The Fundamental Limitation of Frequency Domain Blind Source Separation for Convolutive Mixtures of Speech

Shoko Araki, *Member, IEEE*, Ryo Mukai, *Member, IEEE*, Shoji Makino, *Senior Member, IEEE*, Tsuyoki Nishikawa, and Hiroshi Saruwatari, *Member, IEEE*

***Abstract*—Despite several recent proposals to achieve blind source separation (BSS) for realistic acoustic signals, the separation performance is still not good enough. In particular, when the impulse responses are long, performance is highly limited. In this paper, we consider a two-input, two-output convolutive BSS problem. First, we show that it is not good to be constrained by the condition $T > P$, where T is the frame length of the DFT and P is the length of the room impulse responses. We show that there is an optimum frame size that is determined by the trade-off between maintaining the number of samples in each frequency bin to estimate statistics and covering the whole reverberation. We also clarify the reason for the poor performance of BSS in long reverberant environments, highlighting that the framework of BSS works as two sets of frequency-domain adaptive beamformers. Although BSS can reduce reverberant sounds to some extent like adaptive beamformers, they mainly remove the sounds from the jammer direction. This is the reason for the difficulty of BSS in reverberant environments.**

***Index Terms*—Blind source separation, convolutive mixture, frame size, frequency domain, independent component analysis, reverberant speech.**

I. Introduction

BLIND source separation (BSS) is an approach to estimate original 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 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]. Some approaches consider unmixing systems w_{ij} as FIR filters, and estimate those filters [3], [4]; other approaches transform the problem into the frequency domain to solve an instantaneous BSS problem for every frequency simultaneously [5], [6]. There are a few applications of BSS to mixed speech signals in realistic acoustical environments [7], but the separation performance is still not good enough [8].

Manuscript received September 5, 2001; revised November 11, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Walter Kellermann.

S. Araki, R. Mukai and S. Makino are with the NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan (e-mail: shoko@cslab.kecl.ntt.co.jp; ryo@cslab.kecl.ntt.co.jp; maki@cslab.kecl.ntt.co.jp).

T. Nishikawa and H. Saruwatari are with the Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0192, Japan (e-mail: tsuyo-ni@is.aist-nara.ac.jp; sawatari@is.aist-nara.ac.jp).

Digital Object Identifier 10.1109/TSA.2003.809193

In this paper, we consider a two-input, two-output BSS problem of convolutive mixtures of speech in the frequency domain. The intention of this paper is to clarify the reason why the performance of frequency-domain BSS in long reverberant environments is poor. First, we show that it is not good to be constrained by the condition $T > P$, where T is the frame length of the DFT and P is the length of the room impulse responses. We also clarify the reason for the poor performance of BSS in long reverberant environments, showing that the framework of BSS works as two sets of frequency-domain adaptive beamformers.

In Section II, we summarize the framework of frequency-domain BSS for convolutive mixtures. In Section III, we explain the relationship between the frame size T and the length of the room impulse responses P. In Section IV, we investigate how to choose the frame size for BSS of convolutive mixtures, and show that a longer frame size is not suitable even for long reverberation [9]. In Section V, we discuss the existence of an optimum frame size for frequency-domain BSS. Moreover, we also clarify the reason for the poor performance of BSS in highly reverberant environments. We point out that frequency-domain BSS works as two sets of frequency-domain adaptive beamformers [10]. Cardoso and Souloumiac [11] indicated the connection between blind identification and beamforming in a narrowband context. We discuss this relationship more closely, and provide a physical understanding of frequency-domain BSS. Although BSS can reduce reverberant sounds to some extent [12] like adaptive beamformers, it mainly removes sound from the jammer direction. This understanding explains the reason for the difficulty of BSS in reverberant environments. Section VI concludes this paper.

II. Frequency Domain BSS of Convolutive Mixtures of Speech

The N signals recorded by M microphones are given by

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

where s_i is the source signal from a source i, x_j is the received signal by 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).

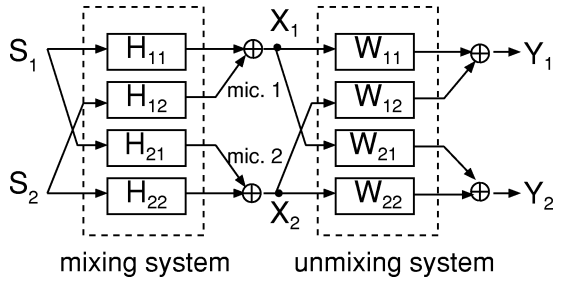

Fig. 1. BSS system configuration.

The frequency-domain approach to convolutive mixtures is to transform the problem into an instantaneous BSS problem in the frequency domain [5], [6]. Using T-point short-time Fourier transformation for (1), we obtain

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

where ω denotes the frequency, m represents the time-dependence of the short-time Fourier transformation, $\boldsymbol{S}(\omega, m) = [S_1(\omega, m), S_2(\omega, m)]^T$ is the source signal vector, and $\boldsymbol{X}(\omega, m) = [X_1(\omega, m), X_2(\omega, m)]^T$ is the observed signal vector. The mixing matrix $\boldsymbol{H}(\omega)$ does not depend on time m. We assume that the (2×2) mixing matrix $\boldsymbol{H}(\omega)$ is invertible, and that $H_{ji}(\omega) \neq 0$.

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 $\boldsymbol{Y}(\omega, m) = [Y_1(\omega, m), Y_2(\omega, m)]^T$ is the estimated source signal vector, and $\boldsymbol{W}(\omega)$ represents a (2×2) unmixing matrix at frequency bin ω. The unmixing matrix $\boldsymbol{W}(\omega)$ is determined so that $Y_1(\omega, m)$ and $Y_2(\omega, m)$ become mutually independent. The above calculation is carried out at each frequency independently.

III. Frame Size for BSS of Convolutive Mixtures

It is commonly believed that the frame size T must be longer than P to estimate the unmixing matrix for a P-point room impulse response. In this paper, we consider that the DFT frame size T is equal to the length of unmixing filter Q.

The reason for the belief that the frame size T must be longer than P is because: i) A linear convolution can be approximated by a circular convolution if $T > 2P$. Therefore, in order to transform (1) into (2) accurately, $T > 2P$ must be held. ii) Signal separation by using a noise cancellation framework with signal leakage into the noise reference was discussed in [13], [14]. In the case of a noise canceller, we should use an FIR filter of length $Q \geq P$; therefore, $T \geq P$. iii) If we want to estimate the inverse system of a system with impulse response P-taps long, we need an inverse system that is Q-taps long, where $Q > P$; therefore, $T > P$.

In [8], [15], however, the authors used $T > Q$ in order to reduce permutation inconsistency and $T \leq P$ to assure that frame size T of the DFT is computed over a stationary window of data. It is important to clarify what filter length is suitable and also why that filter size is suitable for BSS of convolutive mixtures. We investigate this in Section IV and discuss our findings in Section V.

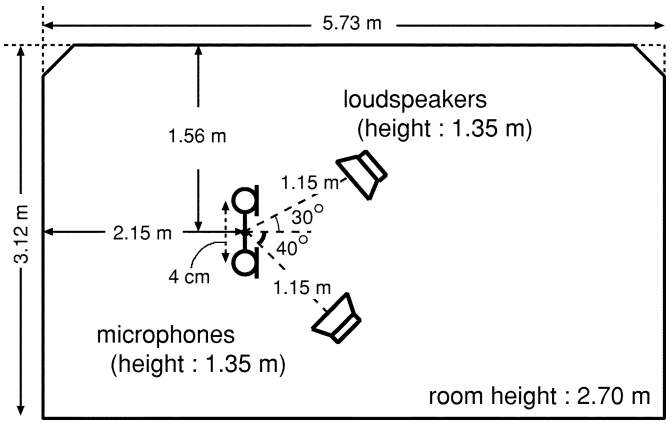

Fig. 2. Layout of a room used in experiments.

IV. Experiments

A. Conditions for Experiments

1) Learning Algorithm: For the calculation of the unmixing matrix $\boldsymbol{W}(\omega)$ in (3), an algorithm based on the minimization of the Kullback-Leibler divergence [6], [16] has been proposed. The optimal $\boldsymbol{W}(\omega)$ is obtained by using the following iterative equation:

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

where $\boldsymbol{Y} = \boldsymbol{Y}(\omega, m)$, $\langle\cdot\rangle$ denotes the averaging operator, i is used to express the value of the ith step in the iterations, and η is the step size parameter. In addition, we define the nonlinear function $\Phi(\cdot)$ as

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

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

For more stable and faster convergence, Amari [17] proposed an algorithm based on the natural gradient. Using the natural gradient, we get the following iterative equation:

$$\boldsymbol{W}_{i+1}(\omega) = \boldsymbol{W}_i(\omega) + \eta\left[\mathrm{diag}\left(\left\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\right\rangle\right) - \left\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\right\rangle\right]\boldsymbol{W}_i(\omega). \tag{6}$$

For the calculation of unmixing matrix $\boldsymbol{W}(\omega)$ in (3), we used this iterative equation (6).

2) Conditions for Experiments: Separation experiments were conducted using speech data convolved with impulse responses recorded in three environments specified by different reverberation times: $T_{\mathrm{R}} = 0$ ms, 150 ms, and 300 ms. Since the sampling rate was 8 kHz, 150 ms, and 300 ms corresponds to $P = 1200$ taps and 2400 taps, respectively. As the original speech, we used two sentences spoken by two male and two female speakers. The investigations were carried out for six combinations of speakers.

The layout of the room we used to measure the impulse responses is shown in Fig. 2. We used a two-element array with an 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.

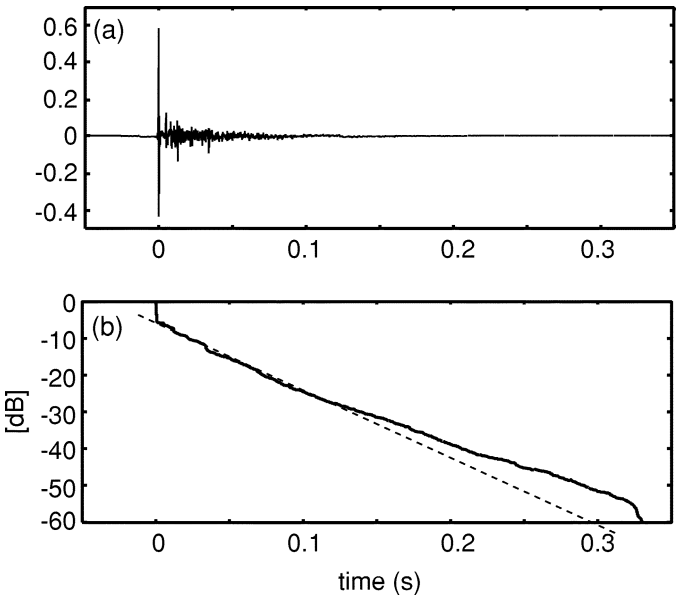

Fig. 3. Example of (a) measured impulse response h_{11} used in experiments and (b) its energy decay curve. $T_{\mathrm{R}} = 300$ ms.

Fig. 3(b) shows the energy decay curve $r(t)$ of an impulse response $h(t)$, which can be obtained by integrating the energy of impulse response as follows:

$$r^2(t) = \int_t^{\infty} h^2(t)\, dt.$$

The reverberation time T_{R} is defined as the time for an energy attenuation of 60 dB.

In these experiments, we varied the frame size T from 32 to 2048 and investigated the performance for each condition. The frame shift was half of the frame size T, the analysis window was a Hamming window, and the step size for adaptation was $\eta = 1.0 \times 10^{-5}$. The learning of $\boldsymbol{W}(\omega)$ using (6) was iterated until the adaptation converged.

In frequency-domain BSS, a scaling and permutation problem occurs, i.e., the estimated source signal components are recovered with a different order and gain in the different frequency bins. To solve this problem, we used the blind beamforming algorithm proposed by Kurita *et al.* [16]: first, from the directivity pattern obtained by $W(\omega)$ we estimate the source directions and reorder the row of $W(\omega)$ so that the directivity pattern forms a null toward the same direction in all frequency bins, then we normalize the row of $W(\omega)$ so that the gains of the target directions become 0 dB.

3) Evaluation Measure: In order to evaluate the performance for different frame sizes T with different reverberation times T_{R}, we used the *signal-to-interference ratio* (SIR), defined as follows:

$$\mathrm{SIR}_i = \mathrm{SIR}_{\mathrm{O}i} - \mathrm{SIR}_{\mathrm{I}i}$$

$$\mathrm{SIR}_{\mathrm{O}i} = 10 \log \frac{\sum_{\omega} |A_{ii}(\omega) S_i(\omega)|^2}{\sum_{\omega} |A_{ij}(\omega) S_j(\omega)|^2} \tag{7}$$

$$\mathrm{SIR}_{\mathrm{I}i} = 10 \log \frac{\sum_{\omega} |H_{ii}(\omega) S_i(\omega)|^2}{\sum_{\omega} |H_{ij}(\omega) S_j(\omega)|^2} \tag{8}$$

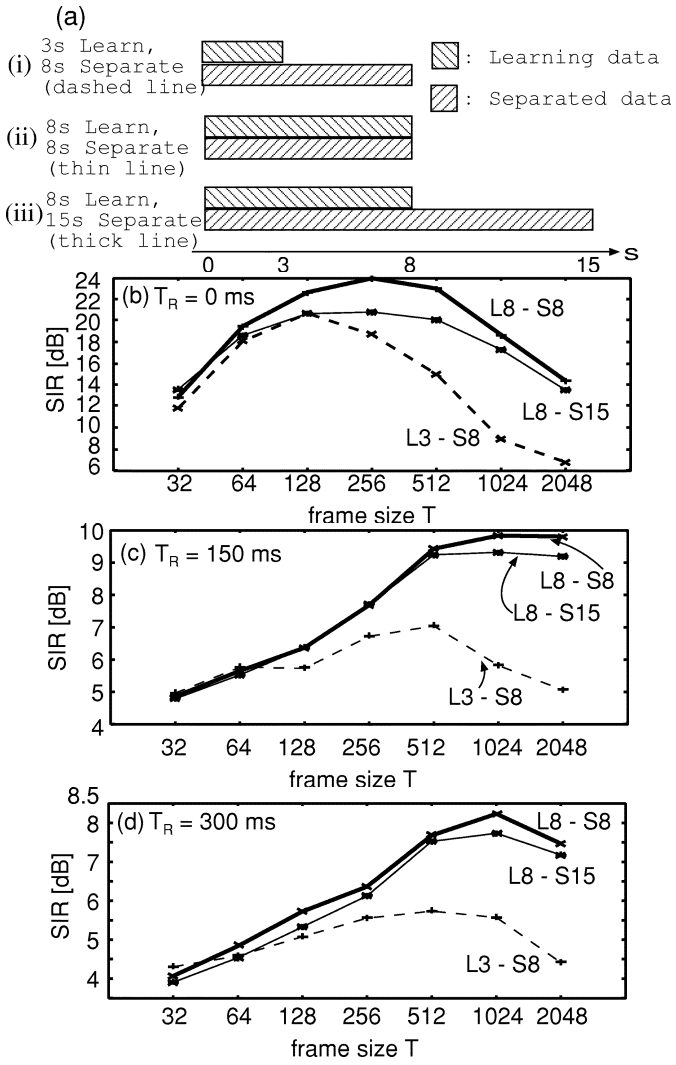

Fig. 4. Results of SIR for different frame sizes: (a) explanation of the test data conditions, (b) nonreverberant test, (c) reverberant test ($T_{\mathrm{R}} = 150$ ms), and (d) reverberant test ($T_{\mathrm{R}} = 300$ ms).

where $\boldsymbol{A}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)$ and $i \neq j$. SIR means the ratio of a target-originated signal to a jammer-originated signal. These values were averaged over all six combinations of the speakers, and SIR_1 and SIR_2 were averaged.

B. Experimental Results

The experimental results are shown in Fig. 4. First we explain the test data condition in Fig. 4(a) as follows; i) The lengths of the mixed speech signals were about 8 s each, the beginning 3 s of the mixed data were used for the learning according to (6), and the entire 8 s data was separated (dashed lines); ii) The entire 8 s of the mixed data for the learning, and the entire 8 s data for the separation (thick lines); iii) The lengths of the mixed data were 15 s, the lengths of the learning data were the beginning 8 s, and the entire 15 s data were separated (thin lines). Note that in cases i) and iii), the separation evaluation was executed for "open" data, i.e., the separated data was not the same as the learning data, and in case ii), the evaluation was done for "closed" data, i.e., the separated data was exactly the same as the learning data.

In the case of 3-s learning (dashed lines), we obtained the best performance with a short frame size both in nonreverberant tests [Fig. 4(b)] and in reverberant tests [Fig. 4(c) and (d)]. A short frame was found to function far better than a long frame size, even for long room reverberation.

For the longer learning data, i.e., the 8-s data, the results were slightly different from the 3-s learning case. In this case, we obtained a better separation performance than for the 3-s learning case. Furthermore, in comparison with the 3-s learning case, the best performance was realized when we used a longer frame size T. With an overly long frame size, however, the performance was poor even when we used longer learning data.

Even for long room reverberation, the condition $T > P$ is not suitable, and a shorter frame size T is best. We will discuss this in the next section.

Moreover, we achieved better performance in "closed" tests than in "open" tests. Since the room impulse response was not changed during the entire experiment, there is little difference between the closed and open tests. If the room impulse response were to slightly change during a real recording, however, the difference there would be larger between the closed and open tests.

V. Analysis and Discussion

A. Optimum Frame Size for Frequency-Domain BSS

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

1) The Case of a Long Frame: In the frequency-domain BSS framework, the signal we can use is not $\boldsymbol{x}(n)$ but $\boldsymbol{X}(\omega, m)$. If the frame size T is long, the number of samples in each frequency bin is small. This causes assumptions to collapse, like the zero-mean and independence assumptions, because this makes the correct estimation of statistics difficult. In this paper, when the number of samples is too small to estimate statistics correctly, we say "the independence assumption is not held" or "the independence decreases/collapses." As an example of such a collapse, we observe that the independence of two source signals goes down when the frame size becomes longer. In order to evaluate the independence of two signals from different points of view, we used two measures. One was the Frobenius norm of the adaptation term in (6), and the other was a correlation coefficient.

Measure 1: Off-Diagonal of $\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle$: To investigate the independence between the two signals, we used the bracketed term of the adaptation equation (6) to define a new measure. Here, we set this term as

$$\boldsymbol{V}(\omega) = \mathrm{diag}\left(\left\langle\Phi\left((\boldsymbol{U})\boldsymbol{U}^{\mathrm{H}}\right)\right\rangle\right) - \left\langle\Phi(\boldsymbol{U})\boldsymbol{U}^{\mathrm{H}}\right\rangle \tag{9}$$

$$= \begin{bmatrix} 0 & v_{12}(\omega) \\ v_{21}(\omega) & 0 \end{bmatrix} \tag{10}$$

where $\boldsymbol{U} = \boldsymbol{U}(\omega, m), v_{ij}(\omega) = -\langle\Phi(U_i)U_j^*\rangle$ and U represents the source signal S, observed signal X or separated signal Y.

When the algorithm converges, $\boldsymbol{V}(\omega)$ becomes the zero matrix, and the two signals become mutually independent. Therefore, we can consider this term as a measure of the independence, i.e., if this term is small, two signals should be highly independent.

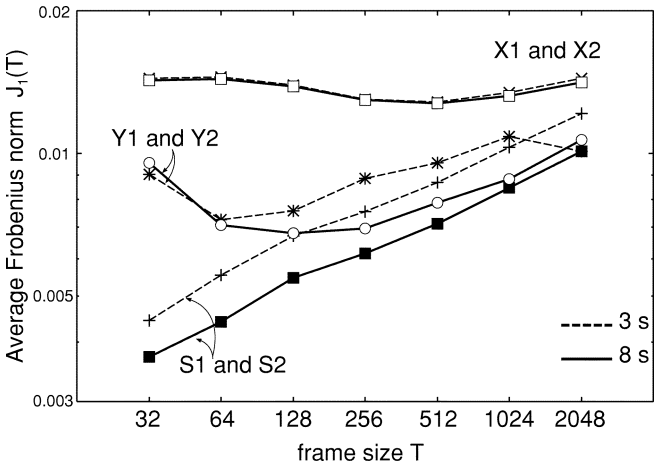

Fig. 5. Average Frobenius norm of off-diagonal terms of $\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle$. The solid lines refer to data of 8 s, and the dashed lines refer to data of 3 s. No reverberation ($T_{\mathrm{R}} = 0$ ms).

The independence measure is the Frobenius norm of $\boldsymbol{V}(\omega)$ averaged over frequency, defined as follows:

$$J_1(T) = \frac{1}{T}\sum_{\omega=1}^{T}\|\boldsymbol{V}(\omega)\|_{\mathrm{F}} = \frac{1}{T}\sum_{\omega=1}^{T}\sqrt{\sum_{i\neq j}|v_{ij}(\omega)|^2}. \tag{11}$$

Fig. 5 shows this measure for each frame size T. As the measure for source signals S_1 and S_2 shows, the independence decreases when the frame size T becomes longer. In these cases, because the number of data samples is smaller, the assumption of independence does not hold for the two source signals. This is a reason why the long frame fails. Besides, the independence is higher when we use 8-s long learning data than when we use 3-s data.

Measure 2: Correlation Coefficient: For a second measure of the independence, we used the correlation coefficient. Although a correlation coefficient does not show independence directly, we use this measure as an index of independence. The independence measure here is the average of the correlation coefficients over all frequency bins

$$J_2(T) = \frac{1}{T}\sum_{\omega}^{T}|r_\omega| \tag{12}$$

where

$$r_\omega = \frac{\sum_m\Big(U_1(\omega, m) - \overline{U_1(\omega)}\Big)\Big(U_2(\omega, m) - \overline{U_2(\omega)}\Big)}{\sqrt{\sum_m\Big[U_1(\omega, m) - \overline{U_1(\omega)}\Big]^2}\sqrt{\sum_m\Big[U_2(\omega, m) - \overline{U_2(\omega)}\Big]^2}}. \tag{13}$$

$\overline{U}$ represents a mean value, and U is the source signal S, observed signal, X or separated signal Y.

Fig. 6 shows the relationship between the frame size T and the average correlation coefficient. The decrease of independence is observed again for the source signals when the frame size T becomes longer. Moreover, the independence is also higher with 8-s learning data, than with 3-s data.

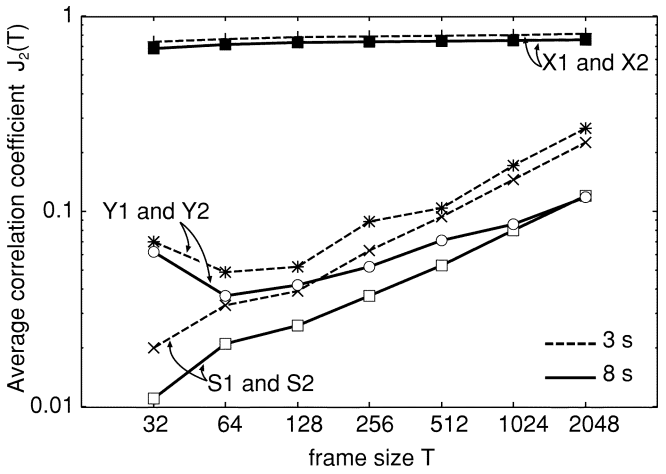

Fig. 6. Relationship between T and the average correlation coefficient. The solid lines refer to data of 8 s, and the dashed lines refer to data of 3 s. $T_{\mathrm{R}} = 0$ ms.

Because the number of samples in each frequency bin becomes smaller when we use a longer frame, the assumption of independence does not hold for the two source signals. This is a reason why the long frame fails.

Moreover, when the filter length becomes longer, the number of coefficients to be estimated increases while the number of samples for learning in each frequency bin decreases. As a result, the estimation error is integrated and the performance worsens. This is another reason for poor performance when we use a longer frame.

2) The Case of a Short Frame: In our experiments, a shorter frame also failed. When we use a short frame, the frame could not cover the reverberation; therefore, the separation performance was limited.

In order to show the reverberation coverage of a frame, we investigated the impulse response of the mixing and unmixing systems in the time domain [12]. We considered a separated signal y_1, target signal s_1, and jammer signal s_2. When the target s_1 is an impulse $\delta(n)$ and jammer signal $s_2 = 0$, we can measure the impulse response h_{T} of the system for the target signal [Fig. 7(a)]. Similarly, when $s_1 = 0$ and $s_2 = \delta(n)$, we can measure the impulse response h_{J} of the system for the jammer [Fig. 7(b)]

$$h_{\mathrm{T}}(n) = w_{11}(n) * h_{11}(n) + w_{12}(n) * h_{21}(n) \qquad (14)$$

$$h_{\mathrm{J}}(n) = w_{11}(n) * h_{12}(n) + w_{12}(n) * h_{22}(n). \qquad (15)$$

The impulse response h_{J} determines the remaining reverberation sound of the jammer that cannot be reduced using BSS, and that worsens the separation performance. Therefore we pay attention to h_{J}.

Fig. 8 shows the jammer signal's impulse response h_{J} for $T = 32$ and 512 and for $T_{\mathrm{R}} = 300$ ms ($P = 2400$). Fig. 8(a) is the observed signal at microphone 1 when the target signal $s_1 = 0$, i.e., the impulse response h_{12}. In the case of $T = 32$ [Fig. 8(b)], the length of the unmixing system is much shorter than the length of the reverberation; accordingly, reverberation longer than the frame cannot be reduced. On the other hand, when $T = 512$ [Fig. 8(c)], the reverberation covered by the frame size is reduced.

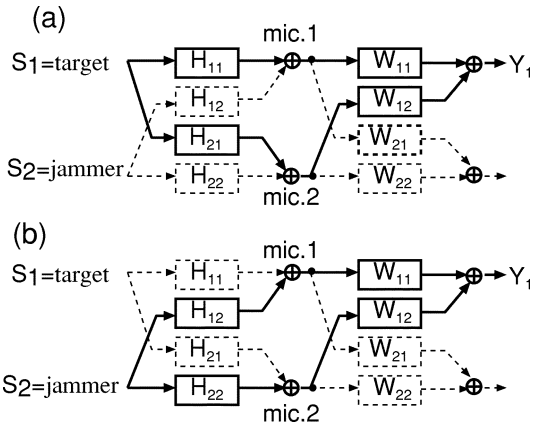

Fig. 7. Definition of system for target and jammer. Signal s_1 is a target and signal s_2 is a jammer. (a) System for the target. (b) System for the jammer.

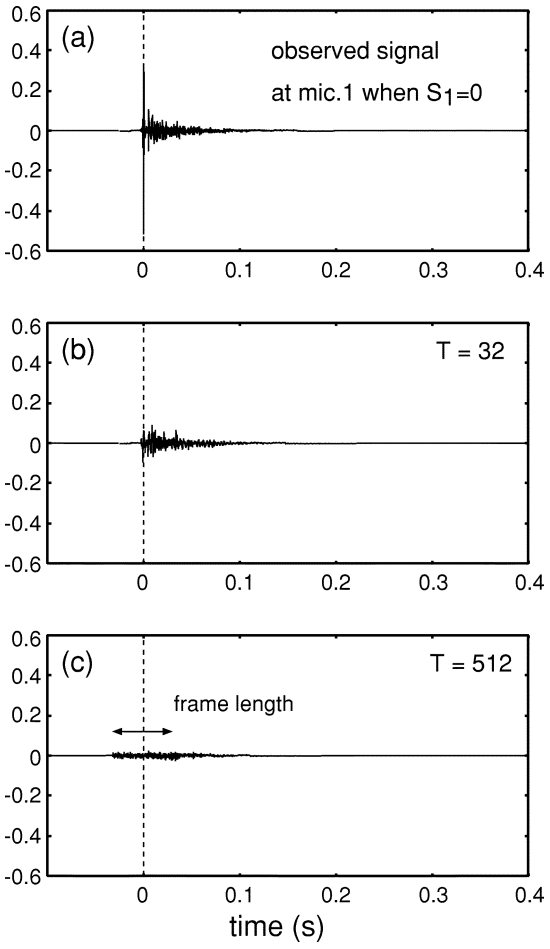

Fig. 8. Impulse response of system for jammer h_{J}.

From Section V-A-1 and Section V-A-2, we conclude that in frequency-domain BSS, there is an optimum frame size determined by a tradeoff between maintaining the assumption of independence and covering the whole reverberation interval.

B. Length of Learning Data and Separation Performance

In Section IV-B, we obtained better performance using 8-s data than using 3-s data. The reason for this result is also explained by Figs. 5 and 6. With 8-s data, the independence was

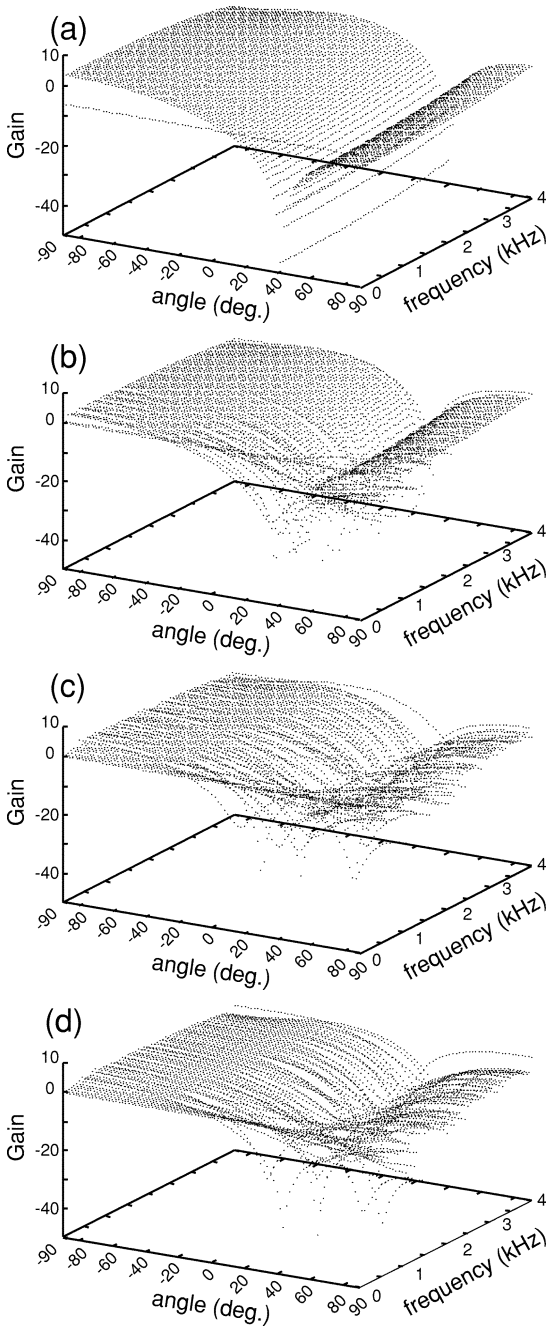

Fig. 9. Directivity patterns obtained by (a) NBF, (b) BSS ($T_{\rm R} = 0$ ms), (c) BSS ($T_{\rm R} = 150$ ms), and (d) BSS ($T_{\rm R} = 300$ ms). Frame size $T = 256$, 3-s learning.

better maintained than with 3-s data. Therefore, we can obtain better performance using the longer data. Furthermore, the optimum frame size changes when we use learning data of different length, because the optimum frame size is determined by the trade-off we mentioned in Section V-A.

C. Physical Understanding of Frequency-Domain BSS

It is well known that an unmixing matrix $\boldsymbol{W}(\omega)$ can at best be obtained up to a scaling and a permutation. Before the permutation and scaling problem, however, we must note that the BSS algorithm cannot solve the dereverberation/deconvolution problem in itself [14].

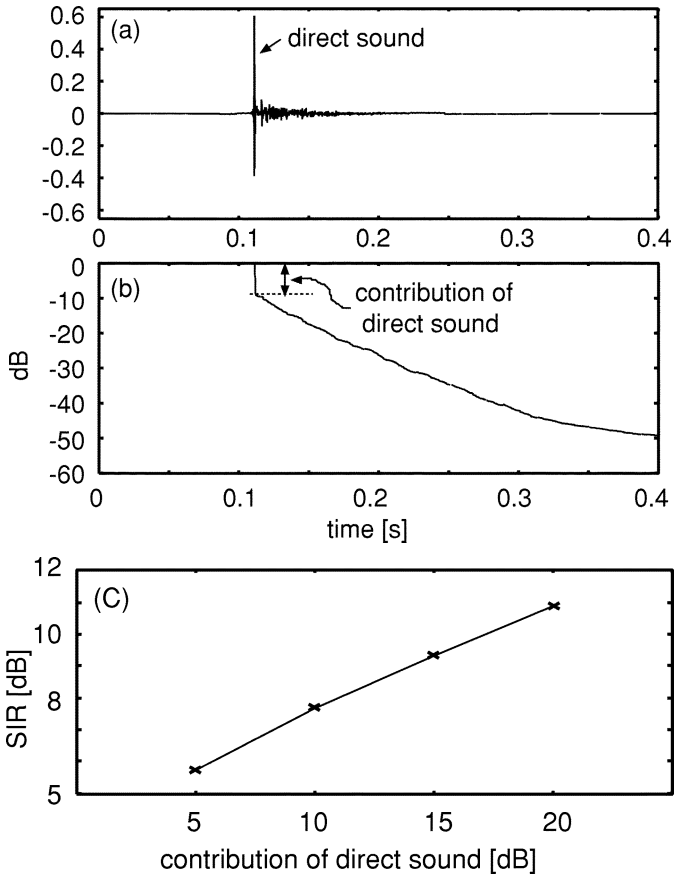

Fig. 10. Relationship between the contribution of a direct sound and the separation performance. $T_{\rm R} = 300$ ms, $T = 512$. (a) Example of an impulse response. (b) Energy decay curve. (c) Separation performance.

In the BSS framework, what an unmixing matrix $\boldsymbol{W}(\omega)$ can do is to minimize the second term of (6), and $\boldsymbol{W}(\omega)$ becomes a solution of

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix} \tag{16}$$

where c_1 and c_2 are arbitrary complex constants. This means that $\boldsymbol{W}(\omega)$ is in general not the inverse matrix of the mixing system $\boldsymbol{H}$. We can understand this unmixing system $\boldsymbol{W}(\omega)$ as two sets of microphone array systems, i.e., two sets of adaptive beamformers (ABFs) [10].

We can form only one null toward the jammer in the case of two microphones. Fig. 9 shows directivity patterns obtained by a null beamformer (NBF) and BSS; Fig. 9(a) is obtained by an NBF that forms a steep null directivity pattern toward a jammer under the assumption of the jammer's direction being known. Fig. 9(b)–(d) are obtained by BSS for (b) $T_{\rm R} = 0$, (c) $T_{\rm R} = 150$ ms, and (d) $T_{\rm R} = 300$ ms. When $T_{\rm R} = 0$, a sharp null is obtained like with an NBF. When $T_{\rm R}$ is long, the directivity pattern is comparatively duller; however, we can still draw a directivity pattern. Although BSS and ABF can reduce reverberant sounds to some extent (see Fig. 8) [12], they mainly remove sound from the jammer direction. This understanding clearly explains the poor performance of BSS in a real room with long reverberation.

Fig. 10 shows the performance when the contribution of the direct sound is changed artificially. The performance increases with the increase of the contribution of the direct sound. This is

the same characteristic as that of an ABF. Because an unmixing system $\boldsymbol{W}(\omega)$ mainly removes sound from the jammer direction, chiefly the direct (largest) sound of the jammer can be separated, and the other reverberant components which arrive from different directions cannot be separated. As a result, separation performance is fundamentally limited.

Moreover, as we have shown in Section IV, a long frame size works poorly in the frequency-domain BSS for speech data of a few seconds. This is because when we use a long frame, the assumption of independence between $S_1(\omega, m)$ and $S_2(\omega, m)$ does not hold at each frequency; this is caused by a small number of samples in each frequency bin. ABF, however, does not use the assumption of independence, so it can achieve good performance using a long frame size, i.e., high frequency resolution. Therefore, the performance of BSS is upper bounded by that of ABF. Note that ABF needs to know the array manifold and the target direction. Note also that ABF can be adapted only when a jammer exists but a target does not exist, whereas BSS can adapt in the presence of target and jammer, and also in the presence of only one active source.

The BSS was shown to outperform an NBF [18], [12]. It is well known that an ABF outperforms an NBF in long reverberation. Our understanding also clearly explains this.

VI. Conclusions

In this paper, we discussed why the separation performance of frequency-domain BSS is poor when there is long reverberation. First, we showed that it is not good to be constrained by the condition $T > P$, where T is the frame size of the 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 independence between the two source signals in each frequency bin when the data length is short, or when a longer frame size T is used. On the other hand, when we use a short frame, we cannot get good performance, because long reverberation cannot be covered by a short frame. Therefore, there is an optimum frame size determined by a trade-off between maintaining the assumption of independence and covering the whole reverberation in frequency-domain BSS.

Next, we showed a physical understanding of frequency-domain BSS. We can understand the frequency-domain BSS system as two sets of microphone array systems i.e., two sets of adaptive beamformers (ABFs). Because ABF and BSS mainly consider sound from the jammer direction by making a null toward jammer, the separation performance is fundamentally limited. Furthermore, we can conclude that the performance of the BSS is upper bounded by that of ABF.

This understanding clearly explains the poor performance of BSS in a real room with long reverberation.

Acknowledgment

The authors thank Dr. S. Katagiri and Dr. K. Shikano for their continuous encouragement. They also thank Dr. F. Asano for detailed discussions. The authors are grateful to the associate editor and the anonymous reviewers, whose constructive and helpful comments led to significant improvements in the manuscript.

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*. New York: Wiley, 2000.
[3] T. W. Lee, *Independent Component Analysis—Theory and Applications*. Norwell, MA: Kluwer, 1998.
[4] M. Kawamoto, A. K. Barros, A. Mansour, K. Matsuoka, and N. Ohnishi, "Real world blind separation of convolved nonstationary signals," in *Proc. Workshop Indep. Compon. Anal. Signal. Sep.*, 1999, pp. 347–352.
[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.
[6] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proc. Workshop Indep. Compon. Anal. Signal. Sep.*, 1999, pp. 365–370.
[7] T. W. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real world signals," *Neural Networks*, vol. 4, pp. 2129–2134, 1997.
[8] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proc. ICASSP2000*, 2000, pp. 1041–1044.
[9] 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, pp. 2737–2740.
[10] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," in *Proc. Eurospeech 2001*, 2001, pp. 2595–2598.
[11] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *Proc. Inst. Elect. Eng.*, vol. 140, no. 6, pp. 362–370, Dec. 1993.
[12] R. Mukai, S. Araki, and S. Makino, "Separation and dereverberation performance of frequency domain blind source separation for speech in a reverberant environment," in *Proc. Eurospeech 2001*, 2001, pp. 2599–2602.
[13] S. Gerven and D. Compernolle, "Signal separation by symmetric adaptive decorrelation: Stability, convergence, and uniqueness," *IEEE Trans. Signal Processing*, vol. 43, pp. 1602–1612, July 1995.
[14] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 405–413, Oct. 1993.
[15] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 320–327, May 2000.
[16] 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*, 2000, pp. 3140–3143.
[17] S. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, pp. 251–276, 1998.
[18] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," in *Proc. ICASSP2001*, 2001, pp. 2733–2736.

Shoko Araki (M'01) received the B.E. and M.E. degrees in mathematical engineering and information physics from the University of Tokyo, Tokyo, Japan, in 1998 and 2000, respectively.

In 2000, she joined NTT Communication Science Laboratories. Her research interests include digital signal processing, array signal processing and blind source separation applied to speech signals.

Ms. Araki is a member of the Acoustical Society of Japan (ASJ).

Ryo Mukai (A'95–M'01) received the B.S. and M.S. degrees in information science from the University of Tokyo, Tokyo, Japan, in 1990 and 1992, respectively.

He joined NTT in 1992. From 1992 to 2000, he was engaged in research and development of processor architecture for network service systems and distributed network systems. Since 2000, he has been with NTT Communication Science Laboratories, where he is engaged in research of blind source separation. His current research interests include digital signal processing and its applications.

Mr. Mukai is a Member of ACM, the Acoustical Society of Japan (ASJ), and the Information Processing Society of Japan (IPSJ).

Shoji Makino (A'89–M'90–SM'99) was born in Nikko, Japan, on June 4, 1956. He received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1979, 1981, and 1993, respectively.

He joined the Electrical Communication Laboratory of Nippon Telegraph and Telephone Corporation (NTT) in 1981. Since then, he has been engaged in research and development of acoustic echo cancellation and adaptive algorithms. He is now a Senior Research Scientist, Supervisor, and Group Leader at the Speech Open Laboratory of the NTT Communication Science Laboratories. His research interests include blind source separation of convolutive mixtures of speech, acoustic signal processing, and adaptive filtering and its applications. He is the author or co-author of more than 170 articles in journals and conference proceedings and has been responsible for more than 140 patents.

Dr. Makino received the Paper Award of the Institute of Electronics, Information, and Communication Engineers of Japan in 2002, the Paper Award of the Acoustical Society of Japan in 2002, the Achievement Award of the Institute of Electronics, Information, and Communication Engineers of Japan in 1997, and the Outstanding Technological Development Award of the Acoustical Society of Japan in 1995. He is a member of the Conference Board of the IEEE Signal Processing Society and an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He is a member of the Technical Committee on Audio and Electroacoustics as well as the Technical Committee on Speech of the IEEE Signal Processing Society. He is the General Chair of the 2003 International Workshop on Acoustic Echo and Noise Control and the Organizing Chair of the 2003 International Conference on Independent Component Analysis and Blind Signal Separation. He served on the Program Committee of the 2003 IEEE International Workshop on Neural Networks for Signal Processing, the Program Committee of the 2004 International Congress on Acoustics, the Program Committee of the 2002 IEEE International Workshop on Neural Networks for Signal Processing, the Organizing Committee of the 2002 China-Japan Joint Conference on Acoustics, and the Technical Committee of the 2001 and 1999 International Workshop on Acoustic Echo and Noise Control. He is a Vice Chair of the Technical Committee on Engineering Acoustics of the Institute of Electronics, Information, and Communication Engineers of Japan. He is a member of the Acoustical Society of Japan and the Institute of Electronics, Information, and Communication Engineers of Japan.

Tsuyoki Nishikawa was born in Mie, Japan, in 1978. He received the B.E. degrees in electronic system and information engineering from Kinki University in 2000 and received the M.E. degrees in information and science from Nara Institute of Science and Technology (NAIST) in 2002. He is now pursuing the Ph.D. degree at Graduate School of Information Science, NAIST.

His research interests include array signal processing and blind source separation.

Mr. Nishikawa is a member of the ASJ.

Hiroshi Saruwatari (M'00) was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively.

He joined Intelligent Systems Laboratory, Secom Co., Ltd., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development on the ultrasonic array system for the acoustic imaging. He is currently an Associate Professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction.

Dr. Saruwatari received the Paper Award from IEICE in 2001. He is a Member of IEICE and the ASJ.