FUNDAMENTAL LIMITATION OF FREQUENCY DOMAIN BLIND SOURCE SEPARATION FOR CONVOLVED MIXTURE OF SPEECH

Shoko Araki † *Shoji Makino* † *Ryo Mukai* † *Tsuyoki Nishikawa* ‡ *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

Despite several recent proposals to achieve Blind Source Separation (BSS) for realistic acoustic signals, the separation performance is still not enough. In particular, when the length of an impulse response is long, the performance is highly limited. In this paper, we consider the reason for the poor performance of BSS in a long reverberation environment. First, we show that it is useless to be constrained by the condition $P \ll T$, where T is the frame size of FFT and P is the length of a room impulse response. We also discuss the limitation of frequency domain BSS, by showing that the frequency domain BSS framework is equivalent to two sets of frequency domain adaptive beamformers.

1. INTRODUCTION

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

To achieve 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.

There have been a lot of proposals to achieve BSS in a realistic room environment, however, the separation performance is still not enough. In this paper, we consider the reason for the poor performance of BSS in a long reverberation environment.

First, we discuss the frame size of FFT used in frequency domain BSS. It is commonly believed that the frame size T must be $P \ll T$ to estimate an unmixing matrix for a P-point room impulse response [5, 6]. We point out that this is not the case for BSS, and show that a smaller frame size is much better, even for long room reverberation.

Next, we discuss the limitation of frequency domain BSS, by showing the equivalence between frequency domain BSS framework and two sets of beamformers. To date, signal separation by using a noise cancellation framework with signal leakage into the noise reference has been discussed in [7, 8]. In these papers it is shown that the least squares criterion is equivalent to the decorrelation criterion of a noise free signal estimate and a signal free noise estimate. 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.

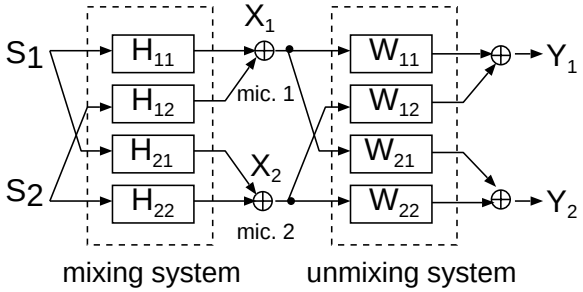


Figure 1: BSS system configuration.

2. FREQUENCY DOMAIN BSS OF CONVOLUTIVE MIXTURES OF SPEECH

The signals recorded by M microphones are given by

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

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 a P-point impulse response from source i to 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 convolutive mixtures is to transform the problem into an instantaneous BSS problem in the frequency domain [3, 4]. Using T-point short time Fourier transformation for (1), we obtain,

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

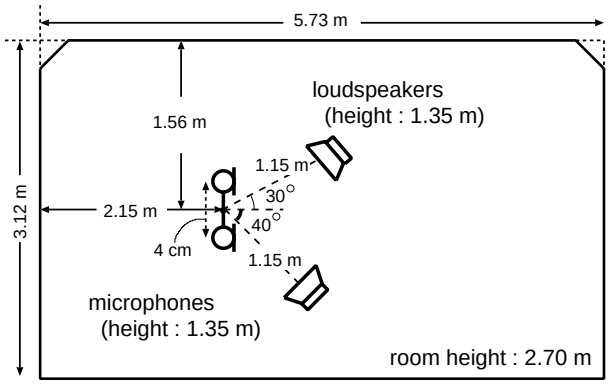


Figure 2: Layout of a room used in experiments.

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 in each frequency independently.

3. EXPERIMENTS

It is commonly believed that the frame size T must be $P \ll T$ to estimate an unmixing matrix for a P-point room impulse response [5, 6]. In this section, we investigate this point for BSS, and show that there is an optimal frame size for BSS [9].

3.1. Conditions for experiments

3.1.1. Learning algorithm

For the calculation of the unmixing matrix $\boldsymbol{W}(\omega)$ in (3), we use an algorithm based on the minimization of the Kullback-Leibler divergence [3, 10]. The optimal $\boldsymbol{W}(\omega)$ is obtained by using the following iterative equation:

$$\begin{aligned}\boldsymbol{W}_{i+1}(\omega) &= \boldsymbol{W}_i(\omega) \\ &+\eta\left[\mathrm{diag}\left(\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right) - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right]\boldsymbol{W}_i(\omega),\end{aligned} \tag{4}$$

where $\boldsymbol{Y}$=$\boldsymbol{Y}(\omega, m)$, $\langle\cdot\rangle$ 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

$$\Phi(\boldsymbol{Y}) = \frac{1}{1+\exp(-\boldsymbol{Y}^{(\mathrm{R})})} + j\,\frac{1}{1+\exp(-\boldsymbol{Y}^{(\mathrm{I})})}, \tag{5}$$

where $\boldsymbol{Y}^{(\mathrm{R})}$ and $\boldsymbol{Y}^{(\mathrm{I})}$ are the real part and the imaginary part of $\boldsymbol{Y}$, respectively.

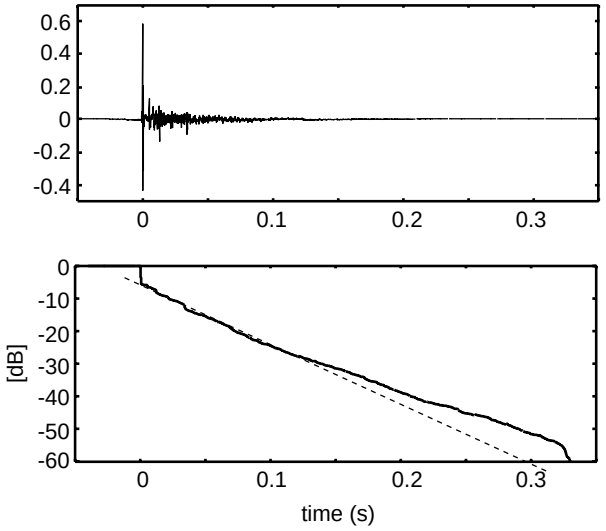


Figure 3: An example of measured impulse response h_{11} used in experiments (T_R = 300 ms).

3.1.2. Conditions for experiments

Separation experiments were conducted using speech data convolved with impulse responses recorded in three environments specified by different reverberation times: T_R = 0 ms, 150 ms (P = 1200), and 300 ms (P = 2400).

The layout of the room we used to measure the impulse responses is shown in Fig. 2. We used a two-element array with inter-element spacing of 4 cm. The speech signals arrived from two directions, −30° and 40°. An example of a measured room impulse response used in our experiments is shown in Fig. 3.

In these experiments, we changed the frame size 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. To solve the permutation problem, we used the blind beamforming algorithm proposed by Kurita et al [10].

3.1.3. Evaluation measure

In order to evaluate the performance for different frame sizes T with different reverberation times T_R, we used the *noise reduction rate* (NRR), defined as the output signal-to-noise ratio (SNR) in dB minus the input SNR in dB.

$$\mathrm{NRR}_i = \mathrm{SNR}_{Oi} - \mathrm{SNR}_{Ii}$$

$$\mathrm{SNR}_{Oi} = 10\log\frac{\sum_\omega |A_{ii}(\omega)S_i(\omega)|^2}{\sum_\omega |A_{ij}(\omega)S_j(\omega)|^2} \tag{6}$$

$$\mathrm{SNR}_{Ii} = 10\log\frac{\sum_\omega |H_{ii}(\omega)S_i(\omega)|^2}{\sum_\omega |H_{ij}(\omega)S_j(\omega)|^2} \tag{7}$$

where $\boldsymbol{A}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)$ and $i \neq j$. These values were averaged for the whole six combinations with respect to the speakers, and NRR_1 and NRR_2 were averaged for the sake of convenience.

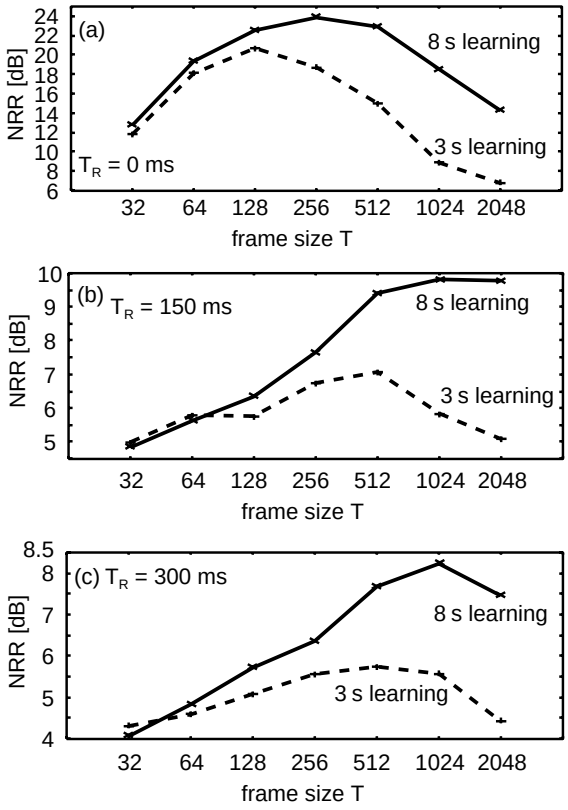


Figure 4: Results of NRR for different frame sizes. The solid lines are for 8 seconds learning, and the dotted lines are for 3 seconds learning. Separation was executed for the 8 seconds long data.

3.2. Experimental results

The experimental results are shown in Fig. 4. The lengths of the mixed speech signals were about eight seconds each. We used the beginning three seconds (dotted lines) or entire eight seconds (solid lines) of the mixed data for learning according to (4), and the entire eight seconds data for separation.

In the case of the three seconds learning, the maximum NRR was obtained when $T = 128$ [Fig. 4(a)] in non-reverberant tests. In reverberant tests, the maximum NRR was obtained using T=512 when T_R=150 ms [Fig. 4(b)], and T_R=300 ms [Fig. 4(c)]. A short frame was found to function far better than a long frame, even for long room reverberation.

For the longer learning data, *i.e.,* the eight seconds data, the results were slightly different. In this case, we obtained a better separation performance than the three seconds learning case. Furthermore, in comparison with the three seconds learning case, the peak performance appeared when we used a longer frame size T. With an overly long frame size, however, the performance become poor even when we used longer learning data.

Even for long room reverberation, the condition $P \ll T$ is useless, and a shorter frame size T is best.

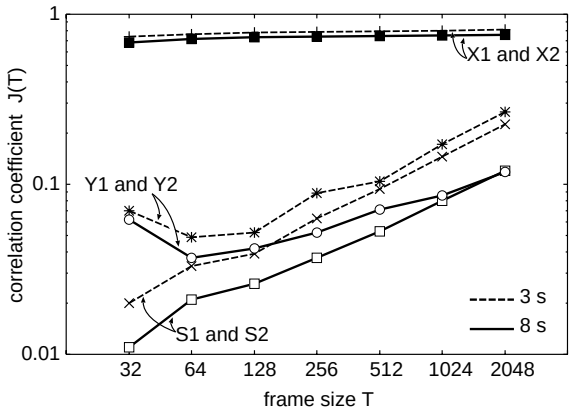


Figure 5: Relationship between T and the correlation coefficient. The solid lines are for data of 8 seconds, and the dotted lines are for data of 3 seconds. T_R=0 ms.

4. DISCUSSION

4.1. Optimum frame size for frequency domain BSS

In the previous section, we showed that a longer frame size T fails. In this section, we discuss the reason why both a short frame and a long frame fail.

In the frequency domain BSS framework, the signal we can use is not $x(n)$ but $\boldsymbol{X}(\omega, m)$. If the frame size T is long, the number of data in each frequency bin becomes few. This causes assumptions to collapse, like the zero mean assumption.

As an example of such a collapse, we observed that the independency of two original signals went down when the frame size became longer. Figure 5 shows the relationship between the frame size T and the correlation coefficient:

$$J(T) = \frac{1}{T}\sum_{\omega}^{T} |r_\omega|, \tag{8}$$

where

$$r_\omega = \frac{\sum_m (U_1(\omega,m) - \overline{U_1(\omega)})(U_2(\omega,m) - \overline{U_2(\omega)})}{\sqrt{\sum_m [U_1(\omega,m) - \overline{U_1(\omega)}]^2}\sqrt{\sum_m [U_2(\omega,m) - \overline{U_2(\omega)}]^2}}. \tag{9}$$

where $\overline{U}$ represents a mean value, and U is original signal S, observed signal, X or separated signal Y. Although a correlation coefficient does not show independency directly, we use this measure as an index of independency. As Fig. 5 shows, the independency decreases when the frame size T becomes longer. In these cases, because the data length gets shorter, the assumption of independency does not hold for the two original sources. This is a reason why the long frame failed.

On the other hand, if we use a short frame, the frame cannot cover the reverberation, therefore, the separation performance is limited.

In frequency domain BSS, the optimum frame size is decided by the trade-off between maintaining the as-

sumption of independency and covering the whole reverberation.

4.2. Length of learning data and separation performance

In 3.2, we obtained a better performance using eight seconds data than using three seconds data. The reason for this result is also explained by Fig. 5. With eight seconds data, independency is better maintained than with three seconds data. Therefore, we can obtain a better performance using the former. Furthermore, the optimum frame size changes when we use learning data of different length, because the optimum frame size is decided by the trade-off we mentioned in 4.1.

5. EQUIVALENCE BETWEEN FREQUENCY DOMAIN BSS AND FREQUENCY DOMAIN ABF

In this section, in order to discuss the fundamental limitation of frequency domain BSS, we show that frequency domain BSS is equivalent to two sets of frequency domain adaptive beamformers (ABF). From the equivalence between BSS and ABF, we can conclude that the performance of BSS is upper bounded by that of ABF [11].

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

In this section, we use a second order statistics (SOS) BSS algorithm for convenience. It is well known that a decorrelation criterion is insufficient to solve problems. In [8], however, it is pointed out that nonstationary signals can provide enough additional information to estimate all W_{ij}. Some authors have utilized SOS for convolutively mixed speech signals [5, 6].

Source signals $S_1(\omega, m)$ and $S_2(\omega, m)$ are assumed to be zero mean and mutually uncorrelated:

$$\begin{aligned} \boldsymbol{R}_S(\omega,k) &= \frac{1}{M}\sum_{m=0}^{M-1}\boldsymbol{S}(\omega,Mk+m)\boldsymbol{S}^*(\omega,Mk+m) \\ &= \boldsymbol{\Lambda}_s(\omega,k), \end{aligned} \tag{10}$$

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

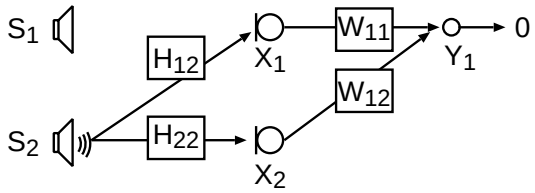
In order to determine $\boldsymbol{W}(\omega)$ so that $Y_1(\omega,m)$ and $Y_2(\omega,m)$ become mutually uncorrelated, we seek a $\boldsymbol{W}(\omega)$ that diagonalizes the covariance matrices $\boldsymbol{R}_Y(\omega,k)$ simultaneously for all k,

$$\begin{aligned} \boldsymbol{R}_Y(\omega,k) &= \boldsymbol{W}(\omega)\boldsymbol{R}_X(\omega,k)\boldsymbol{W}^*(\omega) \\ &= \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)\boldsymbol{\Lambda}_s(\omega,k)\boldsymbol{H}^*(\omega)\boldsymbol{W}^*(\omega) \\ &= \boldsymbol{\Lambda}_c(\omega,k), \end{aligned} \tag{11}$$

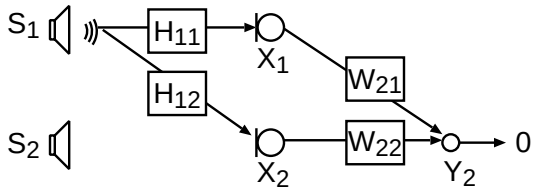
where $\boldsymbol{R}_X$ is the covariance matrix of $\boldsymbol{X}(\omega)$ as follows,

$$\boldsymbol{R}_X(\omega,k) = \frac{1}{M}\sum_{m=0}^{M-1}\boldsymbol{X}(\omega,Mk+m)\boldsymbol{X}^*(\omega,Mk+m), \tag{12}$$

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



(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 6: Two sets of ABF system configurations.

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

$$\begin{aligned} &\arg\min_{W(\omega)}\sum_k ||\text{off-diag}\boldsymbol{W}(\omega)\boldsymbol{R}_X(\omega,k)\boldsymbol{W}^*(\omega)||^2 \\ &s.t.,\ \sum_k \text{diag}||\boldsymbol{W}(\omega)\boldsymbol{R}_X(\omega,k)\boldsymbol{W}^*(\omega)||^2 \neq 0, \end{aligned} \tag{13}$$

where $||x||^2$ is the squared Frobenius norm.

5.2. Frequency domain adaptive beamformer

Here, we consider frequency domain adaptive beamformers (ABF), which can remove jammer signals. Since our aim is to separate two signals S_1 and S_2 with two microphones, two sets of ABF are used (Fig. 6). Note that an ABF can be adapted only when a jammer exists but a target does not exist.

5.2.1. ABF null towards S_2

First, we consider the case of target S_1 and jammer S_2 [Fig. 6(a)]. When target $S_1 = 0$, output $Y_1(\omega,m)$ is expressed as

$$Y_1(\omega,m) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega,m), \tag{14}$$

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

$$\begin{aligned} J(\omega) &= E[Y_1^2(\omega,m)] \\ &= \boldsymbol{W}(\omega)E[\boldsymbol{X}(\omega,m)\boldsymbol{X}^*(\omega,m)]\boldsymbol{W}^*(\omega) \\ &= \boldsymbol{W}(\omega)\mathbf{R}(\omega)\boldsymbol{W}^*(\omega), \end{aligned} \tag{15}$$

where E is the expectation and

$$\mathbf{R}(\omega) = E\left[\begin{array}{cc} X_1(\omega,m)X_1^*(\omega,m) & X_1(\omega,m)X_2^*(\omega,m) \\ X_2(\omega,m)X_1^*(\omega,m) & X_2(\omega,m)X_2^*(\omega,m) \end{array} \right]. \quad (16)$$

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

$$\frac{\partial J(\omega)}{\partial \boldsymbol{W}} = 2\mathbf{R}\boldsymbol{W}^* = 0, \quad (17)$$

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

$$E\left[\begin{array}{cc} X_1X_1^* & X_1X_2^* \\ X_2X_1^* & X_2X_2^* \end{array} \right] \left[\begin{array}{c} W_{11}^* \\ W_{12}^* \end{array} \right] = \left[\begin{array}{c} 0 \\ 0 \end{array} \right]. \quad (18)$$

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

$$W_{11}H_{12} + W_{12}H_{22} = 0. \quad (19)$$

With (19) 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

$$\begin{aligned} Y_1 &= W_{11}X_1 + W_{12}X_2 \\ &= W_{11}H_{11}S_1 + W_{12}H_{21}S_1 = c_1S_1, \end{aligned} \quad (20)$$

which leads to

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

where c_1 is an arbitrary complex constant. Since H_{12} and H_{22} are unknown, the minimization of (15) with adaptive filters W_{11} and W_{12} is used to derive (19) with constraint (21). This means that the ABF solution is derived from simultaneous equations (19) and (21).

5.2.2. ABF null towards S_1

Similarly for target S_2, jammer S_1, and output Y_2 [Fig. 6(b)], we obtain

$$\begin{aligned} W_{21}H_{11} + W_{22}H_{21} &= 0 \quad (22) \\ W_{21}H_{12} + W_{22}H_{22} &= c_2. \quad (23) \end{aligned}$$

5.2.3. Two sets of ABF

By combining (19), (21), (22), and (23), the simultaneous equations for the two sets of ABF are summarized as

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

5.3. Equivalence between BSS and ABF

As we showed in (13), the SOS BSS algorithm works to minimize off-diagonal components in

$$E\left[\begin{array}{cc} Y_1Y_1^* & Y_1Y_2^* \\ Y_2Y_1^* & Y_2Y_2^* \end{array} \right], \quad (25)$$

[see (11)]. Using $\boldsymbol{H}$ and $\boldsymbol{W}$, outputs Y_1 and Y_2 are expressed in each frequency bin as follows,

$$\begin{aligned} Y_1 &= aS_1 + bS_2 \quad (26) \\ Y_2 &= cS_1 + dS_2, \quad (27) \end{aligned}$$

where

$$\left[\begin{array}{cc} a & b \\ c & d \end{array} \right] = \left[\begin{array}{cc} W_{11} & W_{12} \\ W_{21} & W_{22} \end{array} \right] \left[\begin{array}{cc} H_{11} & H_{12} \\ H_{21} & H_{22} \end{array} \right]. \quad (28)$$

5.3.1. 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

$$\begin{aligned} &E[Y_1Y_2^*] \\ &= ad^*E[S_1S_2^*] + bc^*E[S_2S_1^*] + (ac^*E[S_1^2] + bd^*E[S_2^2]) \\ &= 0. \end{aligned} \quad (29)$$

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 (29) to be zero. By squaring the third term and setting it equal to zero

$$\begin{aligned} &(ac^*E[S_1^2] + bd^*E[S_2^2])^2 \\ &= a^2c^{*2}(E[S_1^2])^2 + 2abc^*d^*E[S_1^2]E[S_2^2] + b^2d^{*2}(E[S_2^2])^2 \\ &= 0 \end{aligned} \quad (30)$$

(30) is equivalent to

$$ac^* = bd^* = 0, \quad abc^*d^* = 0. \quad (31)$$

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

$$\left[\begin{array}{cc} W_{11} & W_{12} \\ W_{21} & W_{22} \end{array} \right] \left[\begin{array}{cc} H_{11} & H_{12} \\ H_{21} & H_{22} \end{array} \right] = \left[\begin{array}{cc} c_1 & 0 \\ 0 & c_2 \end{array} \right] \quad (32)$$

This equation is exactly the same as that of the ABF (24).

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

$$\left[\begin{array}{cc} W_{11} & W_{12} \\ W_{21} & W_{22} \end{array} \right] \left[\begin{array}{cc} H_{11} & H_{12} \\ H_{21} & H_{22} \end{array} \right] = \left[\begin{array}{cc} 0 & c_2 \\ c_1 & 0 \end{array} \right] \quad (33)$$

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

Note that the undesirable solutions (*i.e.,* $a = 0, c = c_1, b = 0, d = c_2$ and $a = c_1, c = 0, b = c_2, d = 0$) do not appear, since we assume that $\boldsymbol{H}(\omega)$ is invertible and $H_{ji}(\omega) \neq 0$.

If the uncorrelated assumption between $S_1(\omega)$ and $S_2(\omega)$ collapses, the first and second terms of (29) become the bias noise to get the correct coefficients a, b, c, d.

5.3.2. 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.

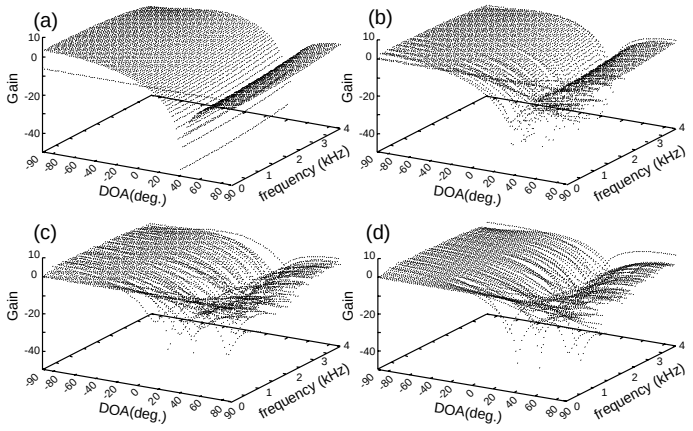


Figure 7: Directivity patterns (a) obtained by an NBF, (b) obtained by BSS (T_R=0 ms), (c) obtained by BSS (T_R=150 ms), and (d) obtained by BSS (T_R=300 ms). DOA means direction of arrival. T=256, three second learning.

5.4. Fundamental limitation of frequency domain BSS

Frequency domain BSS and frequency domain ABF are shown to be equivalent [see (24) and (32)] if the independent assumption ideally holds [see (29)]. We can form only one null towards the jammer in the case of two microphones. Figure 7 shows directivity patterns obtained by a null beamformer (NBF) and BSS; Fig. 7(a) shows a directivity pattern obtained by an NBF that forms a steep null directivity pattern towards a jammer under the assumption of the jammer's direction being known. In Fig. 7, (b),(c), and (d) are drawn by $\boldsymbol{W}$ by BSS: (b) is drawn by $\boldsymbol{W}$ when $T_R = 0$, (c) is drawn by $\boldsymbol{W}$ when $T_R = 150$ ms, and (d) is drawn by $\boldsymbol{W}$ when $T_R = 300$ ms. When $T_R = 0$, a sharp null is obtained like with an NBF. When T_R is long, the directivity pattern is comparatively duller; however, we can draw a directivity pattern. Although BSS and ABF can reduce reverberant sounds to some extent [12], they mainly remove the sounds from the jammer direction. This understanding clearly explains the poor performance of BSS in a real room with long reverberation.

Moreover, as we have shown in section 3, a long frame size works poorly in frequency domain BSS for speech data of a few seconds. This is because when we use a long frame, the assumption of independency between $S_1(\omega)$ and $S_2(\omega)$ does not hold in each frequency; this is caused by the lack of the number of data in each frequency bin. Therefore, the performance of BSS is upper bounded by that of ABF.

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

6. CONCLUSIONS

In this paper, we discuss why the separation performance is poor when there is long reverberation.

First, we show that it is useless to be constrained by the condition $P \ll T$, where T is the frame size of FFT and P is the length of a room impulse response. This is because the lack of data causes the collapse of the assumption of independency between the two original signals in each frequency bin when the data length is short, or when a longer frame size T is used.

Next, we show that frequency domain BSS is equivalent to two sets of frequency domain ABF. Because ABF (and BSS) mainly considers the direct sound by making a null towards jammer direction, the separation performance is fundamentally limited. This understanding clearly explains the poor performance of BSS in a real room with long reverberation.

ACKNOWLEDGEMENTS

We would like to thank Dr. Shigeru Katagiri for his continuous encouragement.

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] 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.

[6] 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.

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

[8] 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.

[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] 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.

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

[12] 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.