Hindawi Publishing Corporation
Advances in Acoustics and Vibration
Volume 2011, Article ID 765429, 10 pages
doi:10.1155/2011/765429

Research Article

Improved Method of Blind Speech Separation with Low Computational Complexity

Kazunobu Kondo,[1] Yu Takahashi,[1] Seiichi Hashimoto,[1] Hiroshi Saruwatari,[2] Takanori Nishino,[3] and Kazuya Takeda[4]

[1] *Corporate Research & Development Center, Yamaha Corporation, 203 Matsunokijima, Iwata 438-0192, Japan*
[2] *Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5, Takayama-cho, Ikoma 630-0192, Japan*
[3] *Graduate School of Engineering, Mie University, 1577, Kurimamachiya-cho, Tsu 514-8507, Japan*
[4] *Graduate School of Information Science, Nagoya University, Chikusa-ku Furou-cho, Nagoya 464-8601, Japan*

Correspondence should be addressed to Kazunobu Kondo, tashin@beat.yamaha.co.jp

Received 15 June 2011; Accepted 19 July 2011

Academic Editor: K. M. Liew

A blind speech separation method with low computational complexity is proposed. This method consists of a combination of independent component analysis with frequency band selection, and a frame-wise spectral soft mask method based on an interchannel power ratio of tentative separated signals in the frequency domain. The soft mask cancels the transfer function between sources and separated signals. A theoretical analysis of selection criteria and the soft mask is given. Performance and effectiveness are evaluated via source separation simulations and a computational estimate, and experimental results show the significantly improved performance of the proposed method. The segmental signal-to-noise ratio achieves 7 [dB] and 3 [dB], and the cepstral distortion achieves 1 [dB] and 2.5 [dB], in anechoic and reverberant conditions, respectively. Moreover, computational complexity is reduced by more than 80% compared with unmodified FDICA.

1. Introduction

In recent years, sophisticated mobile devices with microphones, such as smart phones and digital cameras, have become ubiquitous. The improvement of sound quality when using these devices in noisy environments is widely anticipated. The blind source separation (BSS) technique has received much attention in many industries for speech enhancement applications.

Independent component analysis (ICA) is one of the most researched fields of the BSS method, with many studies focused on fast convergence [1], engineering research [2], and so forth. In particular, frequency domain ICA (FDICA) [3] is expected to achieve significant sound quality improvement for speech enhancement applications, because practical environments are generally reverberant; however, it is computationally complex.

The time-frequency mask method is another current BSS research field, and one of the most common methods used to extract speech signals from underdetermined linear systems of equations is DUET [4]. DUET achieves significant separation performance using sparse signal decomposition and time-frequency clustering. There are many time-frequency mask studies [4–6], and these methods require storage of long-interval observed signals for time-frequency clustering. This means that the time-frequency mask method based on time-frequency clustering requires considerable computational resources.

In this paper, we propose a BSS method based on FDICA which requires fewer computational resources. This is important, because reductions in *battery* power consumption are needed to increase device operating time using current processors. Processors with higher performance and lower power consumption will be common in the near future; thus, it may not seem so important to reduce the computational resources which are required. However, sophisticated mobile devices with many extended and rich media functions, for example, the user interface, must work concurrently with

basic functions such as sound input. Consequently, lowering the power consumption of basic functions would still be valuable, thus the necessity and importance of using less computation to perform FDICA in order to improve sound quality.

In general, FDICA stores observed signals to estimate higher-order statistics; moreover, the time length of the stored signal must be relatively long, for example, over a few seconds. This leads to large memory consumption to perform FDICA. Meanwhile, higher-order statistics must be estimated from separated signals. In addition, the separated signals are different in every iteration, because the separation matrix is continuously updated. This means that the separation process for long-term observed signals must work in *every* iteration. Consequently, this leads to intrinsic computational complexity.

In this paper, we introduce a lower computational complexity BSS method based on FDICA which utilize band selection and a frame-wise spectral soft mask [7]. The sound source, for example, a speech signal, consists of a set of some predominant frequency bands such as formants; therefore, it is natural to assume that limiting frequency bands is important for improving both source separation performance and the efficiency of FDICA learning algorithms. A semiblind source separation (semi-BSS) method based on a similar concept has been proposed by one of the authors [8]. In the semi-BSS method, the band selection process contributes to the reduction of computational complexity. However, a theoretical reason of a band selection criteria, the determinant of a spatial covariance matrix, was not clear in the conventional method [8]. Therefore, we introduce the theoretical analysis of the band selection criteria. Meanwhile, the conventional method achieves practical separation performance; however, the amplitude of the null-beamformer (NBF) separated signal is attenuated at lower frequencies, because the phase difference also becomes small in the low-frequency region. This attenuation always appears in the microphone-array signal processing of NBF, and this is due to the degradation of the output signal.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the conventional method [8]. In Section 3, we introduce the theoretical analysis of the band selection criteria and propose a frame-wise spectral soft mask [7]. In Section 4, we show our experimental results and evaluate the proposed method. Section 5 presents our conclusions and describes future work.

2. Outline of Conventional Semi-BSS Method

In this section, we briefly explain the conventional semi-BSS method [8].

One of the authors has already proposed a semi-BSS method with low computational complexity [8]. Using this method, the target signal is assumed to arrive from a known direction, thus the term "semiblind". First, a separation matrix initialization method using the known direction of the target signal was proposed. This method results in a smaller number of iterations. Second, a band-selection method was proposed which contributes to smaller memory consumption, and an NBF-based separation matrix was used for making substitutions in the nonselected bands. NBF is one of the most common beamformer methods used to block sound sources. Since the FDICA separation matrix corresponds to the coefficients of the beamformer [9], the proposed semi-BSS method achieves significantly improved computational efficiency and practical performance.

Certain bands are selected according to the magnitude of the determinant of a spatial covariance matrix, and the selection occurs according to the largest magnitude. Following band selection, the FDICA separation matrix is obtained for the selected bands using the general FDICA algorithm. The source directions of arrival (DOA) are estimated from the FDICA separation matrix, and the permutation problem is solved using the estimated DOA [9]. The estimated DOA of all the selected bands are averaged, and for nonselected bands, the NBF coefficients, consisting of the averaged DOA, are used as the separation matrix for making substitutions. For the selected bands, the scaling ambiguity is solved using the projection method [10]. Following the completion of the separation matrices for all frequency bands, the separation process obtains the separated signals from the observed signals.

As mentioned in Section 1, the output signals of the NBF are attenuated in the low-frequency region; therefore, the separated signals might be degraded.

3. Proposed Method

3.1. Motivation and Strategy. In the conventional method [8], the NBF is used as the separation matrix for making substitutions in the nonselected bands. However, phase difference is very small in the low-frequency region; thus, the performance of the NBF becomes very poor. In particular, this leads to extreme degradation in mobile devices, because the distance between microphones is small. In contrast to the poor performance of the NBF, a frame-wise soft mask estimate can still be obtained. Therefore, for the nonselected bands, we applied the obtained soft mask instead of the NBF. In addition, because the theoretical analysis of the band selection method in [8] was insufficient, it is introduced in this paper. A block diagram of the proposed method is shown in Figure 1.

3.2. Signal Model. In this section, we formulate a signal model. In addition, we assume in this paper that there are two source signals and two microphones, because the assumed application systems are mobile devices.

The observed signals in the time domain are transformed into the frequency domain by short-time Fourier transform (STFT). The convolutive model describing signal propagation and mixing is formulated in the frequency domain as follows:

$$\mathbf{X}(k,l) = \mathbf{A}(k)\mathbf{S}(k,l), \tag{1}$$

where $\mathbf{X}(k,l) = [X_1(k,l),\ X_2(k,l)]^T$ is an observed signal vector at the microphones and $\mathbf{S}(k,l) = [S_1(k,l),\ S_2(k,l)]^T$

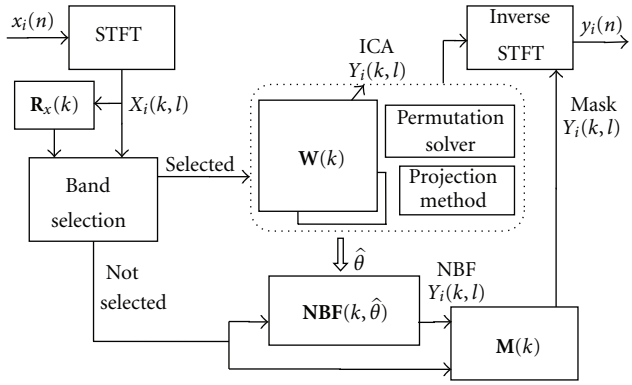


Figure 1: Block diagram of proposed method.

is a source signal vector. $\mathbf{A}(k)$ is a mixing matrix, k is a frequency bin index, l is a frame index, and $(\cdot)^T$ denotes the transpose operator.

In this case, the left half of Figure 2 shows the block diagram of propagation (indices k and l are omitted to simplify the diagram).

In Figure 2, the source signals are located at direction $\theta_i(k)$ (i corresponds to the source number) on each frequency band k, because in the reverberant condition, the source direction deviates on each frequency band. τ_{iS} and $A_i(k)$ are a delay and a gain corresponding to the distance between the source and the center of the microphone. In addition, $\tau_{ij}(k)$ is the delay of each microphone (j corresponds to the microphone number) on the kth frequency band.

Therefore, the mixing matrix $\mathbf{A}(k)$ is formulated as follows:

$$\mathbf{A}(k) = \begin{bmatrix} A_1(k)\mathrm{e}^{-j\omega(k)(\tau_{1S}+\tau_{11}(k))} & A_2(k)\mathrm{e}^{-j\omega(k)(\tau_{2S}+\tau_{21}(k))} \\ A_1(k)\mathrm{e}^{-j\omega(k)(\tau_{1S}+\tau_{12}(k))} & A_2(k)\mathrm{e}^{-j\omega(k)(\tau_{2S}+\tau_{22}(k))} \end{bmatrix}, \tag{2}$$

where $\omega(k)$ is an angular frequency that is equivalent to $2\pi(kF_s/N)$, F_s is the sampling frequency, and N is the size of the FFT.

3.3. Frequency Band Selection. In this section, we introduce a theoretical analysis of band selection criterion. Intuitively, if there is only one source in one of the frequency bands, the rank of the covariance matrix is not full, and the determinant becomes zero; thus, the determinant can describe the number of source signals. Therefore, in the conventional method [8], the determinant of the spatial covariance matrix is used for band selection criterion. On the other hand, the power of source signals is common criterion to describe the degree of spectral influence; thus, we also introduce theoretical analysis of the trace of the spatial covariance matrix and show experimental comparison between the determinant and the trace in the later section.

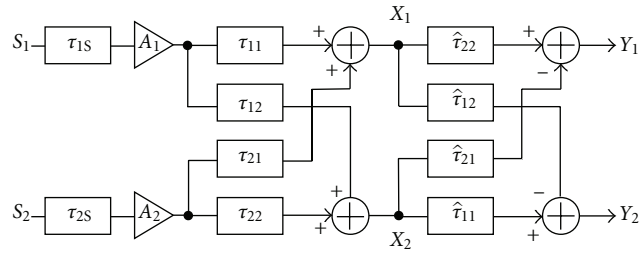


Figure 2: Block diagram of mixing and separation.

3.3.1. Spatial Covariance Matrix. The spatial covariance matrix $\mathbf{R}_x(k)$ is calculated to evaluate the determinant and the trace as follows:

$$\mathbf{R}_x(k) = E_l\Big[\mathbf{X}(k,l)\mathbf{X}^H(k,l)\Big], \tag{3}$$

where $E_l[\cdot]$ is the expectation operator over frame l and $(.)^{\mathrm{H}}$ is the Hermitian operator.

3.3.2. Analysis of the Determinant. The spatial covariance matrix of the observed signals $\mathbf{R}_x(k)$ is transformed using the mixing model in (1) as follows:

$$\begin{aligned} \mathbf{R}_x(k) &= E_l\Big[\mathbf{X}(k,l)\mathbf{X}^H(k,l)\Big] \\ &= \mathbf{A}(k)E_l\Big[\mathbf{S}(k,l)\mathbf{S}^H(k,l)\Big]\mathbf{A}^H(k) \\ &\equiv \mathbf{A}(k)\mathbf{R}_s(k)\mathbf{A}^H(k), \end{aligned} \tag{4}$$

where $\mathbf{R}_s(k)$ is the spatial covariance matrix of the source signals. According to the assumptions of FDICA, each source is independent from each other; therefore, $\mathbf{R}_s(k)$ is a diagonal matrix, and each element is equivalent to the source power $\sigma_i(k)$ on the kth frequency band as follows:

$$\mathbf{R}_s(k) = \begin{bmatrix} \sigma_1(k) & 0 \\ 0 & \sigma_2(k) \end{bmatrix}. \tag{5}$$

The determinant of the spatial covariance matrix, $\det\mathbf{R}_x(k)$, is written with (4) as follows:

$$\begin{aligned} \det\mathbf{R}_x(k) &= \det\Big\{\mathbf{A}(k)\mathbf{R}_s(k)\mathbf{A}^H(k)\Big\} \\ &= \det\mathbf{R}_s(k)\cdot\det\Big\{\mathbf{A}^H(k)\mathbf{A}(k)\Big\}, \end{aligned} \tag{6}$$

meanwhile, the determinant is calculated using the product of the eigenvalues, and for the diagonal matrix, the determinant consists of the product of the diagonal elements. The spatial covariance matrix of the source signals is the diagonal matrix; thus, $\det\mathbf{R}_s(k)$ of (6) can easily be obtained as the product of $\sigma_i(k)$. Substituting (2) into $\det\{\mathbf{A}^H(k)\mathbf{A}(k)\}$ of

(6), we obtain the propagation component of the determinant as follows:

$$\begin{aligned}
&\det \mathbf{A}^H(k)\mathbf{A}(k) \\
&= \det\begin{bmatrix} 2A_1^2(k) & A_1(k)A_2(k)\left\{e^{j\omega(k)(\tau_{21}(k)-\tau_{11}(k))} + e^{j\omega(k)(\tau_{22}(k)-\tau_{12}(k))}\right\} \\ A_1(k)A_2(k)\left\{e^{j\omega(k)(\tau_{11}(k)-\tau_{21}(k))} + e^{j\omega(k)(\tau_{12}(k)-\tau_{22}(k))}\right\} & 2A_2^2(k) \end{bmatrix} \\
&= 2A_1^2(k)A_2^2(k) \\
&\quad \times [1 - \cos\{\omega(k)(\tau_{11}(k) - \tau_{21}(k) - \tau_{12}(k) + \tau_{22}(k))\}],
\end{aligned} \tag{7}$$

consequently, the determinant is obtained as follows:

$$\begin{aligned}
&\det \mathbf{R}_x(k) \\
&\quad = 2A_1^2(k)A_2^2(k) \\
&\qquad \times [1 - \cos\{\omega(k)(\tau_{11}(k) - \tau_{21}(k) - \tau_{12}(k) + \tau_{22}(k))\}] \\
&\qquad \times \sigma_1(k)\sigma_2(k).
\end{aligned} \tag{8}$$

3.3.3. Analysis of the Trace. The trace is the sum of the diagonal elements, and this corresponds to the sum of the eigenvalues. As mentioned in Section 3.3.2, $\mathbf{R}_s(k)$ is the diagonal matrix, and thus, $\mathbf{R}_x(k)$ can be isolated for each source as follows:

$$\mathbf{R}_x(k) = \mathbf{A}_1(k)\mathbf{R}_{s1}(k)\mathbf{A}_1^H(k) + \mathbf{A}_2(k)\mathbf{R}_{s2}(k)\mathbf{A}_2^H(k), \tag{9}$$

where $\mathbf{R}_{si}(k)$ has only one element as follows:

$$\begin{aligned}
\mathbf{R}_{s1}(k) &= \begin{bmatrix} \sigma_1(k) & 0 \\ 0 & 0 \end{bmatrix}, \\
\mathbf{R}_{s2}(k) &= \begin{bmatrix} 0 & 0 \\ 0 & \sigma_2(k) \end{bmatrix},
\end{aligned} \tag{10}$$

and the mixing matrix $\mathbf{A}(k)$ also can be isolated as follows:

$$\begin{aligned}
\mathbf{A}_1(k) &= \begin{bmatrix} A_1(k)e^{-j\omega(k)\tau_{11}(k)} & 0 \\ A_1(k)e^{-j\omega(k)\tau_{12}(k)} & 0 \end{bmatrix}, \\
\mathbf{A}_2(k) &= \begin{bmatrix} 0 & A_2(k)e^{-j\omega(k)\tau_{21}(k)} \\ 0 & A_2(k)e^{-j\omega(k)\tau_{22}(k)} \end{bmatrix}.
\end{aligned} \tag{11}$$

Therefore, the trace of the spatial covariance matrix can be considered for each source, respectively, and the trace is invariant under the cyclic permutations, the trace of the spatial covariance matrix is transformed as follows:

$$\begin{aligned}
\operatorname{tr}\mathbf{R}_x(k) &= \operatorname{tr}\left\{\mathbf{A}_1(k)\mathbf{R}_{s1}(k)\mathbf{A}_1^H(k) + \mathbf{A}_2(k)\mathbf{R}_{s2}(k)\mathbf{A}_2^H(k)\right\} \\
&= \operatorname{tr}\left\{\mathbf{R}_{s1}(k)\mathbf{A}_1^H(k)\mathbf{A}_1(k)\right\} + \operatorname{tr}\left\{\mathbf{R}_{s2}(k)\mathbf{A}_2^H(k)\mathbf{A}_2(k)\right\}.
\end{aligned} \tag{12}$$

The term $\operatorname{tr}\{\mathbf{R}_{si}(k)\mathbf{A}_i^H(k)\mathbf{A}_i(k)\}$ is transformed via the isolated covariance matrix (10) and the isolated mixing matrix (11) as follows:

$$\operatorname{tr}\left\{\mathbf{R}_{si}\mathbf{A}_i^H(k)\mathbf{A}_i(k)\right\} = 2\sigma_i A_i^2(k). \tag{13}$$

Consequently, the trace is obtained as follows:

$$\operatorname{tr}\mathbf{R}_x(k) = A_1^2(k)\sigma_1(k) + A_2^2(k)\sigma_2(k). \tag{14}$$

3.3.4. Band Selection Process. For each criterion, we use the same rule to select frequency bands. Since each criterion (the determinant and the trace) is a real value, the selection is performed according to the largest magnitude of the criterion, until the number of bands selected reaches the designated number. The selection process works as follows.

(1) Calculate the spatial covariance matrix $\mathbf{R}_x(k)$.

(2) Calculate either criterion, $\det \mathbf{R}_x(k)$ or $\operatorname{tr}\mathbf{R}_x(k)$.

(3) Sort the criterion in descending order of magnitude.

(4) Select the designated number of bands from the sorted list of criteria.

The designated number of bands can be defined by system requirements or about 10% of the total number of bands can be selected (this is a rough estimate based on the experimental results discussed in a later section).

3.4. FDICA Separation Matrix. In this section, we explain how to obtain the FDICA separation matrix on the selected bands.

3.4.1. Learning the Separation Matrix on Selected Bands. Following STFT and band selection, the separation matrix $\mathbf{W}(k)$ is obtained using the general FDICA algorithm on the selected bands. A separated signal vector $\mathbf{Y}(k,l) = [Y_1(k,l), Y_2(k,l)]^T$ is formulated as follows:

$$\mathbf{Y}(k,l) = \mathbf{W}(k)\mathbf{X}(k,l). \tag{15}$$

In this paper, we use an iterative FDICA algorithm [11] as follows:

$$\begin{aligned}
&\mathbf{W}_{p+1}(k) \\
&\quad = \mathbf{W}_p(k) - \eta \cdot \text{off-diag}\left\{E_l\left[\phi(k,l)\mathbf{Y}^H(k,l)\right]\right\}\mathbf{W}_p(k),
\end{aligned} \tag{16}$$

where p is an iteration number, η is a step-size, and off-diag$(\cdot)$ denotes the operator at which all diagonal elements are set to zero. $\boldsymbol{\phi}(k,l) \equiv [\phi_1(k,l), \phi_2(k,l)]^T$ denotes the nonlinear function vector. Each function is $\phi_m(k,l) \equiv \text{sgn}(\text{Re}\{Y_m(k,l)\}) + j\,\text{sgn}(\text{Im}\{Y_m(k,l)\})$, and $\text{Re}\{\cdot\}$, $\text{Im}\{\cdot\}$ denote the real and imaginary parts, respectively. The function $\text{sgn}(\cdot)$ is used to obtain the sign of each value.

3.4.2. DOA Estimation from Separation Matrix. As mentioned in Section 2, the DOA of the source signals are estimated from the separation matrix, and the permutation is solved using the estimated DOA [9].

Each element of the separation matrix is represented as follows:

$$\mathbf{W}(k) \equiv \begin{bmatrix} w_{11}(k) & w_{12}(k) \\ w_{21}(k) & w_{22}(k) \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1(k) \\ \mathbf{w}_2(k) \end{bmatrix}, \tag{17}$$

where $\mathbf{w}_i(k) \equiv [w_{i1}(k)\ w_{i2}(k)]$. From the standpoint of array signal processing, we calculate the directivity pattern for each frequency band from $\mathbf{w}_i(k)$ and estimate the DOA of the source signals as follows:

$$\hat{\psi}_i(k) = \arg\min_{\psi} \left\{ \mathbf{w}_i{}^T(k)\boldsymbol{\kappa}(k,\psi) \right\}, \tag{18}$$

where $\boldsymbol{\kappa}(k,\psi) = [1, \mathrm{e}^{j\rho(k,\psi)}]$ ($\rho(k,\psi) \equiv 2\pi(kF_s/N)$ (d/c) $\sin(\psi)$) is a steering vector, ψ is a steering direction, d is the distance between the microphones, and c is the velocity of sound. $\hat{\psi}_i(k)$ is the directional null, and this corresponds to the source direction on each frequency band. Therefore, we solve the permutation problem using the same method in [9], and collected source directions $\hat{\theta}_i(k)$ are averaged over the selected bands to obtain the estimated DOA of the source signal, $\hat{\theta}_i$, as follows:

$$\hat{\theta}_i = \frac{1}{N_b} \sum_{k \in \Xi} \hat{\theta}_i(k), \tag{19}$$

where N_b is the number of bands and Ξ is a set of the selected bands.

3.4.3. Solving the Scaling Problem. The projection method [10] is applied to solve the scaling ambiguity for the FDICA separation matrix, and, as explained in this section, it is shown that the separated signal corresponds to one of observed signals based on the projection method.

The observed signal is transformed via (2) as follows:

$$\begin{aligned} \begin{bmatrix} X_1(k,l) \\ X_2(k,l) \end{bmatrix} &= \begin{bmatrix} X_{11}(k,l) + X_{12}(k,l) \\ X_{21}(k,l) + X_{22}(k,l) \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{e}^{-j\omega(k)\tau_{11}(k)} & \mathrm{e}^{-j\omega(k)\tau_{21}(k)} \\ \mathrm{e}^{-j\omega(k)\tau_{12}(k)} & \mathrm{e}^{-j\omega(k)\tau_{22}(k)} \end{bmatrix} \begin{bmatrix} A_1(k)S_1(k,l) \\ A_2(k)S_2(k,l) \end{bmatrix} \\ &\equiv \mathbf{B}(k)\mathbf{S}'(k,l), \end{aligned} \tag{20}$$

where $\mathbf{B}(k)$ consists of delay factors of the mixing matrix $\mathbf{A}(k)$ and $\mathbf{S}'(k,l)$ is the source signal with the observed amplitude. In other words, $\mathbf{B}(k)$ corresponds to the mixing matrix with the plane wave assumption. If the separation matrix corresponds to the inverse of the mixing matrix $\mathbf{B}(k)$, then the separation matrix formulated via Cramer's formula is as follows:

$$\begin{aligned} \mathbf{W}(k) &= \mathbf{D}(k)\mathbf{B}^{-1}(k) \\ &= \mathbf{D}(k)\frac{1}{\det \mathbf{B}(k)}\widetilde{\mathbf{B}}(k) \\ &\equiv \begin{bmatrix} \lambda_1(k) & 0 \\ 0 & \lambda_2(k) \end{bmatrix} \begin{bmatrix} \mathrm{e}^{-j\omega(k)\tau_{22}(k)} & -\mathrm{e}^{-j\omega(k)\tau_{21}(k)} \\ -\mathrm{e}^{-j\omega(k)\tau_{12}(k)} & \mathrm{e}^{-j\omega(k)\tau_{11}(k)} \end{bmatrix} \\ &\equiv \boldsymbol{\Lambda}(k)\widetilde{\mathbf{B}}(k), \end{aligned} \tag{21}$$

where $\mathbf{D}(k)$ is the matrix of the scaling ambiguity, $\boldsymbol{\Lambda}(k)$ is the scaling matrix that consists of the ambiguity $\mathbf{D}(k)$ and the inverse determinant of the matrix $\mathbf{B}(k)$, and $\widetilde{\mathbf{B}}(k)$ is part of the inverse matrix $\mathbf{B}(k)$.

The projection method [10] solves the scaling ambiguity by considering $\mathrm{diag}\,\mathbf{W}^{-1}(k)\mathbf{W}(k)$, where $\mathrm{diag}(\cdot)$ is the operator which remains diagonal elements of matrix $(\cdot)$ and the other all elements set zero. Therefore, the FDICA separated signal is obtained as follows:

$$\begin{aligned} \mathbf{Y}(k,l) &= \mathrm{diag}\{\mathbf{W}^{-1}(k)\mathbf{W}(k)\}\mathbf{X}(k,l) \\ &= \frac{1}{\det \widetilde{\mathbf{B}}(k)} \begin{bmatrix} \mathrm{e}^{-j\omega(k)\tau_{11}(k)} & 0 \\ 0 & \mathrm{e}^{-j\omega(k)\tau_{22}(k)} \end{bmatrix} \\ &\quad \times \boldsymbol{\Lambda}^{-1}(k)\boldsymbol{\Lambda}(k)\widetilde{\mathbf{B}}(k)\mathbf{B}(k)\mathbf{S}'(k,l) \\ &= \frac{1}{\det \mathbf{B}(k)} \begin{bmatrix} \mathrm{e}^{-j\omega(k)\tau_{11}(k)} & 0 \\ 0 & \mathrm{e}^{-j\omega(k)\tau_{22}(k)} \end{bmatrix} \\ &\quad \times \mathbf{I}\ (\{\det \mathbf{B}(k)\}\ \mathbf{I})\mathbf{S}'(k,l) \\ &= \begin{bmatrix} \mathrm{e}^{-j\omega(k)\tau_{11}(k)} & 0 \\ 0 & \mathrm{e}^{-j\omega(k)\tau_{22}(k)} \end{bmatrix} \mathbf{S}'(k,l) \\ &= \begin{bmatrix} X_{11}(k,l) \\ X_{22}(k,l) \end{bmatrix}. \end{aligned} \tag{22}$$

Consequently, the separated signal corresponds to part of the observed signal via the projection method.

Finally, $\mathrm{diag}\,\mathbf{W}^{-1}(k)\mathbf{W}(k)$ can be considered as the new separation matrix that the scaling ambiguity is solved by, using the projection method.

3.5. Frame-Wise Spectral Soft Mask. In this section, we propose a frame-wise spectral soft mask method which improves separation performance in the low-frequency region.

3.5.1. Tentative Separation and Analysis. We consider the delay $(d/c)\sin(\hat{\theta}_i)$ from the DOA obtained in Section 3.4.2, and this is applied to the NBF to obtain tentative separated signals. As mentioned in Section 3.2, propagation was assumed to be from the source to the center of the microphone. Therefore, the propagation delay τ_{iS} can be omitted without

a loss of generality if we assume that the distance between the microphones is small enough, such as in mobile devices. In this case, the direction only depends on the delay $\tau_{ij}(k)$. In addition, the distance between the source position and the microphone position can be represented as the gain $A_i(k)$. Therefore, the observed signal $X_j(k,l)$ can be written as follows:

$$X_j(k) = A_1(k)S_1(k,l)\mathrm{e}^{-j\omega(k)\tau_{1j}(k)} + A_2(k)S_2(k,l)\mathrm{e}^{-j\omega(k)\tau_{2j}(k)}. \tag{23}$$

Next, we need to consider tentative signal separation. The right half of Figure 2 shows the block diagram of the tentative separation, and this process corresponds to NBF. The output signals of the NBF are formulated as follows:

$$\begin{aligned} Y_1^{\mathrm{NBF}}(k,l) &= X_1(k,l)\mathrm{e}^{-j\omega(k)\hat{\tau}_{22}} - X_2(k,l)\mathrm{e}^{-j\omega(k)\hat{\tau}_{21}}, \\ Y_2^{\mathrm{NBF}}(k,l) &= -X_1(k,l)\mathrm{e}^{-j\omega(k)\hat{\tau}_{12}} + X_2(k,l)\mathrm{e}^{-j\omega(k)\hat{\tau}_{11}}, \end{aligned} \tag{24}$$

where we consider the estimated delay for each channel as follows:

$$\begin{aligned} \hat{\tau}_{11} &= \frac{-(d/c)\sin\left(\hat{\theta}_1\right)}{2}, & \hat{\tau}_{12} &= \frac{(d/c)\sin\left(\hat{\theta}_1\right)}{2}, \\ \hat{\tau}_{21} &= \frac{(d/c)\sin\left(\hat{\theta}_2\right)}{2}, & \hat{\tau}_{22} &= \frac{-(d/c)\sin\left(\hat{\theta}_2\right)}{2}. \end{aligned} \tag{25}$$

Substituting (23) into (24), we obtain the relationship between the source signals $S_1(k,l)$, $S_2(k,l)$ and the tentative separated signal $Y_1^{\mathrm{NBF}}(k,l)$ as follows:

$$\begin{aligned} Y_1^{\mathrm{NBF}}(k,l) &= X_1(k,l)\mathrm{e}^{-j\omega(k)\hat{\tau}_{22}} - X_2(k,l)\mathrm{e}^{-j\omega(k)\hat{\tau}_{21}} \\ &= A_1(k)S_1(k,l)\mathrm{e}^{-j\omega(k)(\tau_{11}(k)+\hat{\tau}_{22})} \\ &\quad + A_2(k)S_2(k,l)\mathrm{e}^{-j\omega(k)(\tau_{21}(k)+\hat{\tau}_{22})} \\ &\quad - A_1(k)S_1(k,l)\mathrm{e}^{-j\omega(k)(\tau_{12}(k)+\hat{\tau}_{21})} \\ &\quad - A_2(k)S_2(k,l)\mathrm{e}^{-j\omega(k)(\tau_{22}(k)+\hat{\tau}_{21})}. \end{aligned} \tag{26}$$

In the reverberant condition, a deviation in the direction of the direct sound is caused by the reflected sound; however, the deviation is quite small, because, in general, direct sound is much stronger than reflected sound. Therefore, we assume that the estimated direction $\hat{\theta}_i$ is approximately equivalent to the source direction θ_i; thus, $\tau_{ij}(k) \approx \hat{\tau}_{ij}$ can be an appropriate assumption. $Y_2^{\mathrm{NBF}}(k,l)$ is calculated in the same way, and consequently, we obtain the approximate relationship as follows:

$$Y_i^{\mathrm{NBF}}(k,l) \approx A_i(k)S_i(k,l)\left\{\mathrm{e}^{-j\omega(\hat{\tau}_{11}+\hat{\tau}_{22})} - \mathrm{e}^{-j\omega(\hat{\tau}_{12}+\hat{\tau}_{21})}\right\}. \tag{27}$$

In the lower-frequency region, $\omega(k)$ is smaller than in the higher frequency region, and thus, the amplitude of the delay section of (27) takes a smaller value due to the lower frequency. This is the reason for the degradation of the separated signals when using the conventional method, as mentioned in Section 2.

3.5.2. Interchannel Separated Signal Mask. First, we consider a cost function as follows:

$$\min_{\alpha_i} E_l\left\{\left(A_i(k)S_i(k,l) - \alpha_i X_i(k,l)\right)^2\right\}, \tag{28}$$

where E_l is the expectation operator and α_i is a mask function. To minimize the cost function (28), the differentials of α_i are considered, and the independence between each source signal is also considered; when we utilize the independence and expectation, cross-correlation between each source signal become zero. Therefore, we obtain the Wiener solution and it is approximated by the tentative separated signals in (27) as follows:

$$\begin{aligned} \alpha_i &= \frac{E_l\left\{A_i^2(k)|S_i(k,l)|^2\right\}}{E_l\left\{A_1^2(k)|S_1(k,l)|^2\right\} + E_l\left\{A_2^2(k)|S_2(k,l)|^2\right\}} \\ &\approx \frac{E_l\left\{\left|Y_i^{\mathrm{NBF}}(k,l)\right|^2\right\}}{E_l\left\{\left|Y_1^{\mathrm{NBF}}(k,l)\right|^2\right\} + E_l\left\{\left|Y_2^{\mathrm{NBF}}(k,l)\right|^2\right\}}. \end{aligned} \tag{29}$$

In this paper, we consider the shortest expectation in order to reduce computational complexity, so the frame-wise soft mask $M_i(k,l)$ is obtained as follows:

$$M_i(k,l) = \frac{\left|Y_i^{\mathrm{NBF}}(k,l)\right|^2}{\left|Y_1^{\mathrm{NBF}}(k,l)\right|^2 + \left|Y_2^{\mathrm{NBF}}(k,l)\right|^2}. \tag{30}$$

The obtained soft mask varies frame by frame, and thus, it can trace temporal changes in each speech source signal.

Finally, the output separated signals are obtained as follows:

$$\mathbf{Y}(k,l) = \begin{cases} \mathbf{W}(k)\mathbf{X}(k,l), & \text{(FDICA)}, \\ \begin{bmatrix} M_1(k,l)X_1(k,l) \\ M_2(k,l)X_2(k,l) \end{bmatrix} & \text{(Soft Mask)}. \end{cases} \tag{31}$$

4. Experimental Results and Estimation of Computational Complexity

In this section, we estimate the computational complexity of the proposed method, and evaluate its performance using a source separation simulation.

4.1. Estimation of Computational Complexity. We estimate the number of operations (multiplication and addition as floating operations) based on (16) to evaluate the efficiency of the proposed method. The parameters of FDICA are shown in Table 3, and the estimate of the computational complexity is shown in Table 1. The number of bands for the estimate is 64, which is determined from the experimental results in Section 4.5. The unit "MOPs" represents the number of mega operations per second.

As shown in Table 1, the proposed method achieves an over 80% reduction in the level of computational complexity compared with unmodified FDICA and is almost equivalent to the conventional method [8].

Table 1: Computational complexity.

	FDICA	Conventional	Proposed
Number of operations (MOPs)	287 (100%)	46 (16%)	48 (17%)

Table 2: Signals for simulation.

	Anechoic	Reverberant
Samp. freq.	8 (kHz)	
Rev. time	—	500 (msec)
Voice type	Male (2), Female (2)	
Location pair	$\{-45, 45\}$, $\{-90, 0\}$, $\{-45, 0\}$ (deg)	

Table 3: FDICA parameters.

FFT size	1024 (sample)
FFT shift	256 (sample)
Learning time	3 (sec)
Iteration	max. 200 (times)
Step size	0.01
Initial matrix	Identity
Permutation solver	DOA [9]
Scaling solver	Projection method [10]

4.2. Simulation Conditions. The speech signals are recorded with two omnidirectional microphones (SHURE SM93) and the distance between them is 3.6 cm. The recording conditions are shown in Table 2 and Figure 3.

The voice signals are played back via loudspeakers and recorded individually, and the observed signals are mixed when the simulation is performed. The parameters of FDICA are shown in Table 3. The number of bands ranges from 384 to 32, because these numbers are roughly the ratio of integers, $3/4, 1/2, \ldots, 1/16$, for the number of frequency bands, which was 513.

As mentioned in Section 1, FDICA must store the observed signals whose lengths are longer than a few seconds. Additionally, the separated signals to estimate higher-order statistics are different in every iteration, because the separation matrix is updated in every iteration. These are the primary reasons that the computational complexity required for FDICA is high; therefore, the smaller number of selections contributes to a lower computational complexity, and thus results in smaller memory consumption as well.

Although the classical FDICA algorithm is used in this paper (see Section 3.4.1), it can be replaced with any state-of-the-art FDICA method. Therefore, the proposed method, which consists of band selection and the use of a frame-wise soft mask for the nonselected bands, is an improvement of the FDICA method, without a loss of generality.

4.3. Comparison between the Determinant and the Trace. In this section, we introduce band-selection performance to show which criteria is more appropriate.

As mentioned in Section 3.3, if there is only one source in one of frequency bands, the rank of $\mathbf{R}_x(k)$ is not full, and the determinant of $\mathbf{R}_x(k)$ becomes zero. According to (14), if there is only one source in one of frequency bands, the trace never becomes zero. In addition, (8) contains the expression $1 - \cos(\{\omega(k)\})$, and this works as a weighting factor, limiting the frequency bands. Therefore, the expression $1 - \cos(\{\omega(k)\})$ means that band selection is unlikely to occur in the low-frequency region via the determinant criterion. As mentioned in Section 1, the phase difference becomes small in the low-frequency region, and this is a disadvantage of using microphone array signal processing methods such as FDICA. In light of these considerations, the trace has a tendency to select bands occurring in the low-frequency region, more so than the determinant. This implies that the determinant is better suited than the trace to select frequency bands.

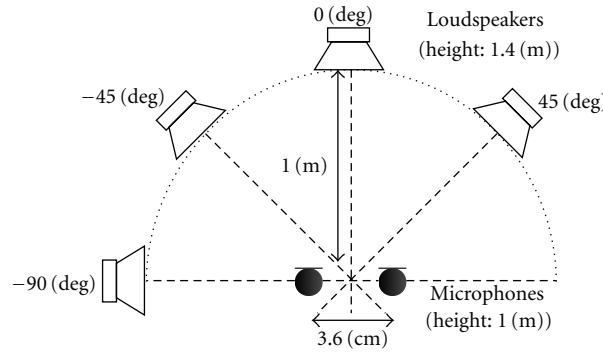


Figure 3: Recording conditions.

We evaluated this tendency via a band-selection experiment in which the DOA is estimated from the separation matrix. In this experiment, we wanted to confirm the difference in the performance of the DOA estimation task using the separation matrix. Therefore, the FDICA algorithm and parameters are the same as in the other experiments, but the simulation is performed only for the anechoic condition and with 32 bands selected. This is because in the reverberant condition, there is some deviation in the estimated DOA, and it is difficult to evaluate the difference in performance from the low-frequency band to the high-frequency band.

The covariance matrix is calculated using recorded speech signals on each frequency band via STFT, then the determinant and trace are calculated, respectively. Following the calculation of the criterion, some bands are selected according to the selection process in Section 3.3.4. For the selected bands, the FDICA separation matrix is iteratively learned using (16), and the DOA of the source signal is estimated via the FDICA separation matrix on each frequency band. In this evaluation, the number of the selected bands is set at 32, because it is easier to confirm the tendency in which bands are selected. The direction of the source signals was set about $\{-45, 45\}$ (deg) (but this direction is not precise, because the loudspeakers were set by hand).

Figure 4 shows the determinant and the trace of the covariance matrix on each frequency band. The horizontal and vertical axis are frequency (Hz), and the normalized value of the determinant and trace, respectively. The determinant

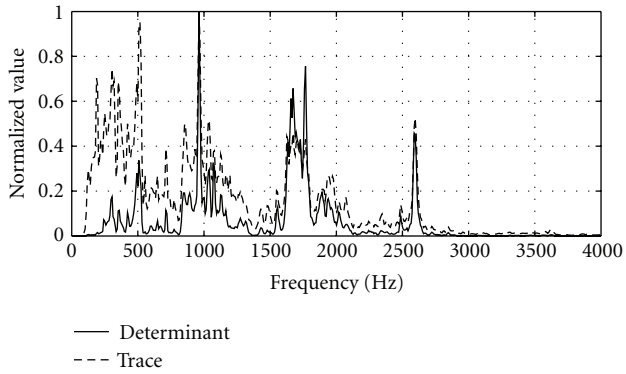


Figure 4: Example of the determinant/trace of the covariance matrix.

and trace are averaged over the different voice types. Normalization is applied to show the characteristics, so each value is divided by the maximum value. Band selection was performed based on the relative value of the determinant or the trace over the frequency bands. Therefore, the normalization is appropriate without a loss of generality. In Figure 4, the values in the low-frequency region show significant differences, especially that the values for the determinant are lower than those for the trace.

Figure 5 shows experimental results of the DOA estimation from the FDICA separation matrix. The DOA estimation might be affected by noise signals in the real world, for example, acoustical ambient noise or electric circuit noise; therefore, the performance of the beamformer is degraded in the low-frequency region. As shown in Figure 5(b), the trace criteria selects more bands in the low-frequency region than the determinant criteria. In addition, since this experiment was performed in the anechoic condition, the estimated directions should not vary among the selected bands. However, the estimated directions via the trace criteria are more varied than the determinant criteria in the low-frequency region, and this fact implies that noise signals affected the DOA estimation, as previously mentioned. Consequently, DOA estimation via the determinant criteria is more efficient and precise than estimation using the trace criteria; in other words, it is advantageous to use the determinant for band selection.

4.4. Evaluation Measure. The performance of the proposed method is evaluated using the segmental signal-to-noise ratio ($\mathrm{SNR}_{\mathrm{seg}}$) [12] and cepstral distortion (CD) [13]. $\mathrm{SNR}_{\mathrm{seg}}$ is a very common method for evaluating noise suppression, for example, in high-efficiency coding such as MP3. In general, $\mathrm{SNR}_{\mathrm{seg}}$ is known to have a better correlation with the perception of noisy speech by humans than entire interval SNR [12]. The proposed method is based on the spectral soft mask method, and thus, the degradation of the separated signals can be estimated. Therefore, $\mathrm{SNR}_{\mathrm{seg}}$ is appropriate for evaluating the proposed method. CD is another measure of the degree of distortion via the cepstrum domain, and this can evaluate distortion of a spectral envelope.

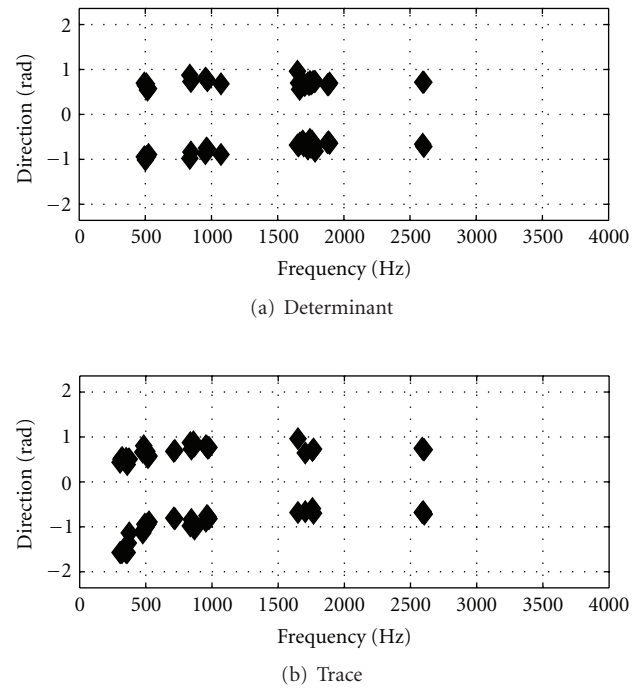


Figure 5: Estimated DOA via the determinant/trace selection criteria.

When we evaluate the proposed method, each frequency domain signal is transformed into a time-domain signal via inverse STFT. As mentioned in Table 3, the projection method [10] is used to solve the scaling ambiguity of FDICA, and this means that each separated signal corresponds to one of the observed signals.

Therefore, we evaluate the performance between one of the observed signals and the separated signal. Each term of the right side in (23) corresponds to one of the observed signals. The observed signal in the time domain is formulated as follows:

$$\begin{aligned} x_1(n) &= x_{11}(n) + x_{21}(n), \\ x_2(n) &= x_{12}(n) + x_{22}(n), \end{aligned} \tag{32}$$

where each separated signal in the time domain is $y_i(n)$ and i is the source number. In this case, $\mathrm{SNR}_{\mathrm{seg}}$ is defined as follows:

$$\mathrm{SNR}_{\mathrm{seg}} \equiv \frac{1}{2}\sum_i \frac{1}{N_{l_s}} \sum_{l_s} \frac{\sum_m x_{ii}^2(m,l_s)}{\sum_m \{x_{ii}(m,l_s) - y_i(m,l_s)\}^2}, \tag{33}$$

where $x_{ij}(m,l)$ and $y_i(m,l)$ are the observed signal and the separated signal in the time domain of frame l_s and time m in the frame, j is the microphone number, N_{l_s} is the number of frames, and the obtained $\mathrm{SNR}_{\mathrm{seg}}$ for each channel are

averaged. We calculate CD from the speech components, and it is defined as follows:

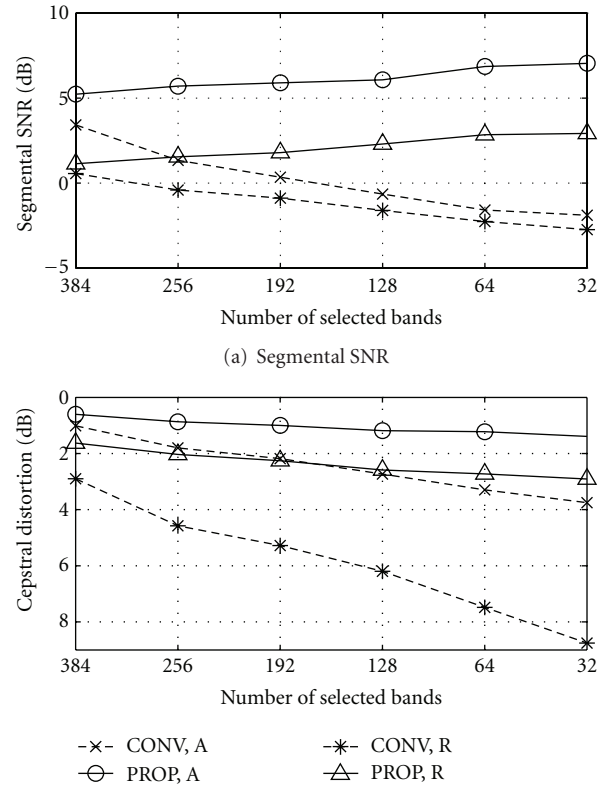
$$\mathrm{CD} \equiv \frac{1}{2}\sum_{i}\frac{20}{N_{l_c}\log 10} \times \left\{\sum_{l_c}\sqrt{\sum_{\nu=1}^{B}2\left(C_{x_{ii}(m,l_c)}(\nu,l_c)-C_{y_i(m,l_c)}(\nu,l_c)\right)^2}\right\}, \tag{34}$$

where $C_{(\cdot)}(\nu, l_c)$ is the νth cepstral coefficient of the signal $(\cdot)$ in frame l_c, and N_{l_c} is the number of frames. The obtained CD values for each channel are averaged ("1/2" means an average for two channels). B is the number of dimensions of the cepstrum used in the evaluation; we set $B = 20$. A small CD value indicates that the sound quality of the separated signal is high.

4.5. Simulated Source Separation Results. Figure 6 shows the performance of the proposed method. The x-axis shows the number of bands, and the y-axis shows $\mathrm{SNR}_{\mathrm{seg}}$ and CD. The y-axis of CD has been flipped, because a lower CD value indicates better sound quality. "CONV" denotes the conventional method [8], and "PROP" denotes the proposed method. "A" and "R" denote anechoic and reverberant conditions. Figure 6 shows that the proposed method is significantly better than the conventional method and that as the number of bands falls, we can confirm that $\mathrm{SNR}_{\mathrm{seg}}$ is improved and CD is degraded using the proposed method. This means that the proposed method has a tradeoff, and that 64 may be the best number of bands to select.

Figure 7 shows the performance comparison of BSS. The x-axis shows $\mathrm{SNR}_{\mathrm{seg}}$, and the y-axis shows CD. Again, the y-axis is turned over, because CD improves with a smaller value. "PROP", "FDICA", and "DUET" denote the proposed method, FDICA and DUET, respectively, while "A" and "R" denote anechoic and reverberant conditions. In this comparison, the number of bands selected is 64. In this experiment, we only used the time difference to perform DUET because source power was identical for each source. In Figure 7, the upper right corner shows better performance, and the lower left corner shows worse performance. Therefore, we can confirm that the sound quality of the proposed method is better than FDICA for $\mathrm{SNR}_{\mathrm{seg}}$ and better than DUET for CD. This tendency is shown in both anechoic and reverberant conditions.

Consequently, the degradation of the conventional method is significantly improved using the proposed method, as shown in Figure 6. The proposed method is somewhat better than both FDICA and DUET. In addition, as mentioned in Section 4.1, the efficiency of the proposed method was confirmed using a computational estimate. Therefore, from these results, both the effectiveness and efficiency of the proposed method are confirmed.



(a) Segmental SNR

(b) Cepstral distortion

Figure 6: Performance of proposed method.

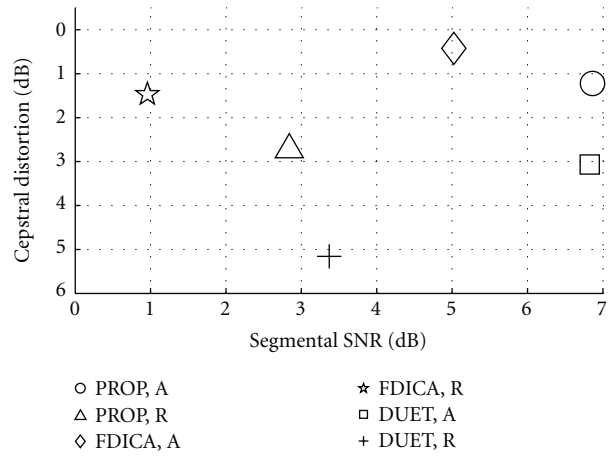


Figure 7: Performance comparison.

5. Conclusion

A blind source separation method involving lower computational complexity was proposed. This method is based on FDICA, with band selection and a frame-wise spectral soft mask derived from the tentative separated signals. The band selection process and its theoretical background were given,

and it was shown that the determinant criteria is more advantageous than the trace criteria for band selection. A theoretical analysis of the tentative separated signals was given, and the proposed frame-wise spectral soft mask was generated from just the power of the tentative separated signals. The efficiency of the proposed method is evaluated using a computational estimate, which showed that only a small amount of additional computation is required compared to the conventional method. In addition, regarding band selection, the determinant of the spatial covariance matrix provided *rough* estimation of the number of sources, when a weighting factor over frequency bands was used to prevent distortion of the beamformer's output. The effectiveness of the proposed method was evaluated via source separation simulations in the anechoic and reverberant conditions; the sound quality of the proposed method is much better than the conventional method and somewhat better than FDICA or DUET.

The proposed method is dramatically lower in computational complexity (80% lower) than the unmodified FDICA method with the number of selected bands at 64 (1/8 of the total number of frequency bands), while significantly improving performance over the conventional method. Although the proposed method is practical enough, future work will include analysis and evaluation in more typical, noisier, environments.

References

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

[2] Y. Takahashi, T. Takatani, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array with independent component analysis for hands-free speech recognition," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC '06)*, September 2006.

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1–3, pp. 21–34, 1998.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1846, 2004.

[5] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2CH BSS using the em algorithm in reverberant environment," in *Proceedings of the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '07)*, pp. 147–150, October 2007.

[6] N. Duong, E. Vincent, and R. Gribonval, "Under-determined convolutive blind source separation using spatial covariance models," in *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 9–12, March 2010.

[7] K. Kondo, Y. Takahashi, S. Hashimoto, H. Saruwatari, T. Nishino, and K. Takeda, "Efficient blind speech separation suitable for embedded devices," in *Proceedings of the 2011 European Signal Processing Conference (EUSIPCO '11)*, Barcelona, Spain, August 2011.

[8] K. Kondo, M. Yamada, and H. Kenmochi, "A semi-blind source separation method with a less amount of computation suitable for tiny dspmodules," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (InterSpeech '09)*, pp. 1339–1342, September 2009.

[9] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluationof blind signal separation method using directivity pattern under reverberant conditions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00)*, vol. 5, pp. 3140–3143, June 2000.

[10] M. Matsumoto and Y. Ikeda, "A method of ica in time-frequency domain," in *Proceedings of the 1999 International Test Conference (ITC '99)*, pp. 365–371, January 1999.

[11] N. Murata and S. Ikeda, "An on-line algorithm for blind source separationon speech signals," in *Proceedings of the International Symposium on Nonlinear Theory and Its Applications (NOLTA '98)*, pp. 923–926, September 1998.

[12] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, New York, NY, USA, 1993.

[13] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, NY, USA, 1993.