Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2012, Article ID 324398, 11 pages
doi:10.1155/2012/324398

Research Article

Practically Efficient Blind Speech Separation Using Frequency Band Selection Based on Magnitude Squared Coherence and a Small Dodecahedral Microphone Array

Kazunobu Kondo,[1] Yusuke Mizuno,[2] Takanori Nishino,[3] and Kazuya Takeda[3]

[1] *Corporate Research & Development Center, Yamaha Corporation, 203 Matsunokijima, Iwata 438-0192, Japan*
[2] *Graduate School of Engineering, Mie University, 1515 Kurimamachiya-cho, Tsu 514-0102, Japan*
[3] *Graduate School of Information Science, Nagoya University, Chikusa-ku Furou-cho, Nagoya 464-8603, Japan*

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

Received 6 July 2012; Revised 13 August 2012; Accepted 20 August 2012

Academic Editor: Xiao-fei Zhang

Small agglomerative microphone array systems have been proposed for use with speech communication and recognition systems. Blind source separation methods based on frequency domain independent component analysis have shown significant separation performance, and the microphone arrays are small enough to make them portable. However, the level of computational complexity involved is very high because the conventional signal collection and processing method uses 60 microphones. In this paper, we propose a band selection method based on magnitude squared coherence. Frequency bands are selected based on the spatial and geometric characteristics of the microphone array device which is strongly related to the dodecahedral shape, and the selected bands are nonuniformly spaced. The estimated reduction in the computational complexity is 90% with a 68% reduction in the number of frequency bands. Separation performance achieved during our experimental evaluation was 7.45 (dB) (signal-to-noise ratio) and 2.30 (dB) (cepstral distortion). These results show improvement in performance compared to the use of uniformly spaced frequency band.

1. Introduction

Speech communication and recognition systems are widely used in the present-day world, generally under reverberant and noisy conditions. An acoustic sound field is described by source signals and impulse responses which correspond to source locations and reflections, in other words, virtual source locations. The voice terminals are usually equipped with microphones, which are used to observe speech signals. In general, the observed signals include some source speech signals, mixed with each other and with the acoustic sound field. Extracting source signals and their locations, which is called encoding an acoustic field, is an important technique for acoustic schemes such as highly realistic communication and speech recognition systems. Blind source separation (BSS) is a useful method used to extract the sound source signals, and frequency domain independent component analysis (FDICA) [1] is one of the most commonly used BSS methods. FDICA achieves significant separation performance; however a permutation problem is one of the fundamental drawbacks to the use of FDICA. Various methods have been proposed to solve this permutation ambiguity such as [2–7] and include using power envelopes of separated signals at neighboring frequency channels, similarity between directivity patterns formed by a separation matrix, and large microphone arrays which surround sound sources. A correlation of power envelopes of separated signals can be observed at neighboring frequency channels under a condition that their frequencies are very close to each other. For wideband sound sources, it is difficult that this assumption can be satisfied.

A blind source separation method using a dodecahedral microphone array (DHMA) system for the acoustic field encoding has been proposed by Ogasawara et al. [8]. This

method introduces a permutation solver based on a combination of amplitude and phase similarities. Experimental results under reverberant conditions show that significant improvement in separation performance can be achieved by using a dodecahedral shaped microphone array, which can effectively use amplitude differences obtained from the surfaces of the array for a permutation correction. However, the BSS method for DHMA uses FDICA to separate the sound sources, which involves a high degree of computational complexity for two reasons. (1) A DHMA can include up to 160 microphones, resulting in a large separation matrix. Estimating the separation matrix requires a very large matrix calculation, and, furthermore, it must be performed with an iterative update. (2) Solving the permutation problem requires calculation of all of the possible combinations of transfer function similarities via hierarchical clustering. Since voice terminals are generally implemented in embedded systems, in order to realize implementation in practical systems, computational complexity must be low. Therefore, it is very important and useful to discover a way to reduce computational complexity while achieving nearly equivalent performance to conventional methods used by speech communication and recognition systems.

For a faster convergence to estimate the separation matrix of FDICA, only a limited number of frequency bands with uniformly spaced intervals can be selected for the estimation [9]. The other bands are interpolated by a direction of arrival estimation and a null beamformer method, and the number of the frequency bands is increased with iterations. Closed-form ICA using second-order statistics (SOS) is most effective method to obtain the separation matrix [10]; however the separation performance is lower than nonclosed-form higher-order ICA. Therefore, the closed-form ICA is used to obtain initial matrix for the higher-order ICA in [11]. These methods are more effective than other methods previously proposed, although they cannot reduce the computational complexity of the permutation correction. A permutation-free ICA has also been proposed [12, 13]. This method considers all of the frequency bands as a single vector in order to estimate the separation matrix; therefore it does not need a permutation solver; however, the size of the matrix of the iterative update is much larger than that of conventional FDICA methods. Joint diagonalization using SOS is known as an efficient BSS algorithm such as [14–16]. DFT transformations in the iterative update prevent a complete decoupling of bandwise frequency components; in other words, this means that permutation solver is not required [16]. However, the backward and forward transformation must be applied in each iteration. On the other hand, to cope with the issue of high computational complexity, a BSS method which restricts the number of frequency bands has been previously proposed by the authors, and it has shown almost equivalent performance to unmodified FDICA [17–19].

In this paper, we propose a BSS method with a frequency band selection method suitable for a DHMA, in order to reduce computational complexity. This method uses the spatial characteristics of a DHMA, and results show separation performance nearly equivalent to the method proposed in [8]. The rest of the paper is organized as follows. In Section 2, we briefly introduce the conventional method. In Section 3, we introduce the proposed method based on the magnitude squared coherence of a DHMA and perform an estimation of computational complexity. In Section 4, we show the results of a source separation experiment using our proposed method. Section 5 presents our conclusions and describes future work.

2. Dodecahedral Microphone Source Separation

2.1. Dodecahedral Microphone Array. Figure 1 shows the dodecahedral microphone array (DHMA). Microphones are installed on ten faces, excluding the top and bottom faces, and 16 holes appear on each face. This means that the DHMA is an agglomerative microphone array; it is small enough that its portability can be considered an advantage. In addition, at each face of array the observed signals have different acoustic features, such as sound pressure levels, arrival times, and influence of diffraction waves. Our DHMA uses small omnidirectional microphones (SONY ECM-77B); six microphones are installed on each face because it is difficult to adjust the characteristics of the microphones in relation to one another. In the conventional method, microphone gain is also adjusted manually.

Characteristics of the DHMA are fully described in [8]; thus here we introduce them only briefly. The DHMA has two remarkable merits: amplitude differences on different faces and spatial aliasing. Firstly, amplitude differences between the microphones installed on different faces can be enlarged, even though the structure of the DHMA is small. Secondly, spatial aliasing related to spatial sampling is not easily achieved, even at high frequencies. In order to obtain large differences in sound pressure using conventional microphone arrays, large distances are needed between microphones or microphone arrays. If the DHMA faces the sound source, acoustic pressure on one of the faces is high, while pressures on the opposite faces are low. When the acoustic pressure distribution observed by the whole surface of the DHMA is compared with a spherical microphone array, the acoustic pressure difference is greater when obtained with a DHMA at frequencies above 6 kHz. This shows the advantage of the DHMA from the viewpoint of acoustic pressure difference.

2.2. Blind Source Separation Using DHMA. The conventional method of BSS is based on FDICA. The permutation problem is solved by the physical characteristics of the DHMA, namely, its dodecahedral shape. In addition, this method does not require prior information, such as the number of sound sources or source locations.

The number of the source signals is estimated using the eigenvalue of the spatial covariance matrix. The number of dimensions of FDICA is reduced using a subspace method (PCA) from the number of microphones. The separation matrix is estimated by FDICA, the scaling problem is solved using the projection method [20], and the permutation

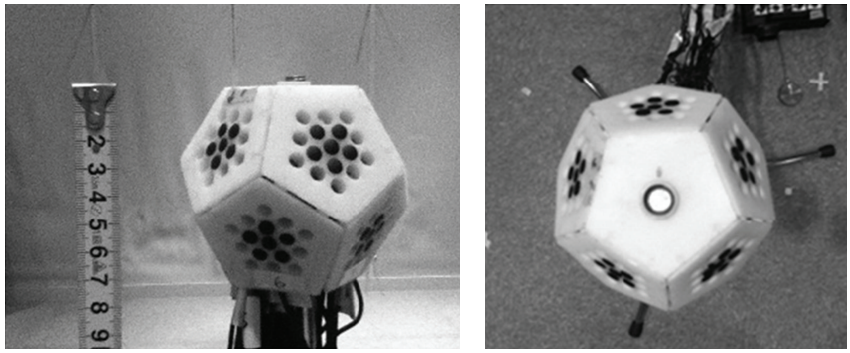

Figure 1: Dodecahedral microphone array (DHMA).

problem is solved by acoustic transfer function clustering. In the case of a DHMA, the transfer function clustering method results in significantly improved permutation correction under reverberant conditions; however this method involves a high degree of computational complexity due to similarity comparisons between all of the transfer functions.

3. Proposed Method

In this paper, we propose a method involving less computational complexity than the conventional method [8]. A block diagram is shown in Figure 2.

3.1. Frequency Band Selection Based on Spatial Coherence. To improve the efficiency and effectiveness of the frequency domain BSS method, frequency band selection has been proposed by the authors in [17–19]. This method showed significant reduction of computational complexity and equivalent performance compared to the unmodified BSS method. Although the classical FDICA algorithm is used in the experiment [17–19], it can be replaced with any state-of-the-art FDICA method with the permutation solver. In other words, the band selection method is not restrained by the update rule of FDICA, so there is no loss of generality.

3.1.1. Magnitude Square Coherence. When estimating the separation matrix of FDICA, it is very important that the training of the separation matrix is performed on highly separable frequency bands when using band selection method. In this paper, magnitude squared coherence (MSC) is considered as a method of selecting the separable frequency bands. MSC corresponds to a measure of the interference between two signals and is formulated as follows:

$$C_{x_1x_2}(k) = \frac{E_l\left[\,|P_{x_1x_2}(k,l)|^2\right]}{E_l[\,|P_{x_1x_1}(k,l)|\,]E_l[\,|P_{x_2x_2}(k,l)|\,]}, \tag{1}$$

where $P_{x_1x_1}(k,l)$ and $P_{x_2x_2}(k,l)$ are the power spectrums of x_1 and x_2, respectively, and $P_{x_1x_2}(k,l)$ is a cross-spectrum. k and l denote frequency band index and frame index, respectively. $E_l[\cdot]$ is the expectation operator over frame l. $C_{x_1x_2}(k)$ is MSC between two signals x_1 and x_2 in the frequency band k. The formulation of MSC is the normalized cross-spectrum on each frequency band, and thus the range of MSC shows $0 \le C_{x_1x_2}(k) \le 1$. In the case of a diffused noise field, MSC is formulated as follows:

$$C_{x_1x_2}(k) = \operatorname{sinc}\left(\frac{2\pi k F_s}{N_F} d_{\text{mic}} c^{-1}\right)^2, \tag{2}$$

where $\operatorname{sinc}(\cdot)$ means the sinc function $(\sin(x)/x)$, F_s is a sampling frequency, N_F is the FFT size, d_{mic} is a microphone distance, and c is the velocity of sound. A theoretical formulation of the diffused noise field is used to evaluate characteristics of the noise field condition [21] or to model the noise field for the postfiltering of the microphone array processing [22]. BSS conducted under the condition of multiple sources can be considered as a diffused noise field. A diffused noise field means that sound waves are randomly arriving from every possible direction, so that observed signals at two microphones have a variety of phase differences according to the direction of the sound sources. In the high frequency region, larger phase differences can be observed than in the low frequency region, and phase differences vary more widely. Therefore, weaker coherence characteristics are observed in the high frequency region, which results in MSC assuming smaller values. Consequently, MSC can evaluate the phase difference between two signals, and we can assume that MSC can contribute to increasing separation performance by allowing us to select frequency bands with small MSC values.

3.1.2. Characteristics of Magnitude Squared Coherence for a DHMA. In this section, we experimentally evaluate the effectiveness of using MSC for a DHMA. As mentioned in Section 2.1, the acoustic pressure distribution of a DHMA is different from that of spherical microphone arrays. In [8], this comparison was made experimentally because the shape of a DHMA makes it too complicated to evaluate this characteristic theoretically. For the same reason, we use experimental evaluation in the current study. Two microphones are arbitrarily selected from the 60 microphones of the DHMA (6 microphones are installed on each face), and the MSC of these two microphones is calculated using the measured impulse responses. The conditions used to evaluate the experimental MSC are shown in Table 1, and

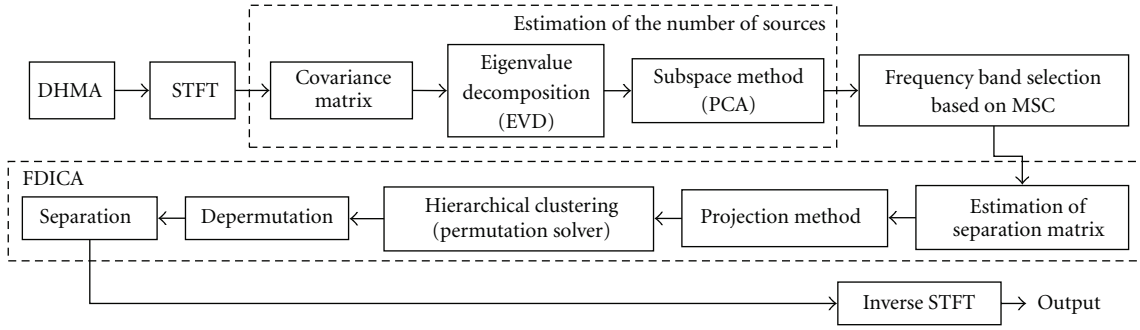

Figure 2: Block diagram of the proposed method.

Table 1: Simulation conditions.

Sampling frequency	40 (kHz)
Source signal	Speech (6 males, 6 females), 4 (sec)
Target frequency region	0–8 (kHz)
Number of sources	12
Velocity of sound	340 (m/sec)
Reverberation time	138 (msec)
Window function	Hann
Window length	1024 (sample)
Shift length	256 (sample)
FFT length (N_F)	1024 (sample)

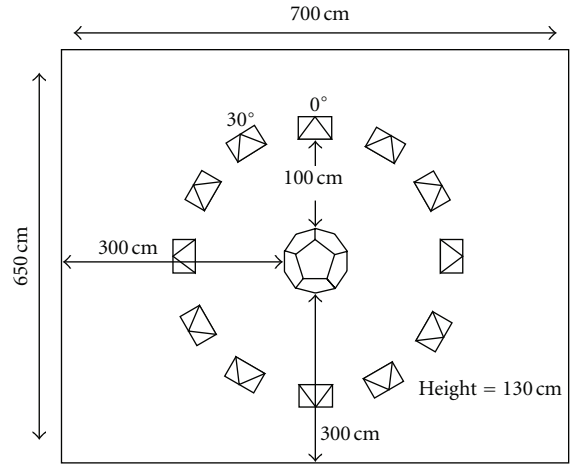

Figure 3: Sources and loudspeaker positions.

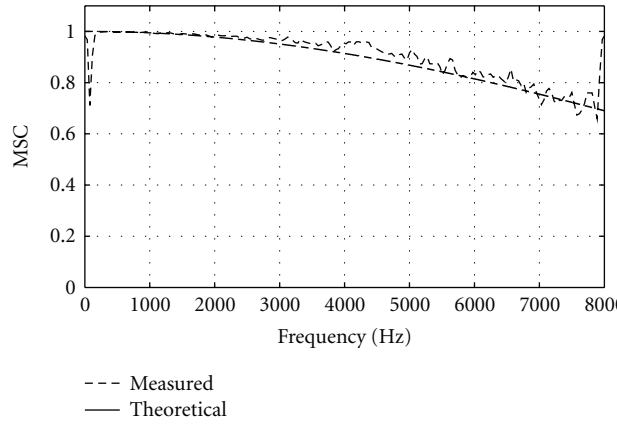

Figure 4: Example of MSC: same face.

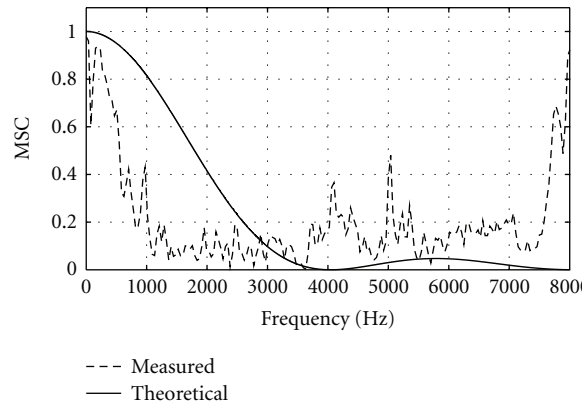

Figure 5: Example of MSC: different faces.

the position of loudspeakers and the DHMA are shown in Figure 3.

Figure 4 shows the MSC between two microphones on the same face of the DHMA (microphone distance $d_{\text{mic}} = 7$ (mm)). A dashed line shows an experimental characteristic from the measured impulse response, and a solid line shows a theoretical characteristic which is formulated using the sinc function. Figure 5 shows MSC on different faces of the DHMA (microphone distance $d_{\text{mic}} = 42$ (mm)). According to Figure 4, when both microphones are on the same face, the experimental MSC is equivalent to the theoretical MSC. On the other hand, when the microphones are on different faces, the experimental MSC is different from the theoretical value. This is due to the shape of the DHMA and the spatial aliasing resulting from the large distance between microphones.

3.1.3. Frequency Band Selection Using Magnitude Squared Coherence. Figure 6 shows the averaged experimental MSC (AEMSC) on the same face of the microphone and on the different faces, respectively. In this paper, we propose a band selection method based on the AEMSC. Small MSC values

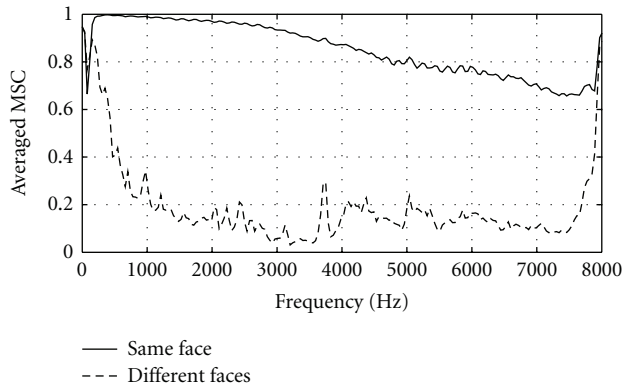

Figure 6: Averaged experimental MSC (AEMSC).

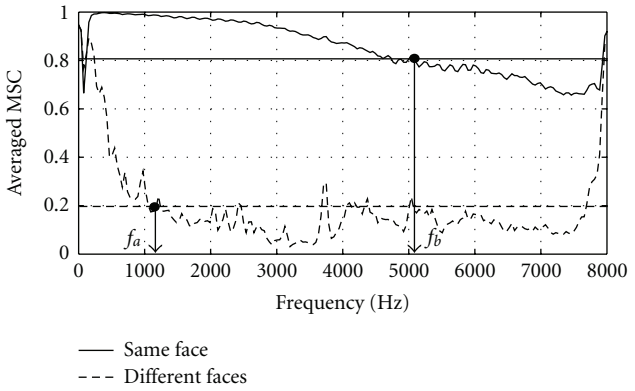

Figure 7: Region of the band selection.

correspond to large phase differences, and thus selection should occur mainly in regions with small MSC values. In order to prevent a bias in which bands are selected, we consider three frequency regions based on MSC, which leads to the selection of a large number of bands with small MSC values. Figure 7 shows frequency regions B_1, B_2, B_3. Boundary frequencies correspond to a mean of the AEMSC shown in Figure 7. f_b is determined by the crosspoint between the AEMSC for the same face and its mean, and f_a is determined in the same manner as the AEMSC of the other face. The values of f_a and f_b in this paper are 1016 (Hz) and 5040 (Hz), respectively.

In Section 4, the number of bands in each frequency region is used as a parameter to evaluate the proposed method. For example, the bands 1/5, 1/3 and 1/2 are selected in regions B_1, B_2, and B_3, respectively, and the total number of frequency bands is 77, reduced from 200 bands, which correspond to the target frequency region 0–8 (kHz), with a sample of FFT 1024.

3.2. Estimating the Number of Source Signals and Dimension Reduction. The source signals are transferred from their locations to the DHMA, and a mixing matrix is described in frequency domain $\mathbf{H}(k)$ which is an $N_S \times N_M$ matrix where N_S and N_M are the number of sources and microphones, respectively. The observed signal is represented by $\mathbf{X}(k,l) = \mathbf{H}(k)\mathbf{S}(k,l)$ for frequency band index k and frame index l. $\mathbf{X}(k,l)$ and $\mathbf{S}(k,l)$ represent the observed and source signal vectors, respectively.

The BSS method using a DHMA is conducted under an overdetermined condition, because a DHMA can include up to 160 microphones. FDICA assumes that the number of source signals is equal to the number of the microphones, and thus an estimation of the number of source signals and a reduction in the number of observed signals are needed to perform FDICA. In addition, the reduced dimensions must exceed the number of actual sound sources, because not only actual sound source signals but also reflection waves are used in FDICA. The spatial covariance matrix is calculated by an expectation of the observed signal $\mathbf{X}(k,l)$, and it is decomposed into eigenvalues. The number of virtual sound sources N_Q that include direct sound sources and early-reflected sources is estimated from eigenvalue diagonal matrix $\Lambda(k)$ as follows:

$$\Lambda(k) = \mathrm{diag}[\lambda_1, \ldots, \lambda_{N_M}], \tag{3}$$

where diag denotes the operator at which all of matrix elements except diagonal elements are set to zero and λ_i is the ith eigenvalue. This corresponds to the estimation of the number of sound sources with directivity, because such sound sources are spatially correlated. Normalization, whose summation of the eigenvalues is 1, is calculated at each frequency as follows:

$$\lambda_m \longleftarrow \frac{\lambda_m}{\sum_{i=1}^{N_M} \lambda_i}, \quad m = 1, \ldots, N_M. \tag{4}$$

The threshold for the normalized eigenvalues evaluates the number of virtual sound sources in each frequency band, and the maximum estimated value in all frequencies is assumed to be N_Q, because the number of estimated sound sources is different in each frequency band. Following the estimation of the number of virtual sources, eigenvectors, which are estimated with eigenvalues, are employed to reduce the number of observed signals using the subspace method. A more detailed description of the process explained previously is given in [8].

3.3. Frequency Domain Independent Component Analysis. Following estimation of the number of virtual sound sources N_Q and band selection based on the MSC, the observed signal is reduced to the N_Q dimension using the subspace method, and FDICA estimates separation matrix $\mathbf{U}(k)$, but only in the selected bands, via an update rule with iteration [23] based on the principle presented in [24]. Separation matrix $\mathbf{W}(k)$ for actual separation is obtained through a combination of a subspace matrix and separation matrix $\mathbf{U}(k)$. The separated signal $\hat{\mathbf{Y}}(k,l)$ is obtained as follows:

$$\hat{\mathbf{Y}}(k,l) = \mathbf{W}(k)\mathbf{X}(k,l). \tag{5}$$

The scaling problem is solved using the projection method [20]. Next, the number of dominant source signals in each frequency band $N_{K(k)}$ is calculated from the separated signals via a threshold operation.

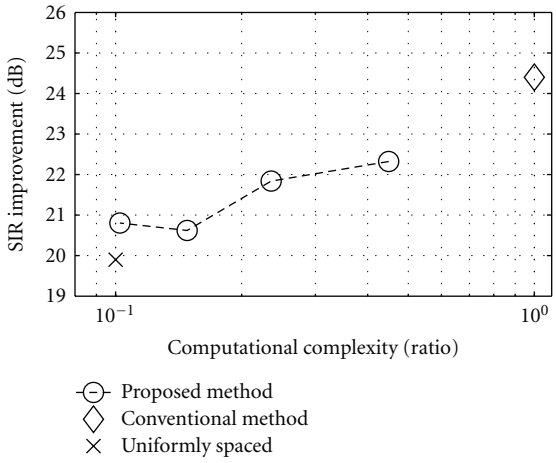

Figure 8: SIR improvement.

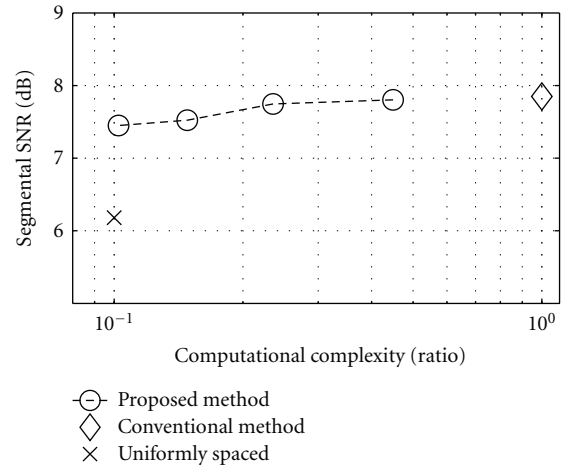

Figure 9: Segmental SNR.

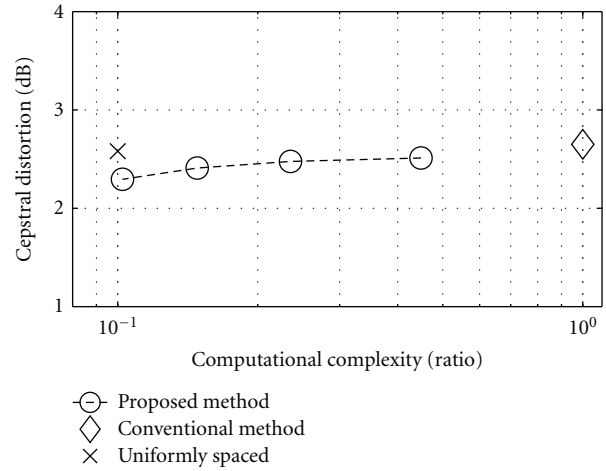

Figure 10: Cepstral distortion.

3.4. Permutation Solver Based on Acoustic Transfer Function Clustering. Solution of the permutation problem affects separation performance significantly, and Ogasawara et al. have proposed a method which combines the acoustic pressure distribution and the relative phase distance [8]. As mentioned in Section 1, the correlation of the separated signals is one of the most common method to solve the permutation problem; however, this method can be applied to narrow band signals that each frequency channel has a very close frequency. In addition, for the acoustic field encoding, not only the direct sound sources but also the early-reflected sources are considered as the source signals, and this implies that similar time differences from different locations are included in the separated signals. The correlation of these signals is very high, and it is difficult to solve the permutation based on the power envelopes of the separated signals. The method proposed in [8] significantly improves our ability to solve the permutation problem for a DHMA; however it uses transfer function clustering, which leads to a high level of computational complexity. In this section, we explain the method proposed by Ogasawara et al. briefly, and estimation of computational complexity is described when the band selection method is applied.

3.4.1. Acoustic Transfer Function Clustering. The Moore-Penrose pseudoinverse of the separation matrix corresponds to the mixing matrix; in other words, it corresponds to the transfer functions between the source signals and the observed signals. The transfer function is estimated as the qth column vector $\mathbf{w}_q^+(k)$ of pseudoinverse $\mathbf{W}^+(k)$.

Two similarities, acoustic pressure distribution $p(\mathbf{w}_q^+(k))$ for amplitude similarity D_a and relative phase distance $\phi(\mathbf{w}_q^+(k))$ for phase similarity D_p, are considered. As mentioned in Section 2, a DHMA has an acoustic pressure difference between each face, and the acoustic pressure distribution shows the characteristics of the directions of the sources. The acoustic pressure distribution is formulated as follows:

$$p\left(\mathbf{w}_q^+(k)\right) = \left[\frac{1}{|M(1)|}\sum_{m\in M(1)}\left|w_{q,m}^+(k)\right|,\ldots,\frac{1}{|M(10)|}\times\sum_{m\in M(10)}\left|w_{q,m}^+(k)\right|\right],$$
$$p\left(\mathbf{w}_q^+(k)\right) \longleftarrow \frac{p\left(\mathbf{w}_q^+(k)\right)}{\sum_q p\left(\mathbf{w}_q^+(k)\right)}, \tag{6}$$

where $M(\mu)$ represents the set of microphones on the μth face and $w_{q,m}^+(k)$ is the transfer function between source q and the mth microphone on face μ. Amplitude similarity D_a is formulated using acoustic pressure distribution $p(\mathbf{w}_q^+(k))$ with the νth centroid $\mathbf{c}_\nu$ as follows:

$$D_a\left(\mathbf{w}_q^+(k), c_\nu\right) = \left\|p\left(\mathbf{w}_q^+(k)\right) - p(\mathbf{c}_\nu)\right\|^2. \tag{7}$$

Phase similarity is calculated using the normalized phase difference $\phi(\mathbf{w}_q^+(k))$ between two microphones, calculated from normalized time difference $\tau_{q,m}(k)$ as follows:

$$\phi\left(\mathbf{w}_q^+(k)\right) = \left[\exp\left(j\tau_{q,1}(k)\right),\ldots,\exp\left(j\tau_{q,N_M}(k)\right)\right],$$
$$\tau_{q,m} = \gamma\frac{\angle w_{q,m}^+(k)}{F_s k/N_F}, \tag{8}$$

where γ is a normalization constant and $\angle$ means the operator which obtain an argument of a complex number. Consequently, the phase similarity D_p is formulated as follows:

$$D_p\left(\mathbf{w}_q^+(k),\mathbf{c}_\nu\right) = \sum_{l=1}^{10}\left|\sum_{m\in M(\mu)}\phi\left(w_{q,m}^+(k)\right)^*\phi(c_{\nu,m})\right|, \tag{9}$$

where $(\cdot)^*$ represents the complex conjugate and $c_{\nu,m}$ represents the νth centroid of the mth microphone. After the normalization of D_a and D_p via their mean and variance, respectively, the combined similarity $\mathcal{J}(\cdot)$ on each frequency band for hierarchical clustering is calculated as follows:

$$\mathcal{J}\left(\mathbf{w}_{q1}^+(k_\alpha),\mathbf{w}_{q2}^+\left(k_\beta\right)\right)$$
$$= \left\{a(k_\alpha)+a\left(k_\beta\right)\right\}D_a + \left\{b(k_\alpha)+b\left(k_\beta\right)\right\}D_p, \tag{10}$$

$$a(k) = \left\{\frac{k/I}{N_F/2}\right\}^\rho, \qquad b(k) = 1-a(k), \tag{11}$$

where I is a parameter to adjust the phase similarity weighting and ρ is a parameter to adjust the boundary frequency between the amplitude and phase similarities.

The similarity described in (10) is calculated with hierarchical clustering as the permutation correction of the BSS method for a DHMA. As mentioned earlier, high computational complexity is one of the drawbacks to performing hierarchical clustering for DHMA BSS. The band selection method mentioned in Section 3.1.3 contributes to reducing computational complexity, not only during estimation of the separation matrix but also during hierarchical clustering for the permutation solution. After permutation correction, the separation matrix on the nonselected bands is linearly interpolated from the separation matrix in the neighboring frequency bands, which has already been estimated. An interpolation of the mixing matrices, which could be estimated from the inverse matrix of the separation matrix, should be appropriate. However, the inverse matrix, which consumes order $\mathcal{O}(n^3)$ complexity (n means a n-by-n square matrix) [25], leads to additional computational complexity. Since the aim of this paper is the reduction of computational complexity, we therefore determine the interpolation of the separation matrix in the neighboring frequency bands as the separation matrix for the non-selected bands.

3.4.2. Estimation of Computational Complexity. In this section, computational complexity is estimated from order $\mathcal{O}(n)$. In our previous papers [17–19], the number of operations, which is calculated using the number of floating operations of multiplication and addition, is used to evaluate computational complexity. From a practical viewpoint, this criterion is valuable for estimating the possibility of implementation with embedded devices, and in addition it is useful to estimate system requirements for manufacturers. On the other hand, especially when there are a large number of microphones, it is difficult to estimate complexity precisely using this method. The scaling ambiguity must be solved when source separation is performed with ICA; the projection method based on the Moore-Penrose pseudoinverse is applied in this paper. The pseudoinverse is obtained using the singular value decomposition (SVD). In our previous research, the assumed number of microphones was only two, the same as the number of source signals; therefore, estimation of the complexity of the projection method was obtained using inverse matrix. However, a DHMA has many microphones, and the source separation algorithm is based on the subspace method and FDICA. Mathematically, this is an overdetermined problem, as the inverse matrix cannot be calculated and the pseudoinverse must be appropriate. In addition, the proposed method uses hierarchical clustering to solve the permutation problem. Consequently, in this paper, $\mathcal{O}(n)$ is used to estimate computational complexity instead of counting the floating operations.

Table 2: Computational complexity.

Method	Complexity
STFT (forward and inverse)	$2\mathcal{O}(N_M N_L N_F \log_2 N_F)$
Covariance matrix	$\mathcal{O}(N_M^2 N_L N_B)$
Eigenvalue decomposition	$\mathcal{O}(4N_M^3/3 + 3N_M^3 N_B)$
Subspace method	$\mathcal{O}(N_Q N_M N_L N_B)$
Separation matrix	$\mathcal{O}(N_Q^3 N_I N_B)$
Projection method	$\mathcal{O}(4N_M^3/3 + 3N_M^3 N_B)$
Hierarchical clustering	$\mathcal{O}(\{N_M N_K(k) N_B\}^2)$

In general, eigenvalue decomposition (EVD) is solved by the Householder method (HHM) and the implicit shifted QL method (ISQL) [25]. In [25], the numerical calculation method for SVD is also introduced and consists of HHM and ISQL. Therefore, in our estimation, the computational complexity of the EVD and pseudoinverse can be estimated by HHM and ISQL. The computational complexity of HHM and ISQL is introduced as $\mathcal{O}(4n^3/3)$ and $\mathcal{O}(3n^3)$ respectively, in [25]. For hierarchical clustering, an efficient algorithm is introduced in [26], and the computational complexity is $\mathcal{O}(n^2)$.

Estimated computational complexity is shown in Table 2 where

(i) N_M: number of microphones,

(ii) N_L: number of frames,

(iii) N_F: FFT size,

(iv) N_B: number of bands which correspond to the Nyquist frequency of speech signals,

(v) N_Q: number of subspaces,

(vi) $N_{K(k)}$: averaged number of separated source signals in each frequency band,

(vii) N_I: number of iterations.

An example, estimation is shown in Table 3 which is calculated from the order of $\mathcal{O}(n)$ in Table 2 using concrete numbers. N_M is 60 as described in Section 2.1; six microphones are installed on each face, and the DHMA has ten faces for microphone arrays. In our experiments described in Section 4, some numbers are given in Table 1; N_L is 625, which corresponds to the length of the source signals as 4 (sec) and N_F is 1024. N_B is 200 in the case of the full band (0–8 (kHz)) and 80 in the case of the proposed method (e.g., estimation). We assume that N_Q and $N_{K(k)}$ are 25 and 15, respectively, via preliminary experiment. To simplify estimation, $N_{K(k)}$ is not varied in every frequency band. Actual iteration is terminated by a convergence test; however to simplify the estimation, we consider that each iteration N_I is 200 times. As shown in Table 3, total estimated complexity has the same order of complexity as hierarchical clustering. The proposed method results in an 84% reduction in computational complexity as a result of reducing number of frequency bands from 200 to 80; however, the number of frequency bands has only been reduced 60%. Hierarchical clustering is based on a bottom-up algorithm, and this is the reason for the large reduction in complexity. Hierarchical clustering needs to calculate similarities between all of the transfer functions, and thus computational complexity depends on the number of initial elements. The reduced complexity of the hierarchical clustering process results in a power of two reduction in complexity compared to reduction of the number of frequency bands.

3.4.3. Comparison to SOS-ICA. In Section 1, the joint diagonalization using SOS, TRINICON [15, 16] is one of the most common methods in this field, is introduced, and it has an advantage to prevent the complete decoupling of the bandwise frequency components; which means that the permutation solver is not required. This is a general way that in each iteration the update equation involves the backward and forward DFT transformation, equation (53) in [16]. This equation also includes a restricting operation for the time sequence of the separation filter, and this is due to the prevention of the complete decoupling of the bandwise frequency components by considering the linear convolution. Using the DFT transformations in the iterative update corresponds to using all of the frequency bands to estimate the separation matrix for linear convolution, even though the speech signal is limited up to 8 (kHz). In contrast, the proposed method allows circular convolution for the efficiency of the computational complexity. Equation (67) in [16] is the simplest update rule of TRINICON with some approximations. Under the same configuration of the estimation in Section 3.4.2, additional parameters for TRINICON are considered that FFT size is four times N_F, the number of bands for the separation matrix is $4N_F/2 + 1$, the number of blocks for the joint diagonalization is $N_J = 20$, and the number of iterations $N_{I'} = 40$. The computational complexity of the update rule equation (67) in [16] is $\mathcal{O}(\{3N_Q^3(4N_F/2+1)N_J\}N_{I'})$, and it is predominant complexity of TRINICON as shown in Table 3. No permutation solver contributes to reduce the total computational complexity; however considering the separation filter in the time domain increases the computational complexity in the iterative update. Note that the author in [15] mentioned necessity of the permutation solver under the independent bandwise separation matrix estimation, and in addition they mentioned in [16] that the applied approximation to the update equation disturbs the perfect permutation correction; in other words, this means that the separation performance is degraded by these approximations. The proposed band selection method is evaluated in the following section, and this method contributes to the reduction of the computational complexity for the permutation solver in which the shape of DHMA is significantly reflected.

4. Experimental Evaluation

In this paper, we compare our proposed method with the conventional method because it is important to evaluate feasibility of the proposed method, which might represent a balance between separation performance and computational complexity.

4.1. Experimental Conditions and Evaluation Measures. The experimental conditions are the same as in Figure 3 and Table 1. The proposed band selection method might cause a degradation in separation performance as a result of the limited number of frequency bands. Separation performance is evaluated by improvement in the signal-to-interference ratio ($\mathrm{SIR}_{\mathrm{imp}}$):

$$\begin{aligned}
\mathrm{SIR}_{\mathrm{imp}}^{(\xi)} &= \mathrm{SIR}_{\mathrm{out}}^{(\xi)} - \mathrm{SIR}_{\mathrm{in}}^{(\xi)},\\
\mathrm{SIR}_{\mathrm{in}}^{(\xi)} &= 10\log_{10}\left[\frac{\sum_t \{x_{J\xi}(t)\}^2}{\sum_t \sum_{(J'\neq\xi)} \{x_{JJ'}(t)\}^2}\right],\\
\mathrm{SIR}_{\mathrm{out}}^{(\xi)} &= 10\log_{10}\left[\frac{\sum_t \{y_{\xi\xi}(t)\}^2}{\sum_t \sum_{(J'\neq\xi)} \{y_{\xi J'}(t)\}^2}\right],
\end{aligned} \tag{12}$$

where $x_{J\xi}(t)$ represents the observed source signal ξ on the Jth microphone and $y_{\xi\xi}(t)$ represents the output signal which corresponds to source signal ξ. $\mathrm{SIR}_{\mathrm{imp}}^{(\xi)}$ is averaged over all of the source signals. The other important factor is the quality of the separated sound. Segmental signal-to-noise ratio ($\mathrm{SNR}_{\mathrm{seg}}$) is a very common measure for evaluating noise suppression. 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 [27]. The proposed method needs to obtain the separation matrix in the nonselected bands through interpolation from the estimated separation matrix in the selected bands, and the degradation of the separated signals can be estimated. Cepstral distortion (CD) [28] is another measure of the degree of distortion via the

Table 3: Estimated computational complexity.

Method	Complexity (conventional)	Complexity (proposed)	Ratio (reduction (%))	Complexity (TRINICON)	Ratio (reduction (%))
STFT (forward and inverse)	7.68*E*8	7.68*E*8	1.0 (0 [%])	3.69*E*9	4.8
Covariance matrix	1.80*E*9	7.20*E*8	0.4 (60 [%])	4.61*E*10	25.6
Eigenvalue decomposition	7.49*E*8	3.00*E*8	0.4 (60 [%])	7.67*E*9	10.3
Subspace method	7.50*E*8	3.00*E*8	0.4 (60 [%])	4.01*E*10	53.5
Separation matrix	2.50*E*9	1.00*E*9	0.4 (60 [%])	3.13*E*11	156.6
Projection method	7.49*E*8	3.00*E*8	0.4 (60 [%])	7.67*E*9	10.3
Hierarchical clustering	5.18*E*11	8.29*E*10	0.16 (84 [%])	—	—
Total	5.26*E*11	8.63*E*10	0.16 (84 [%])	4.19*E*11	0.80 (20 [%])

cepstrum domain, and this can evaluate distortion of a spectral envelope. $\mathrm{SNR}_{\mathrm{seg}}^{(\xi)}$ is formulated as follows:

$$\mathrm{SNR}_{\mathrm{seg}}^{(\xi)} = 10\log_{10}\left[\frac{1}{N_{l_s}}\sum_{l_s}\frac{\sum_t \{x_{J\xi}(t,l_s)\}^2}{\sum_t \{x_{J\xi}(t,l_s) - y_{\xi\xi}(t,l_s)\}^2}\right], \quad (13)$$

where l_s is a frame number and N_{l_s} is the number of frames used to evaluate $\mathrm{SNR}_{\mathrm{seg}}$. We calculate CD from speech components, and it is defined as follows:

$$\mathrm{CD}^{(\xi)} = \frac{20}{N_{l_c}\ln 10}\sum_t \sqrt{\sum_{\kappa=1}^{L} 2\left\{C_{x_{J\xi}(t,l_c)}(\kappa,l_c) - C_{y_{\xi\xi}(t,l_c)}(\kappa,l_c)\right\}^2}, \quad (14)$$

where l_c is a frame number, κ is the index of the cepstrum coefficient and N_{l_c} is the number of frames for CD. $C_{(\cdot)}(\cdot)$ is the cepstrum coefficient, and L is the number of dimensions of the cepstrum used in the evaluation; we set $L = 20$. $\mathrm{SNR}_{\mathrm{seg}}^{(\xi)}$ and $\mathrm{CD}^{(\xi)}$ are also averaged over all of the source signals.

4.2. Experimental Results and Discussion. Results for each performance criterion are shown in Table 4. The first column shows the ratio of selected bands in each frequency region, for example, "(1/5,1/3,1/2)" means one-fifth for lowest frequency region B_1, one-third for middle frequency region B_2, and one-half for highest frequency region B_3. Each region is divided into 1016 (Hz) and 5040 (Hz), respectively. The number of selected bands means the total number of selected frequency bands. Computational complexity is the ratio compared to using the full band, and the method of estimation of computational complexity is described in Section 3.4.2. Therefore, computational complexity is a feature of the number of selected bands in this paper. Figures 8, 9, and 10 show the performance of the proposed method. The x-axis of each figure represents the ratio of computational complexity, and the y-axis represents the evaluation criterion.

The configuration "(1,1,1)" in Table 4 corresponds to the conventional method proposed by Ogasawara et al. [8], and these results are shown at ratio 1.0 (10^0) point on the x-axis of Figures 8–10. The $\mathrm{SIR}_{\mathrm{imp}}$ shows the contribution to separation performance of a large number of selected bands. This result is assumed before the experiment. $\mathrm{SIR}_{\mathrm{imp}}$ degrades as the number of selected bands decreases, and $\mathrm{SNR}_{\mathrm{seg}}$ and CD show different characteristics of this degradation. They show that only a small degradation occurs under 1 (dB); in other words, $\mathrm{SNR}_{\mathrm{seg}}$ and CD show almost equivalent performance using limited number of frequency bands. The proposed method is focused on a *nonuniformly* spaced selection of frequency bands. In addition, the aim of this nonuniformly spaced selection is to utilize the characteristics of the dodecahedral shape of the microphone array. When 64 bands were selected, $\mathrm{SIR}_{\mathrm{imp}}$ of the proposed method shows almost 1 (dB) higher than in the case of uniformly spaced band selection. In particular, $\mathrm{SNR}_{\mathrm{seg}}$ shows significant improvement. On the other hand, CD shows almost equivalent distortion to uniformly spaced band selection. The magnitude squared coherence, theoretically and experimentally described in Section 3.1, reflects this characteristic, which was confirmed by the separation performance results of the experiments; lower values of MSC contribute to separation performance, particularly in the high frequency region that the number of selected bands is larger than the others. Even though computational complexity is reduced by around 90% as compared to using the full band, the acoustic characteristics of the proposed method contribute to improving not only the separation performance but also the quality of the separated sound.

The proposed band selection method for the frequency domain BSS method is advantageous to achieve the trade-off between the separation performance with the significantly low degradation of the sound quality and the computational complexity. On the other hand, the degradation of the separation performance shows a disadvantage of the proposed method compared to the conventional method which uses all of the frequency bands. The joint diagonalization using SOS such as TRINICON shows less computational complexity, around 20% reduction in Section 3.4.3, than the BSS method using the permutation solver. When the number of microphones is small or the voice terminals have a high computation power, it is easy to perform TRINICON. However, the constraint of the linear convolution does not allow to reduce further computational complexity. The voice terminals are generally implemented in the embedded systems as mentioned in Section 1; the lower computational complexity is required to perform the BSS method.

Table 4: Experimental results.

(B_1, B_2, B_3)	Number of selected bands	Computational complexity	SIR improvement (dB)	Segmental SNR (dB)	Cepstral distortion (dB)
(1,1,1) (conventional method)	200	1.0	24.4	7.85	2.65
(1/3,1/2,1)	134	0.45	22.3	7.80	2.51
(1/3,1/2,1/2)	97	0.24	21.8	7.74	2.48
(1/5,1/3,1/2)	77	0.15	20.6	7.52	2.30
(1/5,1/3,1/3)	64	0.1	20.8	7.45	2.30
Uniformly spaced	64	0.1	19.9	6.18	2.58

The proposed band selection method can reduce further computational complexity, and it is easy to achieve over 50% reduction.

5. Conclusion

A blind source separation method with efficient computational complexity for use with an agglomerative DHMA which can encode the acoustic field is proposed. The proposed band selection method uses the spatial characteristics of a DHMA, and a preliminary experiment on magnitude squared coherence describes the criterion of the band selection process. The proposed method uses nonuniformly spaced selection of frequency bands, which contributes to improved separation performance versus uniformly spaced band selection in experiments. Estimated computational complexity was greatly reduced during hierarchical clustering, and thus the total reduction in complexity achieved exceeds the reduction due to limitation of the number of the frequency bands. For example, if the number of frequency bands is reduced by 60%, the total reduction in computational complexity achieved is 84%. Experimental results of the proposed method show practical separation performance compared to the conventional method. In addition, equivalent signal distortion compared with the conventional method is maintained. Band selection is simply based on the spatial characteristics of the DHMA, and therefore any state-of-the-art frequency domain BSS method with the permutation solver can be applied to the proposed method without a loss of generality. However, the method proposed in this paper is only considered for use with an off-line algorithm; therefore future work includes developing an on-line causal method to reduce computational complexity.

References

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

[2] S. Ikeda and N. Murata, "An approach to blind source separation of speech signals," in *proceedings of the International Conference on Artificial Neural Networks (ICANN '98)*, pp. 761–766, September 1998.

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

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

[5] W. Wang, J. A. Chambers, and S. Sanei, "A novel hybrid approach to the permutation problem of frequency domain blind source separation," in *Proceedings of the Independent Component Analysis and Blind Signal Separation (ICA '04)*, pp. 532–539, September 2004.

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

[7] K. Niwa, T. Nishino, and K. Takeda, "Selective listening point audio based on blind signal separation and stereophonic technology," in *Proceedings of the International Workshop on the Principles and Applications of Spatial Hearing (IWPASH' 09)*, p. I10, November 2009.

[8] M. Ogasawara, T. Nishino, and K. Takeda, "A small dodecahedral microphone array for blind source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 229–232, March 2010.

[9] K. Osako, Y. Mori, Y. Takahashi, H. Saruwatari, and K. Shikano, "Fast convergence blind source separation using frequency subband interpolation by null beamforming," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E91-A, no. 6, pp. 1357–1361, 2008.

[10] A. Tanaka, H. Imai, and M. Miyakoshi, "Theoretical foundations of second-order-statistics-based blind source separation for non-stationary sources," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, pp. III600–III603, May 2006.

[11] K. Tachibana, H. Saruwatari, Y. Mori, S. Miyabe, K. Shikano, and A. Tanaka, "Efficient blind source separation combining closed-form second-order ICA and nonclosed-form higher-order ICA," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pp. I45–I48, April 2007.

[12] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[13] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Proceedings of the Independent Component Analysis and Blind Signal Separation (ICA '06)*, pp. 601–608, 2006.

[14] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[15] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of a class of blind source separation algorithms for convolutive mixtures," in *Proceedings of the Independent Component Analysis and Blind Signal Separation (ICA '03)*, pp. 945–950, April 2003.

[16] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2005.

[17] K. Kondo, Y. Takahashi, S. Hashimoto, H. Saruwatari, T. Nishino, and K. Takeda, "Improved method of blind speech separation with low computational complexity," *Journal of Advances in Acoustics and Vibrations*, vol. 2011, Article ID 765429, 10 pages, 2011.

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

[19] 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 European Signal Processing Conference (EUSIPCO '11)*, pp. 2319–2323, August 2011.

[20] S. Ikeda and N. Murata, "Method of ica in time-frequency domain," in *Proceedings of the Independent Component Analysis and Blind Signal Separation (ICA '99)*, pp. 365–371, January 1999.

[21] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, pp. 1167–1170, April 1997.

[22] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field oherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[23] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in *Procedings of the International Workshop on Independent Component a Analysis and Blind Signal Separation*, pp. 371–376, January 1999.

[24] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[25] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK, 1998.

[26] C. F. Olson, "Parallel algorithms for hierarchical clustering," *Parallel Computing*, vol. 21, no. 8, pp. 1313–1325, 1995.

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

[28] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Upper Saddle River, NJ, USA, 1993.