the BSS with Higher Order Statistics (HOS), and will be extended to it shortly.

5. Conclusion

Frequency domain Blind Source Separation (BSS) is shown to be equivalent to two sets of frequency domain adaptive beamformers (ABF). The unmixing matrix of the BSS and the filter coefficients of the ABF converge to the same solution in the mean square error sense if the two source signals are ideally independent. Therefore, we can conclude that the performance of the BSS is upper bounded by that of the ABF. This understanding clearly explains the poor performance of the BSS in a real room with long reverberation. The fundamental difference exists in the adaptation period when they should adapt. That is, the ABF can adapt in the presence of a jammer but the absence of a target, whereas the BSS can adapt in the presence of a target and jammer, and also in the presence of only a target.

6. Acknowledgements

We would like to thank Dr. Shigeru Katagiri and Dr. Kiyohiro Shikano for their continuous encouragement.

7. 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, "Unsupervised adaptive filtering," John Wiley & Sons, 2000.

[3] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," *Proc. ICA99*, pp. 365-370, Jan. 1999.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21-34, 1998.

[5] S. V. Gerven and D. V. Compernolle, "Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness," *IEEE Trans. Speech Audio Processing*, vol. 43, no. 7, pp. 1602-1612, July 1995.

[6] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 405-413, Oct. 1993.

[7] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320-327, May 2000.

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

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

[10] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," *Proc. ICASSP2001*, May 2001.

[11] R. Mukai, S. Araki, and S. Makino, "Separation and dereverberation performance of frequency domain blind source separation for speech in a reverberant environment," *Proc. Eurospeech2001*, Sept. 2001.