Determined and overdetermined convolutive blind source extractions by approximate joint diagonalization

Shinya Saito\(^1\), Kunio Oishi\(^2\) and Toshihiro Furukawa\(^1\)

\(^1\)Department of Management Science, Tokyo University of Science, 6–3–1 Niijuku, Katsushika-ku, Tokyo, 125–8585 Japan
\(^2\)School of Computer Science, Tokyo University of Technology, 1404–1 Katakura, Hachioji, 192–0982 Japan

(Received 24 July 2018, Accepted for publication 19 April 2019)

Abstract: Blind extraction of convolutive speech mixtures can be achieved by the approximate joint diagonalization (AJD) approach. In this paper, we present a least-squares AJD (LS-AJD) algorithm, called fast diagonalization, implemented by minimizing the direct and indirect LS criteria (FDMDI) algorithm. The proposed approach is based on an alternate minimization of the indirect and direct least-squares criteria to the diagonal matrices in the first phase and to the mixing matrix in the second phase, respectively. In our proposed approach, the constrained LS-AJD estimation problem is solved by the method of Lagrange multipliers; moreover, the mixing matrix is estimated by a noniterative method without a nested loop in the second phase. The simulation result demonstrates that overdetermined FDMDI blind source extraction (BSE) provides more effective extracted signals than determined FDMDI BSE in an actual acoustic environment.

Keywords: Blind source extraction (BSE), Overdetermined BSE, Approximate joint diagonalization (AJD), Room acoustics

PACS number: 43.60. –c [doi:10.1250/ast.40.302]

1. INTRODUCTION

Over the past decade, joint diagonalization (JD) of a set of target matrices has been widely applied in blind channel identification, hearing aid equipment, speech enhancement, and blind source separation. In real-world speech and audio applications, there are a limited number of available estimated target matrices. This is called the approximate JD (AJD) problem. An example of those applications is blind source extraction (BSE) for recovering original speech source signals from observed signals without knowing the convolutive mixing process in a noisy and reverberant environment [1–4].

AJD problems are solved by alternating least-squares (ALS) algorithms [4–12]. The ALS algorithm is an approach for minimizing the least-squares (LS) criterion [5,7,9,10]. Generally, the ALS algorithm proceeds in two phases. In the first phase, the ALS algorithm minimizes the criterion with respect to diagonal matrices while keeping a diagonalizing matrix fixed. In the second phase, it minimizes the criterion with respect to the diagonalizing matrix while keeping the diagonal matrices fixed. These two phases are repeated alternately. The primary drawback of the ALS algorithm is its slow convergence. Therefore, acceleration of the convergence of the ALS algorithm is one of the most important tasks for reducing the computational load of the ALS algorithm.

The accuracy of the solution in the AJD problem is dependent on the number of estimated target matrices. A large number of estimated target matrices improves the accuracy of AJD estimates. However, it increases the computational load. Although a small number of estimated target matrices reduces the computational load, it degrades the accuracy of the AJD estimate. Apparently, an undesirable tradeoff exists between computational complexity and accuracy of AJD estimates. In determined BSE, the number of speech sources is assumed to be equal to that of microphones. When the number of microphones exceeds the number of speech sources, this is called overdetermined BSE. It was demonstrated in [13,14] that overdetermined BSE provides higher separation performance than determined BSE. Therefore, when there are a finite number of...
observed signals available, overdetermined BSE is effective for improving the separation performance.

In this paper, we propose an LS-AJD algorithm, denoted as fast diagonalization, implemented by minimizing the direct and indirect LS criteria (FDMDI), for alternately minimizing the indirect LS (ILS) criterion for the diagonal matrices and the constrained direct LS (CDLS) criterion for the mixing matrix. The mixing matrix is obtained from a constrained LS minimizer by a non-iterative method without the nested loop in our FDMDI algorithm. We also demonstrate that FDMDI BSE in overdetermined mixtures achieves higher separation performance than determined FDMDI BSE in a real acoustic environment.

This paper is organized as follows. Section 2 introduces the determined and overdetermined convolutive mixtures and BSE. In Sect. 3, we present our FDMDI algorithm for alternately minimizing two criteria for the diagonal matrices and for the mixing matrix. In Sects. 4 and 5, we compare the performances of our FDMDI algorithm with those of available algorithms in determined and overdetermined convolutive audio mixtures and verify the superiority of FDMDI. Section 6 is the conclusion.

2. FREQUENCY-DOMAIN CONVOLUTIVE BSE

2.1. Determined and Overdetermined Convolutive Mixtures

In the convolutive mixing model from \( N \) sources \( s_1(t), s_2(t), \ldots, s_N(t) \) to \( J \) microphones \( x_1(t), x_2(t), \ldots, x_J(t) \) at time \( t \), the observed signal at the \( i \)th microphone is obtained 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 quasistationary signals from the perspective that their variances are time-varying. However, they can be modeled as approximately stationary processes over a short period of time denoted as an epoch. That is, they are nonstationary between different epochs [9], the sources are independent of each other, \( J \geq N \geq 2 \), and the room impulse response (RIR) \( h_{ij}(t) \) is a stable and causal non-minimum-phase mixing-filter impulse response from the \( j \)th source to the \( i \)th microphone. By definition, the RIR is fixed over the observational interval. The asterisk \( * \) denotes time-domain convolution. The additive white Gaussian noise (AWGN) \( n_i(t) \) with zero mean and variance \( \sigma^2 \) is independent of the sources. If the number of discrete Fourier transform (DFT) points is significantly larger than the length of the RIR \( h_{ij}(t) \), the time-domain convolution is approximately converted to multiplication by the short-time Fourier transform (STFT):

\[
x(m, \omega_k) \approx H(\omega_k) s(m, \omega_k) + n(m, \omega_k),
\]

where \( K \) denotes DFT points, \( \omega_k = 2\pi k/K \) for \( k = 0, 1, \ldots, K-1 \), \( s(m, \omega_k) \) and \( n(m, \omega_k) \) are the STFTs of \( s_j(t) \) and \( n_i(t) \) at epoch \( m \), respectively, \( h_{ij}(t) \) is the DFT of \( h_{ij}(t) \), and \( s(m, \omega_k) = [s_1(m, \omega_k), s_2(m, \omega_k), \ldots, s_N(m, \omega_k)]^T \) is an \( N \times 1 \) vector of sources. The \( J \times N \) mixing matrix of the transfer function from the \( N \) sources to \( J \) microphones, \( H(\omega_k) \), is assumed to have full column rank for all \( \omega_k \). The mixing matrix \( H(\omega_k) \) is determined only when \( J = N \geq 2 \). Meanwhile, the mixing matrix with \( J > N \geq 2 \) is overdetermined. Moreover, the \( J \times 1 \) observed signal vector is defined by \( x(m, \omega_k) = [x_1(m, \omega_k), x_2(m, \omega_k), \ldots, x_J(m, \omega_k)]^T \), and \( n(m, \omega_k) = [n_1(m, \omega_k), n_2(m, \omega_k), \ldots, n_J(m, \omega_k)]^T \) is the \( J \times 1 \) vector of the AWGN. All observed signals occur in the epoch interval \( 1 \leq m \leq M \), where \( M \) is the total number of epochs.

2.2. BSE Based on AJD

In convolutive frequency-domain BSE, the source signals \( s(m, \omega_k) \) are recovered from the observed signals \( x(m, \omega_k) \) by premultiplying \( x(m, \omega_k) \) by the \( N \times J \) unmixing matrix \( W(\omega_k) \):

\[
\hat{s}(m, \omega_k) = W(\omega_k) x(m, \omega_k),
\]

where

\[
W(\omega_k) H(\omega_k) = I(\omega_k) P(\omega_k).
\]

Moreover, \( P(\omega_k) \in \mathbb{R}^{N \times N} \) is a frequency-dependent permutation matrix, and \( I(\omega_k) \in C^{N \times N} \) is an arbitrary scale or phase diagonal matrix. The objective of the BSE techniques is to determine \( W(\omega_k) \), \( P(\omega_k) \), and \( I(\omega_k) \) without the prior knowledge of \( H(\omega_k) \) and \( s(m, \omega_k) \).

Defining

\[
\hat{T}(m, \omega_k) = E[x(m, \omega_k)x(m, \omega_k)^H]
\]

as the \( J \times J \) cross-spectral density matrix of the observed signal at point \( (m, \omega_k) \) [15], \( H(\omega_k) \) and \( \{P_s(m, \omega_k)\}_{m=1}^M \) are determined by diagonalizing a set of \( T(m, \omega_k) \) in an AJD problem. Here, [·] \( \cdot \) and the superscript \( \dagger \) denote the expectation operator and the Hermitian transpose, respectively; moreover, \( I \) denotes the \( J \times J \) identity matrix. Furthermore, the cross-spectral density matrix of the source signal, \( P_s(m, \omega_k) = E[s(m, \omega_k)s(m, \omega_k)^H] \in \mathbb{R}^{N \times N} \), is diagonal, and \( T(m, \omega_k) = H(\omega_k) P_s(m, \omega_k) H(\omega_k)^H \). Defining the matrix \( B(\omega_k) \) and the diagonal matrix \( A(m, \omega_k) \) as

\[
B(\omega_k) = H(\omega_k) P(\omega_k)^{-1} I(\omega_k)^{-1}
\]

\[
A(m, \omega_k) = I(\omega_k) P^{-1}(\omega_k) P_s(m, \omega_k) P(\omega_k)^T I(\omega_k)^{-1}
\]

with the scale constraint \( \|b(\omega_k)\|_2 = 1 \), Eq. (5) can be expressed as
In the LS-AJD problem \([5,10]\), after finding \(B\) frequency-dependent scaling and permutation problems.

3.2. Estimation of Diagonal Matrices

In the LS-AJD problem \([5,10]\), after finding \(B(\omega_k)\) by minimizing the direct LS (DLS) criterion

\[
C_{\text{DLS}}(B(\omega_k), \{A(m, \omega_k)\}_{m=1}^{M}) = \sum_{m=1}^{M} \| \hat{T}(m, \omega_k) - B(\omega_k)A(m, \omega_k)B(\omega_k)^{\dagger} \|^2_F,
\]

subject to the scale constraint \(\|b_i(\omega_k)\|_2 = 1\), \(W(\omega_k)\) is obtained by satisfying Eq. (9), where \(\| \cdot \|_F\) denotes the Frobenius norm. An unmixing-filter impulse response is obtained by converting \(P(\omega_k)^{-1}B(\omega_k)^{-1}W(\omega_k)\) into the time domain by the inverse DFT (IDFT) after solving the frequency-dependent scaling and permutation problems.

3. FDMDI APPROACH TO BSE

In this section, we consider an FDMDI updating rule for BSE. The FDMDI algorithm alternates between two phases.

1. In the first phase, the ILS criterion is minimized with respect to \(\{A(m, \omega_k)\}_{m=1}^{M}\) while fixing the previously estimated \(B(\omega_k)\).

2. In the second phase, the CDLS criterion is minimized with respect to \(B(\omega_k)\) while fixing the previous estimate of \(\{A(m, \omega_k)\}_{m=1}^{M}\).

Assuming that the estimated mixing matrix \(B(\omega_k)^{\dagger}B(\omega_k)\) is nonsingular, the unmixing filter can be selected to force a combined impulse response of the mixing filter with the unmixing filter in tandem to approximate a unit impulse \([12,16]\). The unmixing filter is suitable for prevailing over both interfering source signals and the degradation in quality of the extracted signals in BSE.

3.1. Estimation of Cross-Spectral Density Matrices of Observed Signal

Using the Welch periodogram method \([17]\), the \(M\) estimated cross-spectral density matrices \(\{\hat{T}(m, \omega_k)\}_{m=1}^{M}\) are obtained by dividing the observed signals into \(M\) epochs. The \(M\) estimated target matrices are normalized as

\[
\hat{T}(m, \omega_k) = \frac{\hat{T}(m, \omega_k)}{\max_{\omega_k} \| \hat{T}(m, \omega_k) \|_F}. \tag{11}
\]

3.2. Estimation of Diagonal Matrices

In the first phase, \(M\) diagonal matrices \(\{A(m)\}_{m=1}^{M}\) are estimated by solving the following ILS problem:

\[
C_{\text{ILS}}(\{A(m)\}_{m=1}^{M}) = \sum_{m=1}^{M} \| W\hat{T}(m)W^{\dagger} - A(m) \|_F^2, \tag{12}
\]

where we have omitted the frequency bin index \(\omega_k\) because all the quantities refer to the same frequency bin index. Minimizing \(C_{\text{ILS}}(\{A(m)\}_{m=1}^{M})\) with respect to \(A(m)\) yields

\[
\hat{A}(m) = \text{diag}[WT(m)W^{\dagger}], \tag{13}
\]

where \(\text{diag}[A]\) is the diagonal matrix formed by extracting the diagonal elements of matrix \(A\).

3.3. Estimation of Mixing Matrix

In the second phase, the mixing matrix \(B\) is determined by minimizing the CDLS criterion \([12]\)

\[
C_{\text{CDLS}}(B) = \sum_{m=1}^{M} \| \hat{T}(m) - BA(m)B^{\dagger} \|_F^2 + 2 \sum_{j=1}^{N} \gamma_j(b_j^{\dagger}b_j - 1) \tag{14}
\]

with respect to \(B\) over \(M\) epochs in each frequency bin; here, \(\gamma_j\) is the Lagrange multiplier associated with the scale constraint. Applying the Khatri–Rao (KR) product to Eq. (14), the CDLS criterion is obtained as

\[
C_{\text{CDLS}}(G) = \sum_{m=1}^{M} \| \hat{\theta} - Gd(m) \|_2^2 + 2 \sum_{j=1}^{N} \gamma_j(\text{vec}(I)^T\text{Re}(g_j) - 1), \tag{15}
\]

where

\[
\hat{\theta}(m) = \text{vec}(\hat{T}(m)) \tag{16}
\]

\[
G = [b_1 \otimes b_1, \cdots, b_N \otimes b_N] \tag{17}
\]

\[
d(m) = [\lambda_1(m), \lambda_2(m), \cdots, \lambda_N(m)]^T \tag{18}
\]

\[
Gd(m) = \text{vec}(BA(m)B^{\dagger}) = [B^* \odot B] \cdot d(m). \tag{19}
\]

Here, the superscript ", \(\odot\), and \(\otimes\) denote the complex conjugate, Kronecker product, and KR product, respectively, \(\lambda_i(m)\) denotes the \(i\)th diagonal element of \(A(m)\), and \(g_j\) denotes the \(i\)th column of \(G\). \(\text{Re}(a)\) refers to the real part of \(a\), \text{vec}(\(A\)) denotes the vectorization operator, which forms a \(J^2 \times 1\) column vector by stacking the columns of the \(J \times J\) matrix \(A\).

Defining the matrices \(F\) and \(R\) as

\[
F = \sum_{m=1}^{M} d(m)d(m)^T \tag{20}
\]

\[
R = \sum_{m=1}^{M} \hat{\theta}(m)d(m)^T, \tag{21}
\]

we obtain \(\hat{G}\) by minimizing \(C_{\text{CDLS}}(G)\) using the method of Lagrange multipliers as follows:
\[ \hat{G} = \text{Re}(\hat{G}) + j\text{Im}(\hat{G}), \]  

(22)

where \( j = \sqrt{-1} \), \( \text{Im}(a) \) refers to the imaginary part of \( a \),

\[ \text{Re}(\hat{G}) = \text{Re}(R)F^{-1} - \frac{1}{j} \text{vec}(I) \cdot \{\text{vec}(I)^T \text{Re}(R)f_1, \ldots, \text{vec}(I)^T \text{Re}(R)f_N - 1\} \]  

(23)

\[ \text{Im}(\hat{G}) = \text{Im}(R)F^{-1}, \]  

(24)

and \( f_k \) is the \( k \)th column of \( F^{-1} \).

After minimizing Eq. (15) to obtain \( \hat{g}_i \), the CDLS estimation problem is solved by projecting \( \hat{g}_i \) onto the constraint set \( \Omega \subset \mathbb{C}^{d \times 1} \), defined as [9,18]

\[ \Omega = \{\text{vec}(\Phi) | \Phi = vv^H, v \in \mathbb{C}^{j \times 1}, \|v\|_2 = 1\}. \]  

(25)

\( b_i \) is determined by minimizing \( \|\text{vec}^{-1}(\hat{g}_i) - b_i b_i^H\|_F^2 \) using the power method [19] for \( i = 1, 2, \ldots, N \) once in each ALS iteration.

### 3.4. Construction of Full-Rank Unmixing Matrix

The estimated mixing matrix \( B \) must be nonsingular in the AID problem [20,21]. If the estimated mixing matrix does not have full rank, it is replaced by an appropriate full-rank matrix. The unmixing matrix is derived subsequently from the full-rank matrix. The singular value decomposition (SVD) of \( B \) is expressed as

\[ B = V_r \Sigma_r U_r^H, \]  

(26)

where \( V_r \), \( U_r \), and \( \Sigma_r \) denote the two unitary matrices and a diagonal matrix, respectively:

\[ V_r = [v_1, v_2, \ldots, v_r] \]  

(27)

\[ U_r = [u_1, u_2, \ldots, u_r] \]  

(28)

\[ \Sigma_r = \text{diagonal}(\sigma_1, \sigma_2, \ldots, \sigma_r) \]  

(29)

\[ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0. \]  

(30)

We form an orthonormal set \( \{v_i\}_{i=1}^N \) orthogonal to the orthonormal set \( \{v_i\}_{i=r+1}^N \). Similarly, we form an orthonormal set \( \{u_i\}_{i=1}^N \) orthogonal to the orthonormal set \( \{u_i\}_{i=r+1}^N \). When the rank of \( B^H B \) is \( r \leq N \), \( B \) can be replaced by the full-rank matrix

\[ V \Sigma U^H = \frac{\sqrt{N}}{\sqrt{N} + \delta N} [V_r^T \Sigma_r \begin{bmatrix} 0 & \Sigma_r \end{bmatrix} \delta I] U_r^H \]  

(31)

constructed by the Gram–Schmidt orthonormalization process. Selecting \( \delta > 0 \), \( \sqrt{N}/(\sqrt{N} + \delta N) \) on the right-hand side must satisfy the condition \( \text{tr}[\Sigma] = \sqrt{N} \).

\[ \Sigma_r = \begin{bmatrix} v_{r+1}^T, v_{r+2}^T, \ldots, v_N^T \end{bmatrix} \]  

(32)

\[ U_r = [u_{r+1}, u_{r+2}, \ldots, u_N], \]  

(33)

where \( \text{tr}[A] \) denotes the trace of \( A \). The bar \( \overline{V} \) and the subscript \( r \) denote that \( \{v_i\}_{i=r+1}^N \) is the orthogonal complement of \( \{v_i\}_{i=1}^r \). Replacing \( B \) by \( V \Sigma U^H \) minimizes Eq. (15) because \( \{v_i\}_{i=1}^r \) and \( \{u_i\}_{i=r+1}^N \) are not completely affected by \( \{v_i\}_{i=r+1}^N \) and \( \{u_i\}_{i=r+1}^N \), respectively.

From Eq. (9), the unmixing matrix \( W \) can be estimated by minimizing the squares of the prediction error as

\[ \xi(W) = \|I - BW\|_F^2. \]  

(34)

Minimizing \( \xi(W) \) is equivalent to determining the combined response of the mixing filter with the unmixing filter in tandem [16]. The unmixing matrix is estimated by

\[ W = (B^H B)^{-1} B^H. \]  

(35)

After obtaining \( Gd(m) \) from \( B \) and \( A(m) \) using Eq. (19), this procedure is repeated until Eq. (15) changes by less than a specified convergence criterion \( \varepsilon_C \) between the ALS iterations.

Algorithm 1 summarizes the FDMDI algorithm at each frequency bin.

### Algorithm 1: Summary of FDMDI algorithm.

#### Estimation of the cross-spectral density matrices of the observed signal:

1: Compute \( \hat{T}(m) \) at each epoch using the Welch periodogram method for \( m = 1, 2, \ldots, M \).

#### Normalization of \( \hat{T}(m) \):

2: Normalize \( \hat{T}(m) \) for \( m = 1, 2, \ldots, M \) according to (11).

3: Convert \( \hat{T}(m) \) into \( \bar{t}(m) \) according to (16).

#### Initialization:

4: Set the initial values of \( W(0) \) to the indirect LS estimate at frequency bin \( 0 \). With \( W(0) = I \).

5: while The absolute value of the difference between the current and previous values of \( C_{\text{CDLS}}(\hat{G}) \) in the ALS iterations is larger than \( \varepsilon_C \). do

First phase (Estimation of the diagonal matrices):

6: Minimize \( C_{\text{CDLS}}(\hat{A}(m)) \) to determine \( \hat{A}(m) \) by the method of least squares according to (13).

Second phase (Estimation of the mixing matrix):

7: Compute \( F \) and \( R \) according to (20) and (21), respectively.

8: Compute the inverse of \( F \).

9: Minimize \( C_{\text{CDLS}}(\hat{G}) \) to determine \( \hat{G} \) by the method of Lagrange multipliers according to (22).

10: Minimize \( \|\text{vec}^{-1}(\hat{g}_i) - b_i b_i^H\|_F^2 \) to determine \( b_i \) by the power method for \( i = 1, 2, \ldots, N \).

#### Construction of the full-rank unmixing matrix:

11: if rank \( (B^H B) = r \leq N \) then

12: Construct the orthonormal sets \( v_{r+1}, v_{r+2}, \ldots, v_N \) and \( u_{r+1}, u_{r+2}, \ldots, u_N \) by the Gram–Schmidt process.

13: Replace \( B \) with \( V \Sigma U^H \) as expressed in (31).

14: end if

15: Compute \( W \) from \( B \) according to (35).

16: end while
3.5. Reconstruction of Source Signals
The BSE with FDMDI is accomplished in the following five steps.

a) The unmixing matrix $W$ is determined independently at each frequency bin $\omega_k$ in accordance with Algorithm 1.

b) The frequency-dependent scaling ambiguity is corrected by determining $P^{-1}W$.

c) The permutations over all the frequency bins are aligned by determining $P^{-1}P^{-1}W$.

d) $P^{-1}P^{-1}W$ is multiplied by $e^{-jk\omega}$.

e) An unmixing-filter impulse response is obtained by converting $e^{-jk\omega}P^{-1}P^{-1}W$ into the time domain by IDFT.

The unmixing filter is both a separated filter and an inverse filter. Whenever the mixing-filter impulse response is stable, causal, and non-minimum-phase, the delayed unmixing-filter impulse response prevents noise enhancement [15]. Therefore, the delayed unmixing matrix should be incorporated in BSE. We can use $w_i(t)$ to reconstruct the source signal at the $i$th unmixing-filter output as

$$
\hat{s}_i(t) = \sum_{j=1}^{J} w_{ij}(t) * x_j(t),
$$

where $w_{ij}(t)$ is the IDFT of the delayed unmixing matrix $e^{-jk\omega}P^{-1}P^{-1}W$.

3.6. Complexity Comparison
Whereas the ALSP algorithm [9] is constructed from a loop for estimating each column of the mixing matrix inside another loop for ALS iteration, we apply Gaussian elimination with partial pivoting [19] to determine the inverse of $F$ for directly estimating the mixing matrix in the FDMDI algorithm. Table 1 lists the number of multiplications and divisions per ALS iteration in FDMDI and PARAFAC with enhanced line search (ELS) [22]; here, $p_{r1}$ and $p_{r2}$ denote the number of iterations required for the power method to determine $b_i$ and to construct the full-rank unmixing matrix, respectively. A count of the number of multiplications per ALS iteration required for PARAFAC is illustrated in [22]. The computational requirement of FDMDI is almost identical to that of PARAFAC.

### Table 1 Computational complexity per ALS iteration.

| Method                      | First phase                                                                 | Second phase                                                                 |
|-----------------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------------|
| FDMDI                       | $(J^3 + J^2)M$                                                             | $(N^2 + 2J^2M)M + 5J^3 + 3J^2 + N^3 / 2 - N^2 / 2 + N$                       |
|                             |                                                                            | $+ 3(JN)^2 + (2J^2 + 4J)Np_{r1}$                                            |
|                             |                                                                            | $+ 2(2J^2 + 4J)(N - r)p_{r2}$                                               |
|                             |                                                                            | $(NP_{r1} + 2(N - r)p_{r2} + 1$ divisions)                                   |
| PARAFAC with ELS [22]       | $(NJ(1 + 7N)(J + 2M) + 3J^2M^2 + 11N^3$                                    | $(NJ(1 + 7N)(J + 2M) + 3J^2M^2 + 11N^3$                                    |
|                             | $+ 2N^2(2J + M) + (8N + 10J^2M) / 3$                                       | $+ 2N^2(2J + M) + (8N + 10J^2M) / 3$                                       |

4. SIMULATION RESULTS
In this section, we compare the computational burdens and separation performances of the FDMDI algorithm with those of typical available algorithms: ALSP [9], PARAFAC with ELS [22], ALUJA [23], and GNJD [24]. The performance measure is the input signal-to-interference ratio (SIR) at the $i$th microphone and the output SIR at the $i$th unmixing-filter output defined as

$$
\text{SIR}_{w_i} = 10 \log_{10} \left[ \frac{\sum_i [(h_{ii}(t) * s_i(t))^2]}{\sum_i \left[ \frac{\sum_{j=1}^{N} h_{ij}(t) * s_j(t) + n_i(t)}{2} \right]^2} \right] \quad (37)
$$

$$
\text{SIR}_{y_{ji}} = 10 \log_{10} \left[ \frac{\sum_i [(\gamma_{ij}(t) * s_i(t))^2]}{\sum_i \left[ \frac{\sum_{j=1}^{N} \gamma_{ij}(t) * s_j(t) + \sum_{j=1}^{J} w_{ij}(t) * n_i(t)}{2} \right]^2} \right] \quad (38)
$$

respectively; here, $\gamma_{ij}$ denotes the impulse response of the composite mixing-unmixing filter from the $i$th source to the $j$th unmixing-filter output. In all the simulations, the signal-to-noise ratio (SNR) was determined as the power ratio of the desired source signal to the noise component in the output signals after obtaining an optimum unmixing matrix.
We placed microphones in a modeled image room of size $4.45 \times 3.55 \times 2.50 \text{m}$ by using the image method [27,28] at a sampling rate of 8 kHz. The reverberation times (RT) were 100, 300, 500, 700, and 900 ms. For $N = 3$, the three sources were located at the positions $\alpha$, $\beta$, and $\gamma$ in Fig. 1; here, the center of gravity of the triangle was located at the center of a sphere of radius 1.20 m centered at (2.23, 1.78, 1.20 m). Table 2 lists the source and microphone coordinates corresponding to each label. To prevent the directivity of a linear microphone array from enhancing or attenuating the observed signals, seven microphones were placed at positions A, B, C, D, E, F, and G in Fig. 1. For $J = 3$, the observed signals were obtained from the three microphones placed at A, C, and D in Fig. 1. The five microphones at A, B, C, D, and E in Fig. 1 were used for $J = 5$. For $N = 4$, the four sources were located at positions identical to those in [16]. For $J = 4$, the four microphones were also placed at positions similar to those in [16]. For $J = 5$, we added a microphone located at (2.23, 1.78, 1.32 m) to the microphone configuration for $J = 4$. For $J = 6$, we added a microphone located at (2.23, 1.78, 1.08 m) to the microphone configuration for $J = 5$. The observed signals were obtained by convolving the audio speech data with the RIRs. The observed signals were contaminated with zero-mean AWGN of 20 dB. The duration of the speech datasets was 1,500 s. All the speech data were normalized to approximately identical signal strengths. The STFT was performed with a Hanning window. The parameters were empirically selected as an 8,192-point fast Fourier transform (FFT) with a frame overlap of 80% [16].

In PARAFAC, $B(\omega_0)$ and $\{A(m, \omega_0)\}_{m=1}^M$ were initialized as white noise [22]. In the other ALS algorithms, they were initialized to identity matrices. The initial values of $B(\omega_0)$ and $\{A(m, \omega_0)\}_{m=1}^M$ for all the algorithms, except for ALUJA and GNJD, were set to the LS-AJD estimate in frequency bin $\omega_0$. This initialization procedure was repeated until the LS-AJD estimate in frequency bin $\omega_{K/2}$ was obtained [9]. In ALUJA and GNJD, the initial values of $B(\omega_0)$ and $\{A(m, \omega_0)\}_{m=1}^M$ were set to the identity matrices in each frequency bin for the highest performance. The number of loading matrices $R$ in GNJD was set to 4 [24]. In the FDMDI approach, the convergence criterion was empirically set to $\epsilon_C = 10^{-6}$. To construct the full-rank unmixing matrix, we set $\delta = \sigma_r$. The maximum number of iterations was set to 2,000. The simulation program was coded in the C programming language and executed on an Intel Core i7-2600 3.4 GHz processor in

![Fig. 1 Source–microphone configuration for simulated RIR for $N = 3$.](image-url)
double-precision floating-point format. In all the BSE approaches, the unmixing matrix \( W \) was multiplied by \( e^{-j\pi k} \). In addition to the application of ALUJA and GNJD, the scaling problem was solved by normalizing each row vector of \( W \) in each frequency bin [16]. All the BSE approaches resolved the ambiguity in the frequency-dependent permutation using the correlation coefficient of the power ratios between neighboring frequency bins [29,30].

4.2. Determined Case

To evaluate the computational complexity of the ALS algorithms, we measured the CPU time in both phases of each ALS algorithm. The parameters were set to \( N = J = 3 \) in a 300-ms reverberant environment. Figure 2 shows the CPU time of the first phase in the FDMDI algorithm. FDMDI directly estimates the mixing matrix by computing the inverse of \( F \) through Gaussian elimination with partial pivoting; this involves only the \( J^2 \)-dimensional column vector in Step 8 of Algorithm 1. Consequently, the computational requirement of the FDMDI algorithm does not depend on \( M \), as shown in Fig. 2. However, the computational requirement of Step 7 of FDMDI, which computes \( F \) and \( R \), increases with \( M \). The computational requirement of ELS in the PARAFAC algorithm is included in the second phase. Because noniterative methods are used for estimating the mixing matrix in the second phase, the overall computational requirement for the PARAFAC algorithm is the largest among the tested ALS algorithms. The FDMDI algorithm incurred the lowest computational cost among the tested approaches.

As shown in Tables 3 and 4, because the FDMDI algorithm without the nested loop required less complex computational resources than the ALSP algorithm with the nested loop, the FDMDI algorithm reduces the total number of computations. The reduction with the application of the FDMDI algorithm is attributed to the noniterative method.

Figure 2 also shows the average output SIRs of the BSEs versus the number of epochs. The FDMDI BSE provided the highest separation performance at each epoch. Tables 3 and 4 also present the output SIRs of the tested BSE approaches for \( N = 3 \) and 4, respectively, in the determined mixture. For \( J = N = 3 \) and \( M = 400 \) in the 300-ms reverberant environment, the average improvement in the PESQ score of FDMDI was 0.94 and 0.82 points higher than those of ALUJA and GNJD, respectively.

4.3. Overdetermined Case

Figure 3 shows the output SIRs of the ALS BSEs for \( N = 3 \) in determined and overdetermined mixtures.
reverberant environments shorter than 300 ms, increasing the number of microphones beyond five did not enhance the separation performance of the FDMDI BSE with almost no increase in the number of ALS iterations, but at an additional computational cost, because the LS-AJD estimation errors are predominant and the ambient noise is negligible. As evident in Figs. 2 and 3, the performance of

the overdetermined FDMDI BSE for \( J = 7 \) and a small number of epochs (e.g., \( M = 400 \)), corresponding to the speech data sets of 500 s, was marginally lower than that of the determined FDMDI BSE for a large number of epochs (e.g., \( M = 800 \)), corresponding to speech data sets of 1,000 s. Moreover, for \( M = 800 \) in the 300-ms reverberant environment, the computational requirement was approximately 60% lower in the determined FDMDI BSE than in the overdetermined FDMDI BSE.

Figure 4 shows the performances of the BSE approaches; here, ALUJA and GNJD cannot be directly applied in the case of overdetermined mixtures without a dimension reduction from \( J \) to \( N \) [31]. The FDMDI BSE achieved the highest separation performance among the tested BSE approaches. For an identical number of epochs in long reverberant environments, the overdetermined FDMDI algorithm performed better than the determined FDMDI algorithm. As revealed in Fig. 4, the FDMDI BSE achieved the highest PESQ score among the BSEs in the overdetermined mixtures.

5. SEPARATION PERFORMANCE IN AN ACTUAL ACOUSTIC ENVIRONMENT

Assuming that RIR can be modeled as a linear system, the image method has been widely used for efficiently simulating small-room acoustics from ideal point sources to ideal point microphones. Up to this section, we have been describing the execution of the BSE approaches the artificial RIR for rooms generated by the image method. In an actual acoustic environment, the observed signals are contaminated by nonlinearities such as saturation and distortion. Nonlinear distortion is introduced by loudspeakers and power amplifiers [32–34]. Moreover, real loudspeakers are not ideal omnidirectional point sources, and real microphones are not ideal point microphones. To further visualize the disastrous effects of nonlinearities,
than four loudspeakers. Moreover, when we compare the performances of the FDMDI, GNJD, ALUJA, and PARAFAC BSEs. The other simulation settings were identical to those in the previous experiment. Table 6 illustrates the performances of these algorithms. We evaluated the computational burden of the FDMDI algorithm. We also compared the performance of our algorithm with those of the typical available algorithms in reverberant environments. The simulation results demonstrated that our algorithm outperforms all these algorithms. Finally, we illustrated that the overdetermined FDMDI algorithm exhibits an enhanced separation performance and higher quality of extracted signals with a marginal increase in computational requirements.

Table 6 Average CPU times and iterations per frequency bin and average PESQ scores of FDMDI, GNJD, ALUJA, and PARAFAC BSE approaches for \( N = 5, \epsilon_c = 10^{-12}, M = 1.200, \) and 16,384-point FFT in an actual acoustic environment. The five microphones at positions A, B, C, D, and E in Fig. 5 were used for all the determined BSEs. For \( J = 7, \) the two microphones located at F and G in Fig. 5 were added to the five microphones at A, B, C, D, and E.

We have derived an ALS algorithm for blind extraction of convolutive speech mixtures. The FDMDI approach has the advantage of rapid convergence in comparison to typical available algorithms. We evaluated the computational burden of the FDMDI algorithm. We also compared the performance of our algorithm with those of the typical available algorithms in reverberant environments. The simulation results demonstrated that our algorithm outperforms all these algorithms. Finally, we illustrated that the overdetermined FDMDI algorithm exhibits an enhanced separation performance and higher quality of extracted signals with a marginal increase in computational requirements.

6. CONCLUSION

We have derived an ALS algorithm for blind extraction of convolutive speech mixtures. The FDMDI approach has the advantage of rapid convergence in comparison to typical available algorithms. We evaluated the computational burden of the FDMDI algorithm. We also compared the performance of our algorithm with those of the typical available algorithms in reverberant environments. The simulation results demonstrated that our algorithm outperforms all these algorithms. Finally, we illustrated that the overdetermined FDMDI algorithm exhibits an enhanced separation performance and higher quality of extracted signals with a marginal increase in computational requirements.

Table 5 Locations of loudspeakers and microphones in Fig. 5.

| Method | Label in Fig. 5 | Coordinate [m] | Iterations | CPU time [s] | PESQ score |
|--------|----------------|----------------|------------|-------------|------------|
| FDMDI  | \( s_1(t) \)   | (2.62, 0.64, 1.20) | 0.382 | 29.19 | 1.76 |
| GNJD   | \( s_2(t) \)   | (3.42, 1.80, 1.20) | 5 | 1082.555 | 36.48 | 1.28 |
| ALUJA  | \( s_3(t) \)   | (2.58, 2.92, 1.20) | 7.841 | 107.65 | 1.37 |
| PARAFAC| \( s_4(t) \)   | (1.24, 2.46, 1.20) | 240.918 | 1442.147 | 0.50 |
| FDMDI  | \( s_5(t) \)   | (1.27, 1.05, 1.20) | 29.19 | 1.76 |
| PARAFAC| \( x_1(t) \)   | (2.23, 1.66, 1.20) | 552.538 | 1343.348 | 0.50 |
| PARAFAC| \( x_2(t) \)   | (2.34, 1.74, 1.20) | 447.492 | 20.48 |
| PARAFAC| \( x_3(t) \)   | (2.30, 1.87, 1.20) | 1.036 | 17.01 | 1.90 |
| PARAFAC| \( x_4(t) \)   | (2.15, 1.87, 1.20) | 1.95 | 17.01 |
| PARAFAC| \( x_5(t) \)   | (2.23, 1.86, 1.28) | 20.48 |
| PARAFAC| \( x_6(t) \)   | (2.14, 1.79, 1.12) | 1343.348 | 0.50 |

Fig. 5 Loudspeaker–microphone configuration of acoustic environment experiment in an actual room.

Table 5 Locations of loudspeakers and microphones in Fig. 5.

| Room size: 4.71 × 5.83 × 2.50 m |
|---------------------------------|
| Label in Fig. 5 | Coordinate [m] |
| \( s_1(t) \)   | (2.62, 0.64, 1.20) |
| \( s_2(t) \)   | (3.42, 1.80, 1.20) |
| \( s_3(t) \)   | (2.58, 2.92, 1.20) |
| \( s_4(t) \)   | (1.24, 2.46, 1.20) |
| \( s_5(t) \)   | (1.27, 1.05, 1.20) |
| \( x_1(t) \)   | (2.23, 1.66, 1.20) |
| \( x_2(t) \)   | (2.34, 1.74, 1.20) |
| \( x_3(t) \)   | (2.30, 1.87, 1.20) |
| \( x_4(t) \)   | (2.15, 1.87, 1.20) |
| \( x_5(t) \)   | (2.23, 1.86, 1.28) |
| \( x_6(t) \)   | (2.14, 1.79, 1.12) |
| \( x_7(t) \)   | (2.31, 1.76, 1.12) |

real loudspeakers, and real microphones, in this section, we provide the separation performances of the BSE approaches applied in determined and overdetermined mixtures using real-room recordings at the microphones without measuring RIR.

Five directional loudspeakers and nine omnidirectional microphones were arranged in an empty room as displayed in Fig. 5. The loudspeaker–microphone configuration was highly similar to the source–microphone configuration in Fig. 1, except for the number of loudspeakers and microphones and the room size and form. Table 5 presents the coordinates of the loudspeakers and microphones corresponding to each label. The average reverberation time in the actual room was approximately 350 ms [16]. Measurements were recorded for 1,500 s, corresponding to \( M = 1,200, \) over an 8 kHz bandwidth at a 16 kHz sampling rate. The other simulation settings were identical to those in the previous experiment. Table 6 illustrates the performances of the FDMDI, GNJD, ALUJA, and PARAFAC BSEs. The performances of all the BSEs deteriorated with more than four loudspeakers. Moreover, when we compare the performances of all the BSEs in the actual acoustic environment with those in the simulated environment, the performances degraded in the actual room. Whereas the determined FDMDI exhibited the best CPU time, the overdetermined FDMDI with \( J = 9 \) gave the best PESQ score among the tested BSEs. The second highest PESQ score was achieved by the FDMDI with \( J = 7 \). There is a monotonic improvement in the PESQ for FDMDI as the number of microphones becomes higher than the number of sources. Figure 6 shows the spectrograms of the clean speech signals and the unmixing-filter output signals separated by the FDMDI, ALUJA, and PARAFAC BSEs. A comparison of the spectrograms revealed that the overdetermined FDMDI with \( J = 9 \) recovers the harmonic structure of clean speech signals more effectively than ALUJA and PARAFAC do; here, the average delay between the source and observed signals was approximately 0.004 s and the average delay between the source and separated signals was approximately 0.504 s for all the BSEs.
ments when the number of available observed signals is limited.

ACKNOWLEDGEMENTS

The authors wish to thank the anonymous reviewers for their helpful comments and recommendations.

REFERENCES

[1] L. Shoker, S. Sanei and J. Chambers, “Artifact removal from electroencephalograms using a hybrid BSS-SVM algorithm,” IEEE Signal Process. Lett., 12, 721–724 (2005).

[2] W. D. Clercq, A. Vergult, B. Vanrumste, W. V. Paesschen and S. V. Huffel, “Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram,” IEEE Trans. Biomed. Eng., 53, 2583–2587 (2006).

[3] T. Tsalaile, R. Sameni, S. Sanei, C. Jutten and J. Chambers, “Sequential blind source extraction for quasi-periodic signals with time-varying period,” IEEE Trans. Biomed. Eng., 56, 646–655 (2009).

[4] M. Congedo, C. Gouy-Paillet and C. Jutten, “On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics,” Clin. Neurophysiol., 119, 2677–2686 (2008).

[5] M. Wax and J. Sheinvald, “A least-squares approach to joint diagonalization,” IEEE Signal Process. Lett., 4, 52–53 (1997).

[6] L. Parra and C. Spence, “Convolutonal blind separation of non-stationary sources,” IEEE Trans. Speech Audio Process., 8, 320–327 (2000).

[7] A. Yeredor, “Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation,” IEEE Trans. Signal Process., 50, 1545–1553 (2002). [Online] Available: http://www.eng.tau.ac.il/~arie/

[8] M. Z. Ikram and D. R. Morgan, “Permutation inconsistency in blind speech separation: Investigation and solutions,” IEEE Trans. Speech Audio Process., 13, 1–13 (2005).

[9] K. Rahbar and J. P. Reilly, “A frequency domain method for blind source separation of convolutive audio mixtures,” IEEE Trans. Speech Audio Process., 13, 832–844 (2005).

[10] S. Degerine and E. Kame, “A comparative study of approximate joint diagonalization algorithms for blind source separation in presence of additive noise,” IEEE Trans. Signal Process., 55, 3022–3031 (2007).

[11] G. Chabriel, M. Kleinsteuber, E. Moreau, H. Shen, P. Tichavsky and A. Yeredor, “Joint matrices decompositions and blind source separation: A survey of methods, identification, and applications,” IEEE Signal Process. Mag., 31, 34–43 (2014).

[12] T. Asamizu, S. Saito, K. Oishi and T. Furukawa, “Over-determined blind source separation using approximate joint diagonalization,” Proc. 60th IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS-2017), Boston, USA, pp. 168–171 (2017).

[13] A. Westner and V. M. Bove, “Blind separation of real world audio signals using overdetermined mixtures,” Proc. 1st Int. Conf. Ind. Compon. Anal. Blind Signal Separat., pp. 11–15 (1999).

[14] C. Osterwise and S. L. Grant, “On over-determined frequency domain BSS,” IEEE Trans. Audio Speech Lang. Process., 22, 956–966 (2014).

[15] B. Widrow and S. D. Stearns, Adaptive Signal Processing (Prentice-Hall, Upper Saddle River, 1985).

[16] S. Saito, K. Oishi and T. Furukawa, “Convolutonal blind source separation using an iterative least-squares algorithm for non-

fig. 6 Spectrograms of clean speech signals and unmixing-filter output signals separated by overdetermined FDMDI, PARAFAC, and determined ALUJA BSEs for 16,384-point FFT, M = 1,200, and e_c = 10^{-12} in an actual acoustic environment. The microphones at A, B, C, D, and E in Fig. 5 were used for ALUJA.
orthogonal approximate joint diagonalization.” *IEEE Trans. Audio Speech Lang. Process.*, **23**, 2434–2448 (2015).

[17] P. D. Welch, “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Trans. Audio Electroacoust.*, **15**, 70–73 (1967).

[18] K. Rahbar, J. P. Reilly and J. H. Manton, “Blind identification of MIMO FIR systems driven by quasistationary sources using second order statistics,” *IEEE Trans. Signal Process.*, **52**, 406–417 (2004).

[19] G. H. Golub and C. F. V. Loan, *Matrix Computations* (The Johns Hopkins University Press, Baltimore, 1991).

[20] B. Afsari, “Sensitivity analysis for the problem of matrix joint diagonalization,” *SIAM J. Matrix Anal. Appl.*, **30**, 1148–1171 (2008).

[21] G. Chabriel and J. Barrere, “A direct algorithm for non-orthogonal approximate joint diagonalization,” *IEEE Trans. Signal Process.*, **60**, 39–47 (2012).

[22] M. Rajih, P. Comon and R. A. Harshman, “Enhanced line search: A novel method to accelerate PARAFAC,” *SIAM J. Matrix Anal. Appl.*, **30**, 1128–1147 (2008).

[23] V. Maurandi and E. Moreau, “A decoupled Jacobi-like algorithm for non-unitary joint diagonalization of complex-valued matrices,” *IEEE Signal Process. Lett.*, **21**, 1453–1456 (2014).

[24] X. F. Gong, X. L. Wang and Q. H. Lin, “Generalized non-orthogonal joint diagonalization with LU decomposition and successive rotations,” *IEEE Trans. Signal Process.*, **63**, 1322–1334 (2015).

[25] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio Speech Lang. Process.*, **16**, 229–238 (2008).

[26] “ASJ continuous speech corpus for research (ASJ-JIPDEC),” National Institute of Information. [Online] Available: http://research.nii.ac.jp/src/en/ASJ-JIPDEC.html

[27] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Am.*, 943–950 (1979).

[28] “Room impulse response generator for MATLAB.” [Online] Available: http://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.html

[29] H. Sawada, S. Araki and S. Makino, “Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS,” *Proc. IEEE Int. Symp. Circuits Syst.*, 3247–3250 (2007).

[30] H. Sawada, S. Araki and S. Makino, “MLSP 2007 data analysis competition: Frequency-domain blind source separation for convolutive mixtures of speech/audio signals,” *Proc. 2007 IEEE Workshop Mach. Learn. Signal Process.*, Thessaloniki, Greece, pp. 45–50 (2007).

[31] T. Trainini and E. Moreau, “A coordinate descent algorithm for complex joint diagonalization under Hermitian and transpose congruences,” *IEEE Trans. Signal Process.*, **57**, 4974–4983 (2014).

[32] A. Stenger and W. Kellermann, “Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling,” *Signal Process.*, **80**, 1747–1760 (2000).

[33] A. Guerin, G. Faucon and R. Le Bouquin-Jeannes, “Nonlinear acoustic echo cancellation based on Volterra filters,” *IEEE Trans. Speech Audio Process.*, **11**, 672–683 (2003).

[34] S. Malik and G. Enzner, “State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation,” *IEEE Trans. Audio Speech Lang. Process.*, **20**, 2065–2079 (2012).

Shinya Saito He received his B.S. and M.S. in electronics from the Tokyo University of Technology, Tokyo, Japan, in 2005 and 2007, respectively. He is currently pursuing a Ph.D. in the Department of Management Science at the Tokyo University of Science, Tokyo, Japan. He is a member of IEEE.

Kunio Oishi He received his B.S. degree in electrical engineering from the Chiba Institute of Technology, Chiba, Japan, in 1985, M.S. degree in electronic engineering from the Tokyo University of Agriculture and Technology, Tokyo, Japan in 1987, and Ph.D. in physical electronics Engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1990. From 1990, he has worked in electronics engineering at the Tokyo University of Technology, Tokyo, Japan. He is currently an Associate Professor with the School of Computer Science at the Tokyo University of Technology. He is a member of IEEE and IEICE.

Toshihiro Furukawa He received his B.S. degree and Ph.D. from the Tokyo Institute of Technology, Tokyo, Japan, in 1980 and 1990, respectively. He joined Canon Inc. in 1980, where he designed copy machines and optical disk systems. He was at the Chiba Institute of Technology from 1986 to 1992 as a research associate. From 1992 to 1999, he was an Assistant Professor at the Chiba Institute of Technology, and subsequently at the Fukuoka Institute of Technology. He is currently a Professor of the Department of Management Science at the Tokyo University of Science. His current research interests include adaptive signal processing, wireless communication methods, and machine learning theory. He is a member of IEEE and IEICE.