# AN APPROACH TO CONVOLUTIVE BACKWARD-MODEL BLIND SOURCE SEPARATION BASED ON JOINT DIAGONALIZATION

*Shinya Saito*[1], *Kunio Oishi*[2], *and Toshihiro Furukawa*[1]

[1]Department of Management Science, Tokyo University of Science
1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
[2]School of Computer Science, Tokyo University of Technology
1404-1 Katakura, Hachioji, Tokyo 192-0982, Japan
Email: sinyasaito@ms.kagu.tus.ac.jp, kohishi@cs.teu.ac.jp, furukawa@ms.kagu.tus.ac.jp

## ABSTRACT

A convolutive frequency-domain backward-model blind source separation (BSS) for directly estimating the unmixing matrix by solving a block-by-block least-squares approximate joint diagonalization (AJD) problem is presented. In the new backward-model BSS, the inverse of an exponentially weighted cross-spectral density matrix of the observed signal is calculated at each frequency bin. The expansion of the inverse matrix can lead to a criterion for applying the alternating least-squares with projection (ALSP) algorithm to the backward-model BSS. Introducing the block-processing technique into the least-squares AJD (LS-AJD) problem is effective to reduce computational burden per iteration at each block frame. This new BSS does not need to solve the scaling ambiguity by other methods due to the scale constraint. The interfrequency correlation is used to prevent misalignment permutation for the new BSS. Finally, we compare it with the conventional forward-model BSS in both low and high signal-to-noise ratio (SNR) environments and show that this new BSS improves robustness.

***Index Terms***— Blind source separation (BSS), convolutive audio mixture, joint diagonalization, alternating least-squares (ALS) algorithm, block-processing technique

## 1. INTRODUCTION

Blind source separation (BSS) is a technique to recover source signals from the observed signals that are modeled as an unknown convolutive audio mixture of unknown quasi-stationary source signals, where quasi-stationary signals are modeled as an approximately stationary behavior over short time interval that is called an epoch. Forward-model BSS is to find a mixing matrix by minimizing a least-squares criterion, and then to compute an inverse of the matrix. The minimization of the criterion is mathematically equivalent to jointly approximately diagonalizing the cross-spectral density matrices of the observed signals. Approximate joint diagonalization (AJD) problem [1, 2] is to find the diagonalizing matrix and diagonal matrices. Alternating least-squares (ALS) algorithm [2, 3] is a well-known technique for solving the AJD problem. Recently, the alternating least-squares with projection (ALSP) algorithm for convolutive forward-model BSS in frequency domain has been developed in [1] by expressing the least-squares criterion by Khatri-Rao (KR) product. The convolutive frequency-domain least-squares AJD (LS-AJD) based forward-model BSS [1] is accomplished in the following four steps: 1) Estimate the mixing matrix. 2) Find the unmixing matrix from that. 3) Correct the frequency-dependent arbitrary scaled unmixing matrix. 4) Resolve the frequency-dependent permutation ambiguity.

Backward-model BSS is to directly find an unmixing matrix by minimizing a least-squares criterion. In this paper, we present an approach to LS-AJD-based backward-model BSS of convolutive audio mixtures. The new approach can estimate the unmixing matrix by finding a diagonalizing matrix and diagonal matrices from the inverse of the cross-spectral density matrices of the observed signals. The expansion of the inverse matrix is allowed to adopt the ALSP algorithm in the backward-model BSS. In the new BSS, solving the LS-AJD problem is equivalent to estimating the diagonalizing the matrices and solving the scale problem simultaneously. We introduce the block-processing technique [4] into the LS-AJD problem to reduce computational burden. The interfrequency correlation is used to solve the permutation problem in this new BSS. The new backward-model BSS has the advantage of only two steps: 1) Estimate the unmixing matrix directly. 2) Solve the permutation. The separation performance of the new BSS with setting the number of sensors to that of sources is demonstrated by our real room experiments. Furthermore, we show that this new BSS provides a low level of misadjustment.

## 2. PROBLEM FORMULATION AND LS-AJD-BASED FORWARD-MODEL BSS

In the convolutive mixing model between $N$ sources $s_1(t), s_2(t), \cdots, s_N(t)$ and $J$ sensors $x_1(t), x_2(t), \cdots, x_J(t)$ at time $t$, we obtain the observed signal at the $i$th sensor as $x_i(t) =$

$\sum_{j=1}^{N} h_{ij}(t) * s_j(t) + n_i(t)$, where the sources are zero mean, second-order quasi-stationary signals [1], the $N$ sources are independent of each other, $J \geq N \geq 2$, $h_{ij}(t)$ is the impulse response from the $j$th source to the $i$th sensor without changing over the entire observation interval, the asterisk $*$ denotes time-domain convolution, and the additive white Gaussian noise (AWGN) $n_i(t)$ with mean zero and variance $\sigma^2$ is independent of the sources. The time-domain observed signal is transformed into the time-frequency domain by the short-time Fourier transform (STFT). One can write $x_i(\omega_k, m)$, where $\omega_k = 2\pi k/K$, $k = 0, 1, \cdots, K-1$ and $m$ is a time frame index. Let $T_s$ be shift size between two neighboring windows. All observed signals are available in the time frame interval $1 \leq m \leq M$, where $M$ is the total number of time frames. If $K$ is significantly larger than the length of the mixing-filter impulse response $h_{ij}(t)$, the time-domain convolution is approximately converted to the following multiplication:

$$\mathbf{x}(\omega_k, m) \approx \mathbf{H}(\omega_k)\mathbf{s}(\omega_k, m) + \mathbf{n}(\omega_k, m) \tag{1}$$

where $s_j(\omega_k, m)$ and $n_i(\omega_k, m)$ are the STFTs of $s_j(t)$ and $n_i(t)$ at time frame $m$, $h_{ij}(\omega_k)$ is the discrete Fourier transform (DFT) of $h_{ij}(t)$, $\mathbf{s}(\omega_k, m) = [s_1(\omega_k, m), s_2(\omega_k, m), \cdots, s_N(\omega_k, m)]^T$ is the $N \times 1$ vector of sources, $\mathbf{H}(\omega_k)$ is the $J \times N$ mixing matrix of the transfer function from the $N$ sources to the $J$ sensors, the $J \times 1$ observed signal vector is defined by $\mathbf{x}(\omega_k, m) = [x_1(\omega_k, m), x_2(\omega_k, m), \cdots, x_J(\omega_k, m)]^T$, and $\mathbf{n}(\omega_k, m) = [n_1(\omega_k, m), n_2(\omega_k, m), \cdots, n_J(\omega_k, m)]^T$ is the $J \times 1$ vector of AWGN. The superscript $^T$ denotes transpose. The cross-spectral density matrix of the source signal $\mathbf{P}_s(\omega_k, m) = E[\mathbf{s}(\omega_k, m)\mathbf{s}(\omega_k, m)^H] \in \mathbb{R}^{N \times N}$ is diagonal, where $E[\cdot]$ and the superscript $^H$ denote expectation operation and Hermitian transpose respectively.

In order to separate the sources at each frequency bin $\omega_k$ independently in BSS, pre-multiplication of $\mathbf{H}(\omega_k)$ by the $N \times J$ unmixing matrix $\mathbf{W}(\omega_k)$ yields

$$\mathbf{W}(\omega_k)\mathbf{H}(\omega_k) = \mathbf{\Pi}(\omega_k)\mathbf{D}(\omega_k) \tag{2}$$

where $\mathbf{\Pi}(\omega_k) \in \mathbb{R}^{N \times N}$ is a frequency-dependent permutation matrix and $\mathbf{D}(\omega_k) \in \mathbb{C}^{N \times N}$ is a scale or phase arbitrary diagonal matrix.

Let $\mathbf{P}_x(\omega_k, m) \in \mathbb{C}^{J \times J}$ define the cross-spectral density matrix of the observed signal at point $(\omega_k, m)$

$$\mathbf{P}_x(\omega_k, m) = \mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H + \sigma^2\mathbf{I} \tag{3}$$

where $\mathbf{I}$ denotes the $J \times J$ identity matrix and we assume that $\mathbf{H}(\omega_k)$ and $\mathbf{P}_s(\omega_k, m)$ are nonsingular. If we find a diagonalizing matrix $\mathbf{B}(\omega_k) \in \mathbb{C}^{J \times N}$ and diagonal matrices $\mathbf{\Lambda}(\omega_k, m) \in \mathbb{C}^{N \times N}$ to satisfy $\mathbf{P}_x(\omega_k, m) - \sigma^2\mathbf{I} = \mathbf{B}(\omega_k)\mathbf{\Lambda}(\omega_k, m)\mathbf{B}(\omega_k)^H$ with the scale constraint $\|\mathbf{b}_i(\omega_k)\|_2 = 1$, where $\mathbf{b}_i(\omega_k)$ is the $i$th column of $\mathbf{B}(\omega_k)$ and $\|\cdot\|_2$ denotes Euclidean norm, from (2), the relationship between $\mathbf{B}(\omega_k)$ and $\mathbf{H}(\omega_k)$ becomes $\mathbf{B}(\omega_k) = \mathbf{H}(\omega_k)\mathbf{D}(\omega_k)^{-1}\mathbf{\Pi}(\omega_k)^{-1}$, where $\mathbf{W}(\omega_k)\mathbf{B}(\omega_k) = \mathbf{I}$.

In the frequency-domain forward-model BSS [1], in order to approximate the cross-spectral density matrix of the observed signal, after the $\Xi$ estimated power spectral density matrices are obtained by dividing the all observed signals into $\Xi$ epochs with the Welch periodogram method, the estimation value is normalized. By using the normalized estimation value $\hat{\mathbf{P}}_x(\omega_k, \xi)$, the measurement error at epoch $\xi$ is obtained by

$$\mathbf{E}(\omega_k, \xi) = \hat{\mathbf{P}}_x(\omega_k, \xi) - \mathbf{B}(\omega_k)\mathbf{\Lambda}(\omega_k, \xi)\mathbf{B}(\omega_k)^H. \tag{4}$$

The LS-AJD problem is to find a diagonalizing matrix $\hat{\mathbf{B}}(\omega_k)$ and $\Xi$ associated diagonal matrices $\hat{\mathbf{\Lambda}}(\omega_k, 1), \hat{\mathbf{\Lambda}}(\omega_k, 2), \cdots, \hat{\mathbf{\Lambda}}$ $(\omega_k, \Xi)$ by minimizing the sum of the measurement error squares

$$\hat{\mathbf{B}}(\omega_k), \hat{\mathbf{\Lambda}}(\omega_k, \xi) = \underset{\substack{\mathbf{B}(\omega_k),\ \mathbf{\Lambda}(\omega_k, \xi) \\ \|\mathbf{b}_i(\omega_k)\|_2 = 1}}{\operatorname{argmin}} \sum_{\xi=1}^{\Xi} \|\mathbf{E}(\omega_k, \xi)\|_F^2 \tag{5}$$

with the scale constraint $\|\mathbf{b}_i(\omega_k)\|_2 = 1$ over significantly large number of epochs $\Xi$ at each frequency bin $\omega_k$, where $\|\cdot\|_F$ denotes the Frobenius norm. The unmixing matrix $\hat{\mathbf{W}}(\omega_k)$ is obtained by the pseudo inverse of the matrix $\hat{\mathbf{B}}(\omega_k)$

$$\hat{\mathbf{W}}(\omega_k) = \left(\hat{\mathbf{B}}(\omega_k)^H\hat{\mathbf{B}}(\omega_k)\right)^{-1}\hat{\mathbf{B}}(\omega_k)^H. \tag{6}$$

## 3. BLOCK-PROCESSING LS-AJD CRITERIA FOR BACKWARD-MODEL BSS

The conventional backward-model BSS algorithm for minimizing the sum of the measurement error squares has been proposed in [5], where the measurement error is obtained by $\mathbf{W}(\omega_k)\mathbf{P}_x(\omega_k, m)\mathbf{W}(\omega_k)^H - \mathbf{P}_s(\omega_k, m)$. The ALSP algorithm can not be used to solve the criterion because $\mathbf{P}_x(\omega_k, m)$ is not a diagonal matrix in the form $\mathbf{W}(\omega_k)\mathbf{P}_x(\omega_k, m)\mathbf{W}(\omega_k)^H$. In order to apply the ALSP algorithm to a backward-model BSS, we show that the sum of block-by-block measurement error squares can be expressed by the expansion of the inverse of the cross-spectral density matrix. Whereas a noise-free cross-spectral density matrix can be obtained as $\mathbf{P}_x(\omega_k, m) - \sigma^2\mathbf{I} = \mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H$ for the number of sensors being larger than that of sources, where $\sigma^2$ is the smallest eigenvalue of the matrix $\mathbf{P}_x(\omega_k, m)$, it can not be obtained for the number of sensors being equal to that of sources. Therefore, in this section, we set the number of sensors to that of sources.

We expand the inverse of the cross-spectral density matrix of the observed signal at point $(\omega_k, m)$ as

$$\begin{aligned}\mathbf{P}_x(\omega_k, m)^{-1} &= \left(\mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H + \sigma^2\mathbf{I}\right)^{-1} \\ &= \left(\mathbf{H}(\omega_k)^H\right)^{-1}\mathbf{P}_s(\omega_k, m)^{-1}\mathbf{H}(\omega_k)^{-1} \\ &-\sigma^2\sum_{\ell=0}^{\infty}(-1)^{\ell}\sigma^{2\ell}\left[\left(\mathbf{H}(\omega_k)^H\right)^{-1}\mathbf{P}_s(\omega_k, m)^{-1}\mathbf{H}(\omega_k)^{-1}\right]^{\ell+2}.\end{aligned} \tag{7}$$

We discuss the proof of (7). According to the matrix inversion lemma, the inverse of $\mathbf{P}_x(\omega_k, m)$ can be expressed as

$$\mathbf{P}_x(\omega_k, m)^{-1} = \left(\mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H + \sigma^2\mathbf{I}\right)^{-1}$$
$$= \left(\mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H\right)^{-1} - \sigma^2\left(\mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H\right)^{-1}$$
$$\cdot \left[\mathbf{I} + \sigma^2\left(\mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H\right)^{-1}\right]^{-1}$$
$$\cdot \left(\mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H\right)^{-1}. \tag{8}$$

If $\sigma^2$ is smaller than the smallest eigenvalue of $\mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)$ $\mathbf{H}(\omega_k)^H$, the series will be convergent as follows:

$$\sum_{\ell=0}^{\infty}\left[-\sigma^2\left(\mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H\right)^{-1}\right]^{\ell}$$
$$= \left[\mathbf{I} + \sigma^2\left(\mathbf{H}(\omega_k)\mathbf{P}_s(\omega_k, m)\mathbf{H}(\omega_k)^H\right)^{-1}\right]^{-1}. \tag{9}$$

Hence, substituting (9) in the second term on the right side of (8), we get (7). If we find a diagonalizing matrix $\mathbf{W}(\omega_k) \in \mathbb{C}^{N\times N}$ and diagonal matrices $\mathbf{\Lambda}(\omega_k, m)^{-1} \in \mathbb{C}^{N\times N}$ to satisfy

$$\mathbf{P}_x(\omega_k, m)^{-1} + \sigma^2\sum_{\ell=0}^{\infty}(-1)^{\ell}\sigma^{2\ell}\left[\left(\mathbf{H}(\omega_k)^H\right)^{-1}\mathbf{P}_s(\omega_k, m)^{-1}\mathbf{H}(\omega_k)^{-1}\right]^{\ell+2}$$
$$= \mathbf{W}(\omega_k)^H\mathbf{\Lambda}(\omega_k, m)^{-1}\mathbf{W}(\omega_k)$$

with the scale constraint $||\mathbf{w}_i(\omega_k)||_2 = 1$, where $\mathbf{w}_i(\omega_k)$ is the $i$th row of $\mathbf{W}(\omega_k)$, from (2), the relationship between $\mathbf{W}(\omega_k)$ and $\mathbf{H}(\omega_k)^{-1}$ becomes $\mathbf{W}(\omega_k) = \mathbf{\Pi}(\omega_k)\mathbf{D}(\omega_k)\mathbf{H}(\omega_k)^{-1}$. Let $\mathcal{E}(\omega_k, m)$ define a measurement error at point $(\omega_k, m)$

$$\mathcal{E}(\omega_k, m) = \hat{\mathbf{P}}_x(\omega_k, m)^{-1} - \mathbf{W}(\omega_k)^H\mathbf{\Lambda}(\omega_k, m)^{-1}\mathbf{W}(\omega_k) \tag{10}$$

where a procedure to obtain $\hat{\mathbf{P}}_x(\omega_k, m)^{-1}$ is given later. We use the LS-AJD problem to find a diagonalizing matrix $\hat{\mathbf{W}}(\omega_k)$ and $M$ associated diagonal matrices $\hat{\mathbf{\Lambda}}(\omega_k, 1)^{-1}, \hat{\mathbf{\Lambda}}(\omega_k, 2)^{-1}, \cdots, \hat{\mathbf{\Lambda}}(\omega_k, M)^{-1}$ by minimizing the sum of the measurement error squares

$$\hat{\mathbf{W}}(\omega_k), \hat{\mathbf{\Lambda}}(\omega_k, m)^{-1} = \underset{\substack{\mathbf{W}(\omega_k),\ \mathbf{\Lambda}(\omega_k, m)^{-1} \\ ||\mathbf{w}_i(\omega_k)||_2 = 1}}{\operatorname{argmin}} \sum_{m=1}^{M} ||\mathcal{E}(\omega_k, m)||_F^2 \tag{11}$$

with the scale constraint $||\mathbf{w}_i(\omega_k)||_2 = 1$ over significantly large number of time frames $M$ at each frequency bin $\omega_k$. We use the exponentially weighted cross-spectral density matrix of the observed signal to provide the effect of a short-term memory in the estimate of $\mathbf{P}_x(\omega_k, m)$ at point $(\omega_k, m)$

$$\hat{\mathbf{P}}(\omega_k, m) = \beta\hat{\mathbf{P}}(\omega_k, m-1) + \mathbf{x}(\omega_k, m)\mathbf{x}(\omega_k, m)^H \tag{12}$$

where the forgetting factor $\beta$ is a positive constant close to, but less than unity. $\hat{\mathbf{P}}(\omega_k, 0) = c\mathbf{I}$ and $c$ is a small positive constant. $\mathbf{P}_x(\omega_k, m)^{-1}$ is estimated by normalizing $\hat{\mathbf{P}}(\omega_k, m)^{-1}$ as $\hat{\mathbf{P}}_x(\omega_k, m)^{-1} = \hat{\mathbf{P}}(\omega_k, m)^{-1}/||\hat{\mathbf{P}}(\omega_k, m)^{-1}||_F$, where $\hat{\mathbf{P}}(\omega_k, m)^{-1}$ is the inverse of $\hat{\mathbf{P}}(\omega_k, m)$.

**Table 1**. Procedure for estimating the unmixing matrix.

1. Update $\hat{\mathbf{P}}(\omega_k, m)$ at each frame, as given by (12).
2. Compute $\hat{\mathbf{P}}(\omega_k, \tau L)^{-1}$ at each block frame using the (13).
3. Estimate $\mathbf{W}(\omega_k)$ by the ALSP algorithm at each block frame as shown in (14), (18), (20), and (22) in [1].
4. Go to step 1 for $\tau = 1, 2, \cdots, \lfloor M/L \rfloor$.

The significantly large $M$ increases computational burden to compute the inverse matrix $\hat{\mathbf{P}}(\omega_k, m)^{-1}$, $M$ diagonal matrices $\hat{\mathbf{\Lambda}}(\omega_k, 1)^{-1}, \hat{\mathbf{\Lambda}}(\omega_k, 2)^{-1}, \cdots, \hat{\mathbf{\Lambda}}(\omega_k, M)^{-1}$, and the diagonalizing matrix $\hat{\mathbf{W}}(\omega_k)$ at each time frame. In order to reduce computational burden, we introduce the block-processing technique [4] to find the LS-AJD estimate by inverting the exponentially weighted cross-spectral density matrix of the observed signal at each block time frame

$$\hat{\mathbf{P}}_x(\omega_k, \tau L)^{-1} = \frac{\hat{\mathbf{P}}(\omega_k, \tau L)^{-1}}{||\hat{\mathbf{P}}(\omega_k, \tau L)^{-1}||_F};\ \tau = 1, 2, \cdots, \lfloor M/L \rfloor \tag{13}$$

where block length $L$ is set to unity or positive integer larger than unity and $\lfloor x \rfloor$ is the largest integer less than or equal to $x$. Let us define a measurement error using the block-processing technique, similar to (10), at each block frame

$$\mathcal{E}(\omega_k, \tau L) = \hat{\mathbf{P}}_x(\omega_k, \tau L)^{-1} - \mathbf{W}(\omega_k)^H\mathbf{\Lambda}(\omega_k, \tau L)^{-1}\mathbf{W}(\omega_k). \tag{14}$$

By using the $\lfloor M/L \rfloor$ block-by-block inverse matrices $\hat{\mathbf{P}}_x(\omega_k, \tau L)^{-1}$, the block-processing LS-AJD problem is to minimize the sum of the block-by-block measurement error squares

$$\hat{\mathbf{W}}(\omega_k), \hat{\mathbf{\Lambda}}(\omega_k, \tau L)^{-1} = \underset{\substack{\mathbf{W}(\omega_k),\ \mathbf{\Lambda}(\omega_k, \tau L)^{-1} \\ ||\mathbf{w}_i(\omega_k)||_2 = 1}}{\operatorname{argmin}} \sum_{\tau=1}^{\lfloor M/L \rfloor} ||\mathcal{E}(\omega_k, \tau L)||_F^2 \tag{15}$$

with the scale constraint $||\mathbf{w}_i(\omega_k)||_2 = 1$. Consequently, the block-processing LS-AJD estimate is a diagonalizing matrix $\hat{\mathbf{W}}(\omega_k)$ and $\lfloor M/L \rfloor$ diagonal matrices $\hat{\mathbf{\Lambda}}(\omega_k, L)^{-1}, \hat{\mathbf{\Lambda}}(\omega_k, 2L)^{-1}, \cdots, \hat{\mathbf{\Lambda}}(\omega_k, L\lfloor M/L \rfloor)^{-1}$.

The procedure for solving (15) is shown in Table 1. The ALSP algorithm alternates the two phases [1]. On phase one, the method of least squares minimizes the sum of the measurement error squares to find $\hat{\mathbf{W}}(\omega_k)^H \odot \hat{\mathbf{W}}(\omega_k)$, where $\odot$ denotes the KR product and $N$ matrices of size $N^2 \times \lfloor M/L \rfloor$ are used to rewrite the criterion. It requires $O\left(\lfloor M/L \rfloor N^3\right)$ computational operations per iteration. $\hat{\mathbf{W}}(\omega_k)$ is estimated from $\hat{\mathbf{W}}(\omega_k)^H \odot \hat{\mathbf{W}}(\omega_k)$ by the power method. The power method is continued until $\hat{\mathbf{w}}_j(\omega_k)$ changes by less than $\epsilon_P$ between iterations. The method of least squares and the power method are repeated until $\hat{\mathbf{W}}(\omega_k)^H \odot \hat{\mathbf{W}}(\omega_k)$ changes by less than $\epsilon_G$ between iterations. On the other phase, the $\lfloor M/L \rfloor$ pseudo-inverse matrices of size $N \times N^2$ are calculated to find $\hat{\mathbf{\Lambda}}(\omega_k, m)^{-1}$ by the method of least squares. The ALSP algorithm is repeated until $\sum_{\tau=1}^{\lfloor M/L \rfloor} ||\mathcal{E}(\omega_k, \tau L)||_F^2$ changes by less than $\epsilon_C$ between iterations. It requires $O\left(\lfloor M/L \rfloor N^3\right)$ computational operations per iteration.

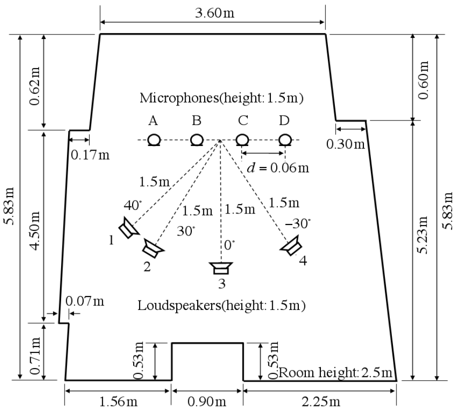


**Fig. 1**. Experimental setup (300-ms reverberation).

## 4. EXPERIMENTAL RESULTS

The configuration of the room is shown in Fig.1. After measuring impulse responses of the room at 8 kHz and 16-bit sampling format, observed signals were obtained by convolving audio speech data with the impulse response. The data set consisted of 12 seconds long speech for $J = N = 2$, 30 seconds for $J = N = 3$, and 100 seconds for $J = N = 4$. The direction-of-arrivals (DOAs) of the sources were 30° and −30° on a two-microphone linear array, 30°, 0°, and −30° on a three-microphone linear array, and 40°, 30°, 0°, and −30° on a four-element linear array. The Hanning window was used for the STFT. The parameter was chosen empirically as 2048-point FFT, the forgetting factor of $\beta = 0.99$, and the small positive constant of $c = 10^{-2}$. The ALSP algorithm continuously ran until the change between iterations was less than $\epsilon_P = 10^{-15}$ and $\epsilon_G = \epsilon_C = 10^{-6}$. The new backward-model BSS was compared with the conventional forward-model BSS [1]. The difference between the new and the conventional BSSs is how to estimate the unmixing matrix. In the conventional BSS, the scale problem was solved by normalizing each row vector of $\hat{\mathbf{W}}(\omega_k)$ at each frequency bin. The new and the conventional BSSs solved the permutation problem by the approach based on correlation [6].

Let input signal-to-interference ratio (SIR) at the $i$th observed signal and output SIR at the $i$th output signal define

$$\mathrm{SIR}_{x_i} = 10\log_{10}\frac{E\left[(h_{ii}(t) * s_i(t))^2\right]}{E\left[\left(\sum_{\substack{j=1\\ j\neq i}}^{N} h_{ij}(t) * s_j(t) + \sum_{j=1}^{J} n_j(t)\right)^2\right]}$$

$$\mathrm{SIR}_{y_i}=10\log_{10}\frac{E\left[(\gamma_{ii}(t)*s_i(t))^2\right]}{E\left[\left(\sum_{\substack{j=1\\ j\neq i}}^{N}\gamma_{ij}(t)*s_j(t) + \sum_{j=1}^{J}\hat{w}_{ij}(t)*n_j(t)\right)^2\right]}$$

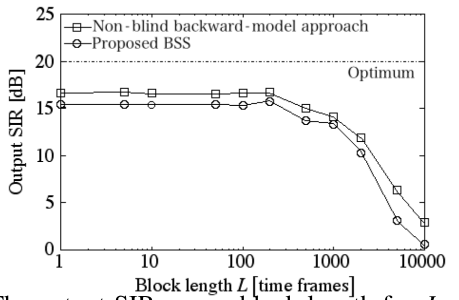


**Fig. 2**. The output SIR versus block length for $J = N = 2$, $M = 12000$, $T_s = 8$, and SNR ≈ 20 dB.

where $\gamma_{ij}(t)$ and $\hat{w}_{ij}(t)$ are the IDFTs of the matrices $\Gamma(\omega_k) = \mathbf{D}(\omega_k)^{-1}\mathbf{\Pi}(\omega_k)^{-1}\hat{\mathbf{W}}(\omega_k)\mathbf{H}(\omega_k)$ and the unmixing matrices $\hat{\mathbf{W}}(\omega_k)$. The SNR is determined by the ratio of the desired signal power and the power of the interference plus noise component in the output signal after obtaining an optimum unmixing matrix $\mathbf{W}_{\mathrm{opt}}(\omega_k)$ and an optimum permutation $\mathbf{\Pi}_{\mathrm{opt}}(\boldsymbol{\omega_k})$. When the room impulse response $h_{ij}(t)$ is available, after calculating $\mathbf{H}(\omega_k)^{-1}$, the optimum unmixing matrix is obtained by normalizing each row vector of $\mathbf{H}(\omega_k)^{-1}$ that is mathematically equivalent to solving the scale problem. Similarly, when $\mathbf{H}(\omega_k)$ and $\mathbf{W}_{\mathrm{opt}}(\omega_k)$ are available, the optimum permutation is obtained by

$$\mathbf{\Pi}_{\mathrm{opt}}(\omega_k) = \underset{\mathbf{\Pi}(\omega_k)}{\mathrm{argmax}}\left\|\mathrm{diag}\left(\mathbf{\Pi}(\omega_k)\mathbf{C}_{\mathrm{opt}}(\omega_k)\right)\right\|_F^2 = \mathbf{I} \qquad (16)$$

where $\mathbf{C}_{\mathrm{opt}}(\omega_k) = \mathbf{W}_{\mathrm{opt}}(\omega_k)\mathbf{H}(\omega_k)$ and $\mathrm{diag}(\mathbf{A})$ denotes the diagonal matrix of $\mathbf{A}$. Thus, the SNR is equal to the optimum output SIR.

First, the averaged output SIR versus block length is illustrated in Fig.2. The averaged input SIR was about 0.01 dB. Fig.2 also depicts the overall output SIR of the nonblind depermutation backward-model approach, when $\mathbf{H}(\omega_k)$ is available, that is, $\hat{\mathbf{W}}(\omega_k)$ is permuted by

$$\mathbf{\Pi}(\omega_k) = \underset{\mathbf{\Pi}(\omega_k)}{\mathrm{argmax}}\left\|\mathrm{diag}\left(\mathbf{\Pi}(\omega_k)\hat{\mathbf{C}}(\omega_k)\right)\right\|_F^2 \qquad (17)$$

after obtaining the block-processing LS-AJD estimate $\hat{\mathbf{W}}(\omega_k)$, where $\hat{\mathbf{C}}(\omega_k) = \hat{\mathbf{W}}(\omega_k)\mathbf{H}(\omega_k)$. We could achieve a good separation performance for short block length almost same as for $L = 1$ whereas the performance is degraded for block length longer than 200 time frames. The separation performance for $L = 200$ is just below the nonblind de-permutation approach while the proposed BSS for $L = 1$ is 200 times as complex as that for $L = 200$.

Second, we applied the new BSS on multiple sources. Comparison of the nonblind de-permutation forward- and backward-model approaches and the new BSS, and the conventional BSS for $J = N = 2$ is shown in Fig.3(a), where $T_e$ denotes the epoch size for the conventional forward-model approach. In the SNR higher than 15 dB, both nonblind de-permutation approaches could not achieve the optimum output SIR because both LS-AJD estimation errors are dominant and the ambient noise is negligible. Meanwhile, the

**Table 2**. Comparison of the output SIR with the conventional backward-model BSS [5] for $J = N = 2$, $L = 200$, and SNR $\approx$ 30 dB. The reverberation time was about 270 ms [7].

| Parra's method | Proposed BSS |
|---|---|
| 1.65 dB | 16.72 dB |

background noise dominates both LS-AJD estimation errors in the low SNR range. Compared to the conventional BSS, the new BSS is at most up 1.31 dB on whether it is a robust approach. Fig.3(b) compares the two nonblind de-permutation approaches and the two BSSs for $J = N = 3$. In the SNR higher than 10 dB, both nonblind de-permutation approaches could not achieve the optimum output SIR. Note that the output SIR of the proposed BSS is 0.08 to 1.2 dB above the conventional BSS in both low and high SNR environments. Fig.3(c) shows the result of comparison of the nonblind de-permutation approaches and the BSSs for $J = N = 4$. It can be observed that the new BSS is about 0.55 to 1.62 dB higher than the conventional BSS.

Finally, in Table 2, we compare the new BSS with the conventional gradient-based backward-model BSS [5] with parameter setting described in [7]. We only changed the impulse response of a room artificially generated by the image method, where the size was 16.6 × 11.2 × 8.0 ft, two sources and sensors were located at the same position as in [7], and the reflection coefficient was set to 0.7. The averaged input SIR was about −0.67 dB. The new BSS performs quite well.

## 5. CONCLUSION

We have proposed an approach to backward-model BSS of convolutive audio mixtures based on a block-by-block LS-AJD problem. The inverse of an exponentially weighted cross-spectral density matrix of the observed signal is calculated at each block frame. The ALSP algorithm could be applied to the backward-model BSS by the expansion of the inverse matrix. It was shown that the block-processing technique is effective to reduce computational burden. This new BSS does not need to solve the scaling ambiguity because of the scale constraint. The interfrequency correlation was used to overcome the permutation ambiguity. The experimental results have shown that the new BSS is superior in its performance to the conventional forward-model BSS in both low and high SNR environments.

## 6. REFERENCES

[1] K. Rahbar and J. P. Reilly, "A frequency domain method for blind source separation of convolutive audio mixtures," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 832–844, Sept. 2005.

[2] S. Dégerine and E. Kame, "A comparative study of approximate joint diagonalization algorithms for blind source separation in presence of additive noise," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 3022–3031, June 2007.

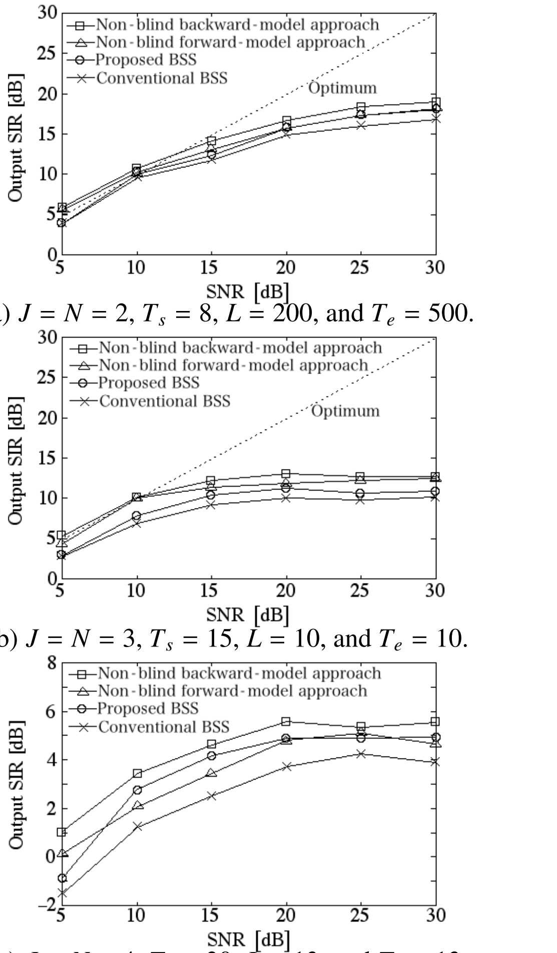
[3] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 810–823, Mar. 2000.

[4] J. C. Lee and C. K. Un, "Performance analysis of frequency-domain block LMS adaptive digital filters," *IEEE Trans. Circuits and Syst.*, vol. CAS-36, no. 2, pp. 173–189, Feb. 1989.

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

[6] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals,"*Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[7] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: investigation and solutions,"*IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 1–13, Jan. 2005.



(a) $J = N = 2$, $T_s = 8$, $L = 200$, and $T_e = 500$.

(b) $J = N = 3$, $T_s = 15$, $L = 10$, and $T_e = 10$.

(c) $J = N = 4$, $T_s = 30$, $L = 13$, and $T_e = 13$.

**Fig. 3**. The output SIR versus SNR. The overall input SIR was about −0.14 dB for $J = N = 2$, −3.52 dB for $J = N = 3$, and −5.87 dB for $J = N = 4$.