EQUIVALENCE BETWEEN FREQUENCY DOMAIN BLIND SOURCE SEPARATION AND FREQUENCY DOMAIN ADAPTIVE BEAMFORMING

Shoko Araki † *Shoji Makino* † *Ryo Mukai* †
Yoichi Hinamoto ‡ *Tsuyoki Nishikawa* ‡ *Hiroshi Saruwatari* ‡

† NTT Communication Science Laboratories, NTT Corporation
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 (ABFs). 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, the performance of the BSS is limited by that of the ABF. This understanding gives an interpretation of BSS from physical point of view.

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 at each input channel. BSS is applicable to the achievement of noise robust speech recognition and high-quality hands-free telecommunication. 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]. In this paper, we consider the BSS of convolutive mixtures of speech in the frequency domain [2].

In earlier works, Kurita et al. [3] and Parra et al. [4] utilized the relationship between BSS and Adaptive Beamformers (ABF) to achieve a better performance of BSS. However, they did not discuss this relationship theoretically.

Signal separation by using a noise cancellation framework with signal leakage into the noise reference was discussed in [5, 6]. This study showed 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) framework. The equivalence and differences between the BSS and ABF are discussed.

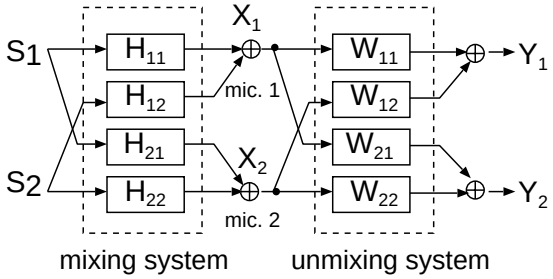


Figure 1: BSS system configuration.

2. FREQUENCY DOMAIN BSS OF CONVOLUTIVE MIXTURES OF SPEECH

In this paper, $\boldsymbol{S}(\omega,m) = [S_1(\omega,m),\cdots,S_N(\omega,m)]^T$, $\boldsymbol{X}(\omega,m) = [X_1(\omega,m),\cdots,X_M(\omega,m)]^T$, and $\boldsymbol{Y}(\omega,m) = [Y_1(\omega,m),\cdots,Y_N(\omega,m)]^T$ are the time-frequency representations of the source signals, observed signals and output signals (estimated source signals) respectively, which are obtained by frame-by-frame discrete Fourier transform (DFT). ω is the frequency index and m denotes the position of the frame with width T. We consider a two-input, two-output convolutive BSS problem, *i.e.,* $N = M = 2$ (see Fig. 1) without a loss of generality.

In frequency domain BSS [2], the separation is performed using only the information of observed signals $\boldsymbol{X}(\omega,m) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega,m)$, under the assumption that the source signals are mutually independent in each frequency bin ω. Here, $\boldsymbol{H}(\omega)$ is a (2×2) mixing matrix comprising components $H_{ji}(\omega)$, which are Fourier transforms of the P-point impulse responses from a source i to a microphone j. We assume that $\boldsymbol{H}(\omega)$ is invertible, and $H_{ji}(\omega) \neq 0$.

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

$$\boldsymbol{Y}(\omega,m) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega,m), \tag{1}$$

where $\boldsymbol{W}(\omega)$ 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)

A decorrelation criterion is sufficient to estimate all W_{ij} for non-stationary signals [6]. Previously, [7] and [8] utilized the SOS for mixed speech signals.

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 time blocks k,

$$\begin{aligned} \boldsymbol{R}_Y(\omega,k) &= \boldsymbol{W}(\omega)\boldsymbol{R}_X(\omega,k)\boldsymbol{W}^*(\omega) \\ &= \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)\boldsymbol{\Lambda}_s(\omega,k)\boldsymbol{H}^*(\omega)\boldsymbol{W}^*(\omega) \\ &= \boldsymbol{\Lambda}_c(\omega,k), \end{aligned} \quad (2)$$

where $*$ denotes the conjugate transpose, $\boldsymbol{R}_X$ is the covariance matrix of $\boldsymbol{X}(\omega)$, *i.e.*,

$$\boldsymbol{R}_X(\omega,k) = \frac{1}{M}\sum_{m=0}^{M-1} \boldsymbol{X}(\omega, Mk+m)\boldsymbol{X}^*(\omega, Mk+m), \quad (3)$$

$\boldsymbol{\Lambda}_s(\omega, k)$ is the covariance matrix of the source signals, which is a different diagonal matrix for each k, 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-squares problem,

$$\begin{aligned} &\underset{W(\omega)}{\operatorname{argmin}} \sum_k ||\text{off-diag}\boldsymbol{W}(\omega)\boldsymbol{R}_X(\omega,k)\boldsymbol{W}^*(\omega)||^2 \quad (4) \\ &\text{subject to } \sum_k \text{diag}||\boldsymbol{W}(\omega)\boldsymbol{R}_X(\omega,k)\boldsymbol{W}^*(\omega)||^2 \neq 0, \end{aligned}$$

where $||\cdot||^2$ is the squared Frobenius norm.

3. FREQUENCY DOMAIN ADAPTIVE BEAMFORMER

Here, we consider the frequency domain adaptive beamformer (ABF), which can remove a jammer signal. Since our aim is to separate two signals S_1 and S_2 with two microphones, we use two sets of ABFs (see Fig. 2). Note that the ABF can be adapted when only a jammer exists but a target does not exist, and that the direction of the target or the impulse responses from the target to microphones should be known.

3.1. ABF for a target S_1 and a jammer S_2

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

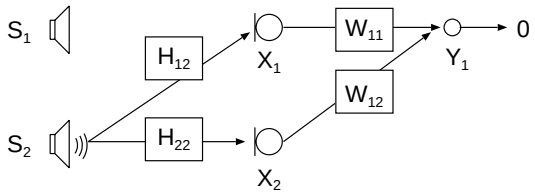
$$Y_1(\omega,m) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega,m), \quad (5)$$

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

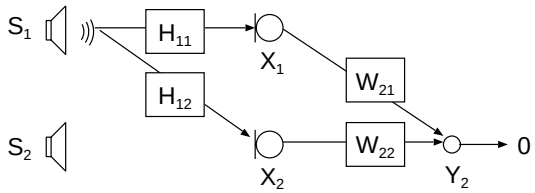
To minimize jammer $S_2(\omega, m)$ in output $Y_1(\omega, m)$ when target $S_1 = 0$, mean square error $J(\omega)$ is introduced as

$$\begin{aligned} J(\omega) &= E[Y_1^2(\omega,m)] \\ &= \boldsymbol{W}(\omega)E[\boldsymbol{X}(\omega,m)\boldsymbol{X}^*(\omega,m)]\boldsymbol{W}^*(\omega) \\ &= \boldsymbol{W}(\omega)\mathbf{R}(\omega)\boldsymbol{W}^*(\omega), \end{aligned} \quad (6)$$

where $E[\cdot]$ is the expectation operator and



(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.

$$\mathbf{R}(\omega) = E\begin{bmatrix} X_1(\omega,m)X_1^*(\omega,m) & X_1(\omega,m)X_2^*(\omega,m) \\ X_2(\omega,m)X_1^*(\omega,m) & X_2(\omega,m)X_2^*(\omega,m) \end{bmatrix}. \quad (7)$$

By differentiating the cost function $J(\omega)$ with respect to $\boldsymbol{W}$ and setting the gradient equal to zero, we obtain [hereafter (ω, m) and (ω) are omitted for convenience],

$$\frac{\partial J(\omega)}{\partial \boldsymbol{W}} = 2\mathbf{R}\boldsymbol{W}^* = 0. \quad (8)$$

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

$$W_{11}H_{12} + W_{12}H_{22} = 0. \quad (9)$$

With Eq. (9) only, we have a trivial solution W_{11}=W_{12}=0. Therefore, an additional constraint should be added to ensure target signal S_1 in output Y_1, *i.e.*,

$$Y_1 = (W_{11}H_{11} + W_{12}H_{21})S_1 = c_1S_1, \quad (10)$$

which leads to

$$W_{11}H_{11} + W_{12}H_{21} = c_1, \quad (11)$$

where c_1 is an arbitrary complex constant. The ABF solution is derived from simultaneous equations Eqs. (9) and (11).

3.2. ABF for a target S_2 and a jammer S_1

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

$$\begin{aligned} W_{21}H_{11} + W_{22}H_{21} &= 0 \quad (12) \\ W_{21}H_{12} + W_{22}H_{22} &= c_2. \quad (13) \end{aligned}$$

3.3. Two sets of ABFs

By combining Eqs. (9), (11), (12), and (13), we can summarize the simultaneous equations for two sets of ABFs as follows,

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}. \quad (14)$$

4. EQUIVALENCE BETWEEN BLIND SOURCE SEPARATION AND ADAPTIVE BEAMFORMERS

As we showed in Eq. (4), the SOS BSS algorithm works to minimize off-diagonal components in

$$E\left[\begin{array}{cc} Y_1Y_1^* & Y_1Y_2^* \\ Y_2Y_1^* & Y_2Y_2^* \end{array}\right], \tag{15}$$

[see Eq. (2)]. Using $\boldsymbol{H}$ and $\boldsymbol{W}$, outputs Y_1 and Y_2 are expressed in each frequency bin as

$$Y_1 = aS_1 + bS_2 \tag{16}$$

$$Y_2 = cS_1 + dS_2, \tag{17}$$

where

$$\left[\begin{array}{cc} a & b \\ c & d \end{array}\right] = \left[\begin{array}{cc} W_{11} & W_{12} \\ W_{21} & W_{22} \end{array}\right]\left[\begin{array}{cc} H_{11} & H_{12} \\ H_{21} & H_{22} \end{array}\right], \tag{18}$$

and they show the paths in Fig. 3.

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

$$\begin{aligned} &E[Y_1Y_2^*] \\ &= ad^*E[S_1S_2^*] + bc^*E[S_2S_1^*] + (ac^*E[S_1^2] + bd^*E[S_2^2]) \\ &= 0. \end{aligned} \tag{19}$$

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 Eq. (19) to zero for all time blocks k. This leads to

$$ac^* = bd^* = 0, \quad abc^*d^* = 0. \tag{20}$$

CASE 1: $a = c_1, c = 0, b = 0, d = c_2$

$$\left[\begin{array}{cc} W_{11} & W_{12} \\ W_{21} & W_{22} \end{array}\right]\left[\begin{array}{cc} H_{11} & H_{12} \\ H_{21} & H_{22} \end{array}\right] = \left[\begin{array}{cc} c_1 & 0 \\ 0 & c_2 \end{array}\right] \tag{21}$$

This equation is identical with the Eq. (14) in ABF.

CASE 2: $a = 0, c = c_1, b = c_2, d = 0$

$$\left[\begin{array}{cc} W_{11} & W_{12} \\ W_{21} & W_{22} \end{array}\right]\left[\begin{array}{cc} H_{11} & H_{12} \\ H_{21} & H_{22} \end{array}\right] = \left[\begin{array}{cc} 0 & c_2 \\ c_1 & 0 \end{array}\right] \tag{22}$$

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

CASE 3: $a = 0, c = c_1, b = 0, d = c_2$

$$\left[\begin{array}{cc} W_{11} & W_{12} \\ W_{21} & W_{22} \end{array}\right]\left[\begin{array}{cc} H_{11} & H_{12} \\ H_{21} & H_{22} \end{array}\right] = \left[\begin{array}{cc} 0 & 0 \\ c_1 & c_2 \end{array}\right] \tag{23}$$

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

CASE 4: $a = c_1, c = 0, b = c_2, d = 0$

$$\left[\begin{array}{cc} W_{11} & W_{12} \\ W_{21} & W_{22} \end{array}\right]\left[\begin{array}{cc} H_{11} & H_{12} \\ H_{21} & H_{22} \end{array}\right] = \left[\begin{array}{cc} c_1 & c_2 \\ 0 & 0 \end{array}\right] \tag{24}$$

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 because we assume that $\boldsymbol{H}(\omega)$ is invertible, and $H_{ji}(\omega) \neq 0$.

The BSS can adapt, even if there is only one active source. In this case, only one set of ABF is achieved [9].

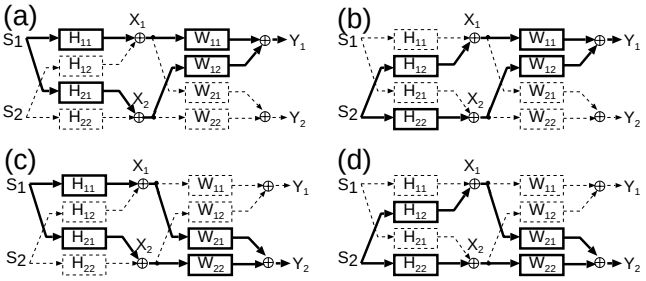


Figure 3: Paths in Equation (18).

5. SIMULATIONS AND DISCUSSIONS

5.1. Limitation of frequency domain BSS

Frequency domain BSS and frequency domain ABF are equivalent [see Eqs. (14) and (21)] in the mean square error sense if the independent assumption ideally holds [see Eq. (19)]. If not, the first and second terms of Eq. (19) behave as a bias noise in getting the correct coefficients a, b, c, d. We have shown in [10], that a long frame size works poorly in frequency domain BSS for speech data of a few seconds. This is because the number of data in each frequency bin becomes few and the assumption of independency between $S_1(\omega, m)$ and $S_2(\omega, m)$ does not hold in each frequency when we use a long frame [11]. Therefore, the upper bound of the performance of BSS is given by that of ABF.

Figure 4 shows the separation performances of BSS and ABF. We performed simulations for two different reverberation time T_R = 0 ms and 300 ms. The room size was 5.73 m × 3.12 m × 2.70 m and the distance between the loudspeakers and microphones was 1.15 m. We used a two-element array with an inter-element spacing of 4 cm. The speech signals arrived from two directions, $-30°$ and $40°$. The length of speech data was about eight seconds. We used the beginning three seconds of the data for learning and the entire eight seconds data were separated. We changed the frame size for DFT, T from 32 to 2048 and investigated the performance for each condition. The sampling rate was 8 kHz, the frame shift was half of frame size T, and the analysis window was a Hamming window. In order to evaluate the performance, we used the signal-to-interference ratio (SIR), defined as the output signal-to-noise ratio (SNR) in dB minus the input SNR in dB. These values were averaged for the whole six combinations with respect to the speakers. As for an ABF, we used the ABF proposed by Frost [12].

In the BSS case, when the frame size was too long, the separation performance got worse. This is because the independency assumption collapses in each frequency when the frame size is long. On the other hand, ABF does not use the assumption of independency of the source signals. In the ABF case, therefore, the separation performance increased as the frame size became longer. Figure 4 confirms that the performance of the BSS is limited by that of the ABF.

5.2. Physical interpretation of BSS

Now, we can understand the behavior of BSS as two sets of ABFs. Figure 5 shows directivity patterns obtained

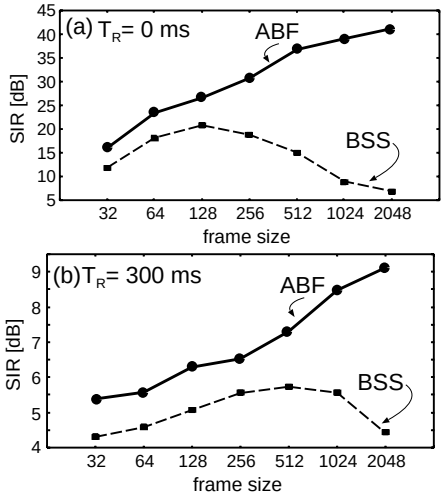


Figure 4: Results of SIR for different frame sizes. The solid lines are for ABF and the broken lines are for BSS. (a) Non-reverberant test, (b) reverberant test (T_R=300 ms).

by BSS and ABF. In Fig. 5, (a) and (b) show directivity patterns by $\boldsymbol{W}$ obtained by BSS, and (c) and (d) show directivity patterns by $\boldsymbol{W}$ obtained by ABF. When $T_R = 0$, a sharp spatial null is obtained by both BSS and ABF [see Figs. 5(a) and (c)]. When T_R = 300 ms, the directivity pattern becomes duller[see Figs. 5(b) and (d)].

BSS removes the sound from jammer direction and reduces reverberation of the jammer signal to some extent [13] in the same way as ABF. This understanding clearly explains the poor performance of the BSS in a real acoustic environment with a 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 [13, 14]. It is well known that an adaptive beamformer outperforms a null beamformer in long reverberation. Our understanding also clearly explains the result.

Note that fundamental differences exist in the adaptation period (*i.e.*, when they should adapt), data length needed to adapt the filters, and necessity of the knowledge of the target signals.

6. CONCLUSION

We gave an interpretation of BSS from physical point of view showing the equivalence between frequency domain Blind Source Separation (BSS) and two sets of frequency domain adaptive beamformers (ABFs). 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, the performance of the BSS is limited by that of the ABF. Moreover, we can understand the behavior of BSS as two sets of ABFs. BSS reduces reverberation of the jammer signal to some extent in the same way as ABF. That is, BSS mainly removes the sound from jammer direction. This understanding clearly explains the poor performance of the BSS in a real acoustic environment with long reverberation.

ACKNOWLEDGEMENTS

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

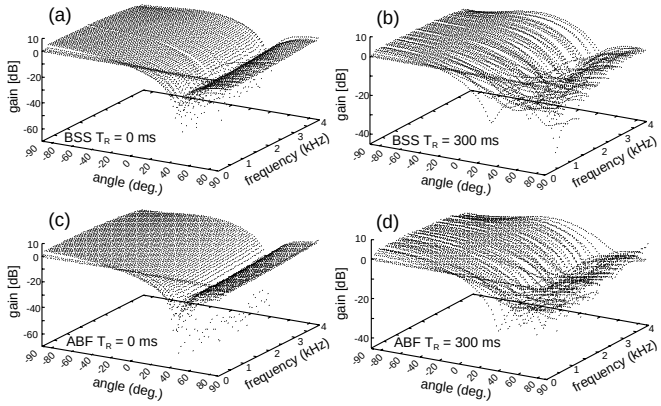


Figure 5: Directivity patterns (a)obtained by BSS (T_R=0 ms), (b) obtained by BSS (T_R=300 ms), (c)obtained by ABF (T_R=0 ms) and (d) obtained by ABF (T_R=300 ms).

REFERENCES

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

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

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

[4] L. Parra and C. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *Proc. NNSP2001*, pp. 273-282, Sept. 2001.

[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, "Multichannel 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, June 2000.

[9] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," *Proc. Eurospeech2001*, pp. 2595-2598, Sept. 2001.

[10] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," *Proc. ICASSP2001*, vol. 5, pp. 2737-2740, May 2001.

[11] S. Araki, S. Makino, R. Mukai, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolved mixture of speech," *Proc. ICA2001*, Dec. 2001 (to appear).

[12] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, No. 8 , Aug. 1972.

[13] 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*, pp. 2599-2602, Sept. 2001.

[14] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," *Proc. ICASSP2001*, pp. 2733-2736, May 2001.