Combined Approach of Array Processing and Independent Component Analysis for Blind Separation of Acoustic Signals

Futoshi Asano, *Member, IEEE*, Shiro Ikeda, *Member, IEEE*, Michiaki Ogawa, Hideki Asoh, *Member, IEEE*, and Nobuhiko Kitawaki, *Member, IEEE*

***Abstract*—In this paper, two array signal processing techniques are combined with independent component analysis (ICA) to enhance the performance of blind separation of acoustic signals in a reflective environment. The first technique is the subspace method which reduces the effect of room reflection when the system is used in a room. Room reflection is one of the biggest problems in blind source separation (BSS) in acoustic environments. The second technique is a method of solving permutation. For employing the subspace method, ICA must be used in the frequency domain, and precise permutation is necessary for all frequencies. In this method, a physical property of the mixing matrix, i.e., the coherency in adjacent frequencies, is utilized to solve the permutation. The experiments in a meeting room showed that the subspace method improved the rate of automatic speech recognition from 50% to 68% and that the method of solving permutation achieves performance that closely approaches that of the correct permutation, differing by only 4% in recognition rate.**

***Index Terms*—Array signal processing, blind signal separation, independent component analysis, permutation, room reflection.**

I. Introduction

It IS indispensable to separate acoustic signals and to pick up signals of interest for applications such as automatic speech recognition (ASR) when they are used in a real environment. The framework of blind source separation (BSS) based on independent component analysis (ICA) is attractive since it can be used to separate multiple signals without any previous knowledge of the sound sources and sound environment such as the configuration of microphones (e.g., [1]) that is necessary in conventional microphone-array signal processing (e.g., [2]). However, when applying BSS to an acoustical mixture problem such as a number of people talking in a room, the performance of the BSS system is greatly reduced by the effect of the room reflections/reverberations and ambient noise [3].

One method of reducing the effect of room reflections is to employ directional microphones in the BSS framework [4]. In this method, however, the system cannot track the movement of the sound sources. Another way to reduce the effect of room reflections is to employ an acoustic beamformer such as delay-and-sum (DS) beamformer (e.g., [5]). However, for designing the beamformer, an array response database consisting of transfer functions from possible source locations to microphones or, at least, the configuration of the microphone-array is required. This information is not available in the BSS framework.

The authors previously proposed an alternative approach, the subspace method, for reducing the effect of room reflections and ambient noise [6]. In this method, room reflections are separated from direct components in the eigenvalue domain of the spatial correlation matrix based on the spatial extent of the acoustic signals. Then, the eigenvectors corresponding to the eigenvalues of the direct components are used as a filter which selects the subspace in which the direct components lie and discards the subspace filled with the energy of reflections. As described in this paper, the subspace method works as a self-organizing beamformer focusing on the target sources and does not require any previous knowledge of the array or sound field. Therefore, the subspace method can be used in the framework of BSS. This is understood from the fact that the subspace method is a special case of principal component analysis (PCA) with $M \gg D$, where M and D denote the number of nodes (channels) of the input and the output of PCA, respectively [7]. PCA is known as a method of unsupervised learning which does not require any previous knowledge. In this paper, a combined approach of the subspace method and ICA is proposed. In this method, the subspace method is utilized as a pre-processor of ICA which reduces room reflections in advance, the remaining direct sounds then being separated by ICA.

For combining the subspace method with ICA, the frequency-domain ICA [8], [9] must be employed, since the subspace method works in the frequency-domain. The biggest obstacle in the frequency-domain ICA is the permutation and scaling problem. In the frequency-domain ICA, the input signal is first transformed into the frequency domain by the Fourier transform. By using this transformation, a convolutive mixture problem is reduced to a complex but instantaneous mixture problem. This instantaneous mixture problem is then solved at each frequency independently. In usual instantaneous ICA, arbitrary permutation and scaling of the output is allowed. However, in the frequency-domain processing for a convolutive mixture such as that employed in this paper, different permutations at different frequencies lead to re-mixing of signals in

Manuscript received January 20, 2002; revised August 13, 2002. This work was supported by the IEEE. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hsaio-Chuan Wang.

F. Asano and H. Asoh are with the National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan (e-mail: f.asano@aist.go.jp; h.asoh@aist.go.jp).

S. Ikeda is with Kyushu Institute of Technology, Japan (e-mail: shiro@brain.kyutech.ac.jp).

M. Ogawa and N. Kitawaki are with Tsukuba University, Tsukuba, Japan (e-mail: kitawaki@is.tsukuba.ac.jp).

Digital Object Identifier 10.1109/TSA.2003.809191

1063-6676/03$17.00 © 2003 IEEE

the final output. Also, different scaling at different frequencies leads to distortion of the frequency spectrum of the output signal.

For the scaling problem, the method proposed by [9], in which the separated output is filtered by the inverse of the separation filter, shows good performance. On the other hand, for the permutation problem, a method using the correlation between the spectral envelope at different frequencies (denoted as Inter-frequency Spectral Envelope Correlation (IFSEC), hereafter in this paper. See Appendix I for a brief explanation.) has been proposed [9], but has been reported to sometimes fail when the input signals have similar envelopes [3]. In this paper, a new approach for solving the permutation problem is proposed. This method utilizes the *coherency* of the mixing matrices in several adjacent frequencies and is thus denoted as Inter-Frequency Coherency (IFC) in this paper.

This paper is organized as follows: In Section II, the model of the sound environment treated in this paper is described. In Section III, an outline of the proposed BSS system is presented. Moreover, some portions of the system which were proposed in the previous studies but which are necessary for understanding the following sections are briefly described. In Section IV, the subspace method for reducing room reflections is detailed. In Section V, a new method for solving the permutation is proposed. In Section VI, results of experiments using real data to evaluate the proposed system are reported.

II. Model of Signal

Let us consider the case when there are D sound sources in the environment. By observing this sound field with M microphones and taking the short-term Fourier transform (STFT) of the microphone inputs, we obtain the input vector

$$\mathbf{x}(\omega, t) = [X_1(\omega, t), \ldots, X_M(\omega, t)]^T. \tag{1}$$

Here, $X_m(\omega, t)$ is STFT of the input signal in the tth time frame at the mth microphone. The symbol $\cdot^T$ denotes the transpose. In this paper, the input signal is assumed to be modeled as

$$\mathbf{x}(\omega, t) = \mathbf{A}(\omega)\mathbf{s}(\omega, t) + \mathbf{n}(\omega, t). \tag{2}$$

Matrix $\mathbf{A}(\omega)$ is termed the mixing matrix, its (m, n) element, $A_{m,n}(\omega)$, being the transfer function from the nth source to the mth microphone as

$$A_{m,n}(\omega) = H_{m,n}(\omega)e^{-j\omega\tau_{m,n}}. \tag{3}$$

The symbol $H_{m,n}(\omega)$ is the magnitude of the transfer function. The symbol $\tau_{m,n}$ denotes the propagation time from the nth source to the mth microphone. Vector $\mathbf{s}(\omega, t)$ consists of the source spectra as $\mathbf{s} = [S_1(\omega, t), \ldots, S_D(\omega, t)]^T$, where $S_n(\omega, t)$ denotes the spectrum of the nth source. The first term, $\mathbf{A}(\omega)\mathbf{s}(\omega, t)$, expresses the directional components in $\mathbf{x}(\omega, t)$. On the other hand, the second term, $\mathbf{n}(\omega, t)$, is a mixture of less-directional components, which includes room reflections and ambient noise.

TABLE I
Outline of the Entire BSS System

	Operation	Filter Matrix	Section in this paper
1	Taking STFT of the input		II
2	Applying the subspace method	$\mathbf{W}$	IV
3	Solving the instantaneous mixture by ICA	$\mathbf{U}$	III.B
4	Solving permutation	$\mathbf{P}$	V
5	Solving scaling problem	$\tilde{\mathbf{B}}_m^+$	III.C
6	Filtering	$\mathbf{F}$	III.D

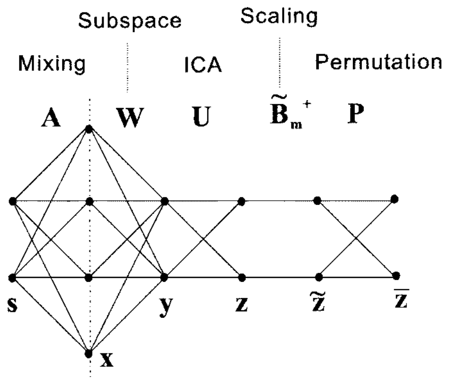


Fig. 1. Proposed BSS filter network.

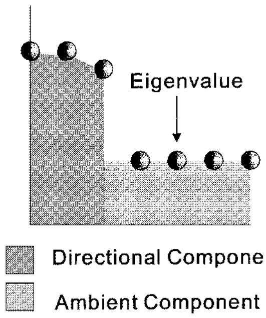


Fig. 2. Typical eigenvalue distribution.

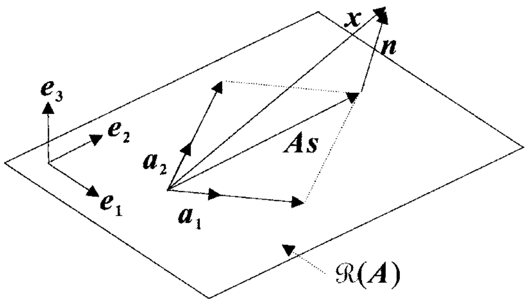
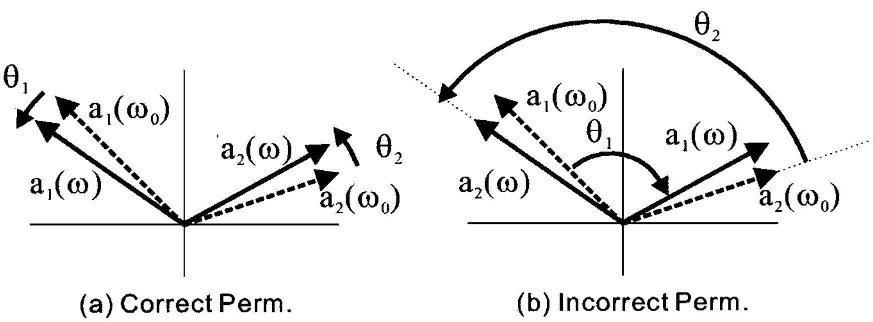


Fig. 3. Relation of vectors.

III. BSS System

A. Entire System

The flow of the proposed BSS system is summarized in Table I, which lists each stage of the system, the obtained filter matrices at each stage and the corresponding section in this paper. A block diagram of the system is depicted in Fig. 1.



(a) Correct Perm. (b) Incorrect Perm.

Fig. 4. Rotation of the location vector. (a) Correct permutation and (b) incorrect permutation.

First, STFT of the multichannel input signal, $\mathbf{x}(\omega, t)$, is obtained with an appropriate time shift and window function. Once STFT is obtained, the Fourier coefficients at each frequency are treated as a complex time series. By doing this, the convolutive mixture problem is reduced to a complex but instantaneous mixture problem [9].

Next, the subspace method is applied to the input vector $\mathbf{x}(\omega, t)$ to obtain the subspace filter $\mathbf{W}(\omega)$. In this stage, room reflections and ambient noise are reduced in advance of the application of ICA. It should be noted that the node of the filter network is reduced from M to D in this stage as depicted in Fig. 1.

The instantaneous ICA is then applied to the output of the subspace stage, $\mathbf{y}(\omega, t)$ to obtain the filter matrix $\mathbf{U}(\omega)$. For the sake of convenience, the product of $\mathbf{W}(\omega)$ and $\mathbf{U}(\omega)$

$$\mathbf{B}(\omega) = \mathbf{U}(\omega)\mathbf{W}(\omega) \tag{4}$$

is termed the separation filter, hereafter.

After obtaining this separation filter, the permutation and the scaling problem must be solved. In this stage, the output of the separation filter is processed with the permutation matrix $\mathbf{P}(\omega)$ and the scaling matrix $\tilde{\mathbf{B}}_{\tilde{m}}^{+}(\omega)$.

Finally, the filter matrices obtained in the above stages are transformed into the time domain, and the input signal is processed with this time-domain filter network.

B. ICA Algorithm

In this subsection, the ICA algorithm used in this paper is briefly described. In this paper, the Infomax algorithm with feed-forward architecture [10], [11] extended to complex data [8] is used. In this stage, the input signal (the output of the subspace filter) $\mathbf{y}(\omega, t)$ is processed with the filter matrix $\mathbf{U}(\omega)$ as

$$\mathbf{z}(\omega, t) = \mathbf{U}(\omega)\mathbf{y}(\omega, t). \tag{5}$$

The learning rule is written as

$$\mathbf{U}(\omega, t+1) = \mathbf{U}(\omega, t) + \eta\left[\mathbf{I} - \varphi(\mathbf{z}(\omega, t))\mathbf{z}^H(\omega, t)\right]\mathbf{U}(\omega, t) \tag{6}$$

where the score function for the complex data $\varphi(\mathbf{z})$ is defined as [12]

$$\varphi(\mathbf{z}) = [\varphi(z_1), \ldots, \varphi(z_d), \ldots, \varphi(z_D)]^T \tag{7}$$

$$\varphi(z_d) = 2\tanh(G \cdot \mathrm{Re}(z_d)) + 2j\tanh(G \cdot \mathrm{Im}(z_d)). \tag{8}$$

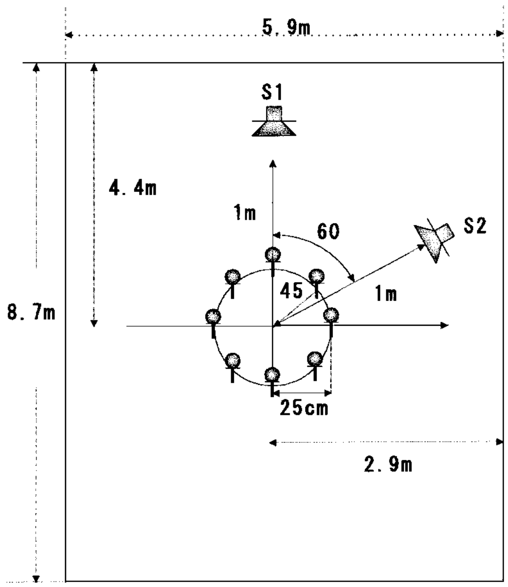


Fig. 5. Configuration of microphone array and sound sources.

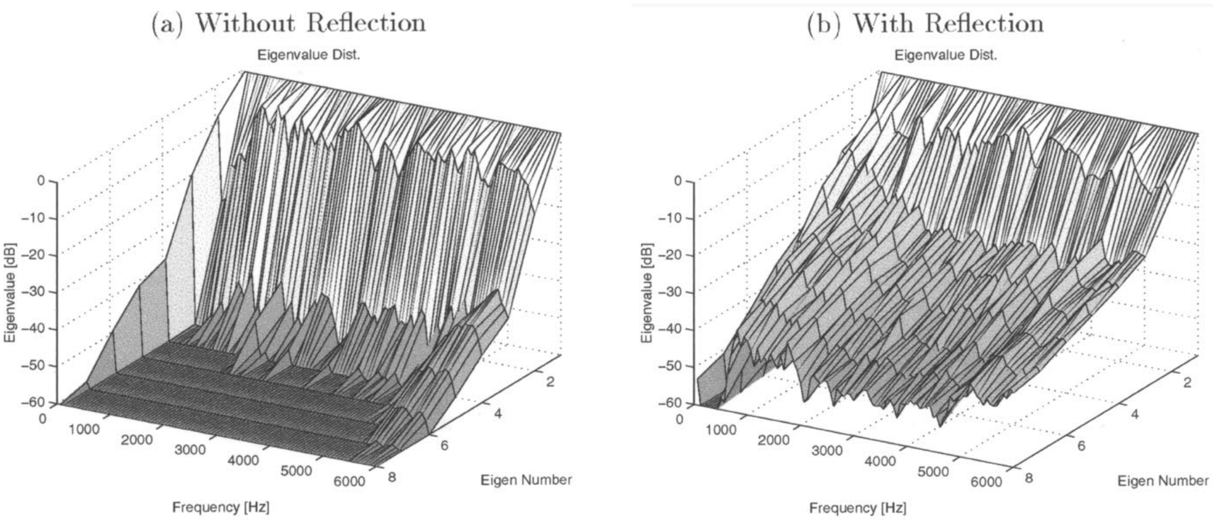
TABLE II
PARAMETERS OF ASR

Recognition Engine	HTK Ver.3.0
Model	HMM with continuous density
Feature vector	12th order MFCC + 12th order Δ MFCC + Δ log power
Training data	20,000 clean Japanese sentences
Recognition Task	492 Japanese words
Noise robust technique	none

TABLE III
PARAMETERS OF THE BSS SYSTEM

Sampling frequency	16 kHz
Length of STFT	512
Shift of STFT	16
Window function	hamming
Learning rate, η	0.0001
Gain for score function, G	100
Number of microphones, M	8
Number of sources, D	2
Reference range in permutation, K	5

The symbol z_d is the dth element of the vector $\mathbf{z}(\omega, t)$. The matrix $\mathbf{I}$ is an identity matrix. The symbol $\cdot^H$ denotes the Hermitian transpose. The constant η is termed the learning rate. The


(a) Without Reflection

(b) With Reflection

Fig. 6. Eigenvalue distribution.

symbol G is the gain constant for the nonlinear score function, assuming that the magnitude of $\mathbf{y}(\omega, t)$ is normalized.

C. Scaling Problem

In [9], it was proposed that the scaling problem be solved by filtering individual output of the separation filter by the inverse of $\mathbf{B}(\omega)$ separately. In this paper, the pseudoinverse of $\mathbf{B}(\omega)$, denoted as $\mathbf{B}(\omega)^+$, is used instead of the inverse of $\mathbf{B}(\omega)$ since $\mathbf{B}(\omega)$ is not square due to employment of the subspace method.

The nth component of $\mathbf{z}(\omega, t)$, $z_n(\omega, t)$ is filtered by $\mathbf{B}(\omega)^+$ separately as

$$\tilde{\mathbf{z}}_n(\omega, t) = \mathbf{B}(\omega)^+[0, \ldots, 0, z_n(\omega, t), 0, \ldots, 0]^T \tag{9}$$

where $\tilde{\mathbf{z}}_n(\omega, t) = [\tilde{z}_{1,n}(\omega, t), \ldots, \tilde{z}_{M,n}(\omega, t)]^T$ and $\tilde{z}_{m,n}(\omega, t)$ corresponds to the recovered signal of the nth source observed at the mth microphone. The operation (9) is equivalent to

$$\tilde{z}_{\tilde{m},n}(\omega, t) = B^+_{\tilde{m},n} z_n(\omega, t) \tag{10}$$

where $B^+_{\tilde{m},n}$ denotes the $(\tilde{m}, n)$th element of $\mathbf{B}(\omega)^+$. The symbol $\tilde{m}$ denotes an arbitrary microphone number. Equation (10) can be written in the matrix-vector notation as

$$\tilde{\mathbf{z}}(\omega, t) = \tilde{\mathbf{B}}^+_{\tilde{m}} \mathbf{z}(\omega, t) \tag{11}$$

where $\tilde{\mathbf{B}}^+_{\tilde{m}}(\omega)$ is a $D \times D$ diagonal matrix

$$\tilde{\mathbf{B}}^+_{\tilde{m}}(\omega) = \mathrm{diag}[B^+_{\tilde{m},1}, \ldots, B^+_{\tilde{m},D}] \tag{12}$$

and $\tilde{\mathbf{z}}(\omega, t) = [\tilde{z}_{\tilde{m},1}, \ldots, \tilde{z}_{\tilde{m},D}]^T$.

D. Filtering

Using the matrices obtained above, the final filtering matrix in the frequency domain can be written as

$$\mathbf{F}(\omega) = \mathbf{P}(\omega)\tilde{\mathbf{B}}^+_{\tilde{m}}(\omega)\mathbf{B}(\omega). \tag{13}$$

The filtering is conducted in the time domain to avoid time-domain aliasing. The time domain filters are obtained as the inverse Fourier transform of $\mathbf{F}(\omega)$ as

$$f_{n,m}(i) = \mathrm{IDFT}[F_{n,m}(\omega)]w(i) \tag{14}$$

where IDFT$[\cdot]$ operator denotes the inverse DFT. The symbols $F_{n,m}(\omega)$ and $f_{n,m}(i)$ denote the (n, m)th element of the frequency domain filter $\mathbf{F}(\omega)$ and its time domain correspondence, respectively. The symbol $w(i)$ denotes the windowing function. The multiplication by the windowing function is necessary for guaranteeing convergence of the impulse response of the filters in the time domain and avoiding time-domain aliasing.

IV. Subspace Method

A. Spatial Correlation Matrix

The spatial correlation matrix is defined as

$$\mathbf{R}(\omega) = E[\mathbf{x}(\omega, t)\mathbf{x}^H(\omega, t)]. \tag{15}$$

Since the subspace method is conducted at each frequency independently, the frequency index ω is omitted in this section for the sake of simplicity in notation.

Assuming that $\mathbf{s}(t)$ and $\mathbf{n}(t)$ are uncorrelated, $\mathbf{R}$ can be written as

$$\mathbf{R} = \mathbf{A}\mathbf{Q}\mathbf{A}^H + \mathbf{K}. \tag{16}$$

Matrix $\mathbf{Q} = E[\mathbf{s}(t)\mathbf{s}^H(t)]$ is the cross-spectrum matrix of the sources $\mathbf{s}(t)$. Matrix $\mathbf{K} = E[\mathbf{n}(t)\mathbf{n}^H(t)]$ is the correlation matrix of $\mathbf{n}(t)$. When $\mathbf{n}(t)$ includes room reflections of $\mathbf{s}(t)$, $\mathbf{s}(t)$ and $\mathbf{n}(t)$ are correlated and the above assumption does not hold. However, when the window length of STFT is short and the time interval between the direct sound and the reflection exceeds this window length, this assumption holds to some extent in a practical sense. A typical example of this is that a consonant portion of speech is overlapped by the reflections of a preceding vowel portion.

B. Properties of the Subspace Method

By taking the generalized eigenvalue decomposition of $\mathbf{R}$ as [13]

$$\mathbf{R} = \mathbf{K}\mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^{-1} \tag{17}$$

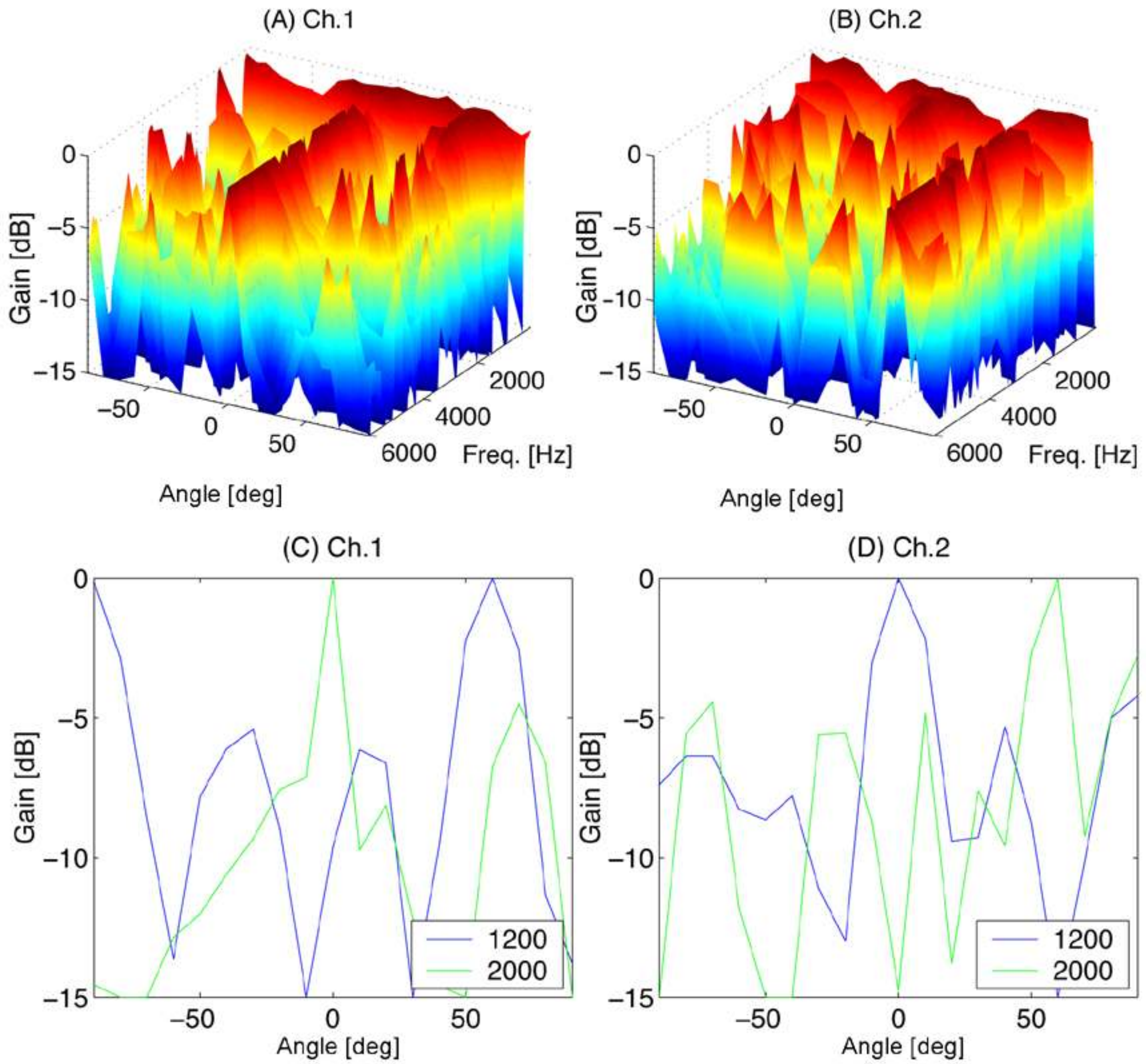


Fig. 7. Directivity pattern obtained by the subspace method. (a), (b): Whole frequency range and (c), (d): a slice of (a) and (b) at 1200 and 2000 Hz.

we have the eigenvector matrix $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_M]$ and the eigenvalue matrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$, where $\mathbf{e}_m$ and λ_m are the eigenvector and the eigenvalue, respectively. As described in Section II, the noise $\mathbf{n}(t)$ includes the reflection/reverberation of the signals and, thus, the correlation matrix of the noise, $\mathbf{K}$, cannot be observed separately. Therefore, in this paper, $\mathbf{K} = \mathbf{I}$ is assumed. This assumption is equivalent to the case in which the standard eigenvalue decomposition, $\mathbf{R} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1}$, is employed in the subspace method. In a physical sense, this corresponds to the assumption that $\mathbf{n}(t)$ is spatially white (e.g., [5]).

Based on the structure of $\mathbf{R}$ and the assumptions described above, the eigenvalues and eigenvectors have the following properties [6], [13], [14].

Property1) The energy of the D directional signals $\mathbf{s}(t)$ is concentrated on the D dominant eigenvalues.

Property2) The energy of $\mathbf{n}(t)$ is equally spread over all eigenvalues.

Property3) $\Re(\mathbf{A}) = \Re(\mathbf{E}_s)$, where $\mathbf{E}_s = [\mathbf{e}_1, \ldots, \mathbf{e}_D]$ denotes the eigenvectors corresponding to the D dominant eigenvalues.

Property4) $\Re(\mathbf{A}) = \Re(\mathbf{E}_n)^{\perp}$, where $\mathbf{E}_n = [\mathbf{e}_{D+1}, \ldots, \mathbf{e}_M]$ denotes the eigenvectors corresponding to the other $M - D$ eigenvalues.

The notation $\Re(\mathbf{A})$ denotes the space spanned by the column vectors of $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_D]$, i.e., $\Re(\mathbf{A}) = \mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_D)$. The notation $\Re(\mathbf{E}_n)^{\perp}$ denotes the orthogonal complement of $\Re(\mathbf{E}_n)$. The subspaces $\Re(\mathbf{E}_s)$ and $\Re(\mathbf{E}_n)$ are termed signal subspace and noise subspace, respectively. The vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_D\}$ and $\{\mathbf{e}_{D+1}, \ldots, \mathbf{e}_M\}$ become the basis of the signal subspace and the noise subspace, respectively. A typical eigenvalue distribution and the corresponding energy distribution that reflects Properties 1 and 2 are depicted in Fig. 2 ($M = 7$ and $D = 3$ is assumed). The relation of vectors that reflects Properties 3 and 4 is depicted in Fig. 3 ($M = 3$ and $D = 2$ is assumed).

C. Subspace Filter

In the subspace method, the input signal $\mathbf{x}(t)$ is processed as

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{18}$$

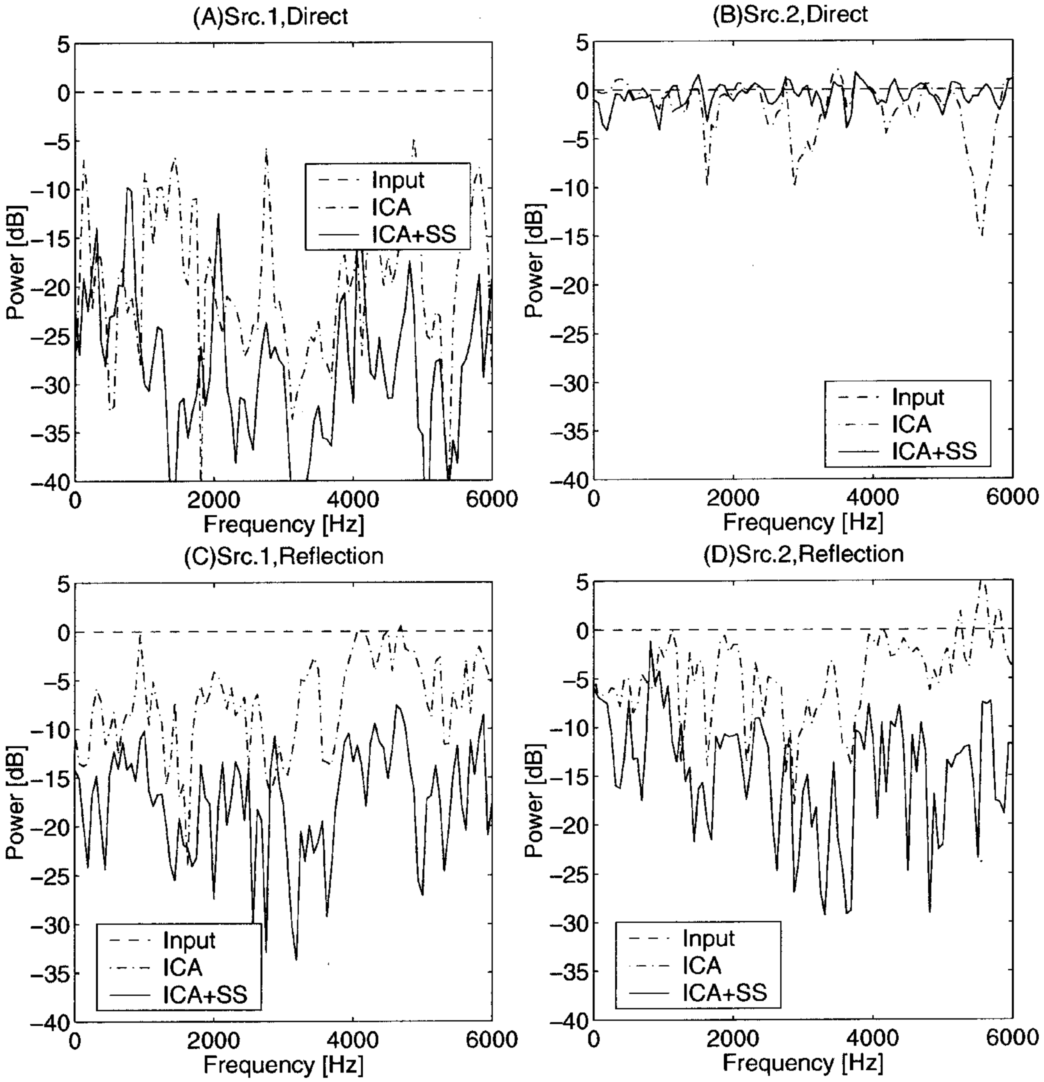


Fig. 8. Relative spectra of direct sound/reflection at the input/output of the system. (a) Source #1, direct sound; (b) Source #2, direct sound; (c) Source #1, reflection; and (d) Source #2, reflection. Regarding the output spectra (ICA and ICA + SS), the spectra at the output channel #2 are shown. SS: subspace method.

where the subspace filter is defined as

$$\mathbf{W} = \mathbf{\Lambda}_s^{-1/2}\mathbf{E}_s^H \tag{19}$$

where $\mathbf{\Lambda}_s = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$. The term $\mathbf{\Lambda}_s^{-1/2}$ is a normalization factor, the same as that used in PCA [1]. The term $\mathbf{E}_s^H$ plays a main role in the subspace filter that reduces the energy of $\mathbf{n}(t)$ in the noise subspace as described in Appendix II.

V. Permutation

A. Structure of Mixing Matrix

When the mixing matrix $\mathbf{A}(\omega)$ has the form of the model shown in (3), the nth column vector (location vector of the nth source) in the mixing matrix at the frequency ω and that at the adjacent frequency $\omega_0 = \omega - \Delta\omega$ are written as

$$\mathbf{a}_n(\omega) = \begin{pmatrix} e^{-j\omega\tau_{1n}} \\ \vdots \\ e^{-j\omega\tau_{Mn}} \end{pmatrix}, \qquad \mathbf{a}_n(\omega_0) = \begin{pmatrix} e^{-j(\omega-\Delta\omega)\tau_{1n}} \\ \vdots \\ e^{-j(\omega-\Delta\omega)\tau_{Mn}} \end{pmatrix}. \tag{20}$$

Here, $H_{m,n}(\omega) = 1$ in (3) is assumed for the sake of simplicity. From (20), it can be seen that the location vector $\mathbf{a}_n(\omega)$ is $\mathbf{a}_n(\omega_0)$ which is rotated by the angle θ_n as depicted in Fig. 4(a). Based on this relation (coherency) of the location vectors at the adjacent frequencies, the relation of the mixing matrix can be written as [15], [16]

$$\mathbf{A}(\omega) = \mathbf{T}(\omega, \omega_0)\mathbf{A}(\omega_0) \tag{21}$$

where the matrix $\mathbf{T}(\omega, \omega_0)$ is the rotation matrix. When the difference in frequency $\Delta\omega$ (frequency resolution of STFT) is sufficiently small

$$\mathbf{A}(\omega) \simeq \mathbf{A}(\omega_0), \qquad \mathbf{T}(\omega, \omega_0) \simeq \mathbf{I} \tag{22}$$

and the angle between the location vectors at ω and ω_0, θ_n, is small. Based on this, θ_n is expected to be the smallest for the correct permutation as depicted in Fig. 4.

B. Method for Solving Permutation

Based on the above discussion, permutation is solved so that the sum of the angles $\{\theta_1, \ldots, \theta_D\}$ between the location vec-

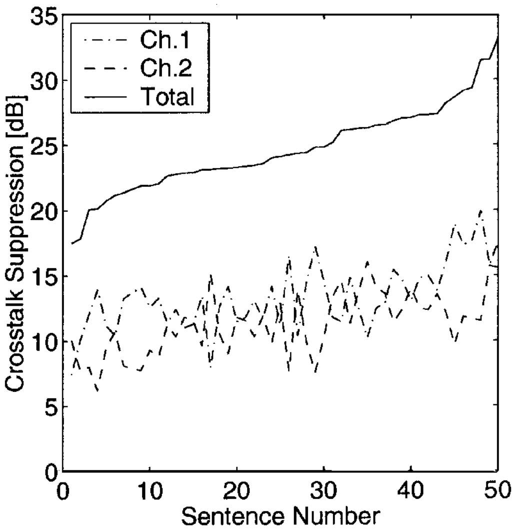


Fig. 9. Cross-talk suppression.

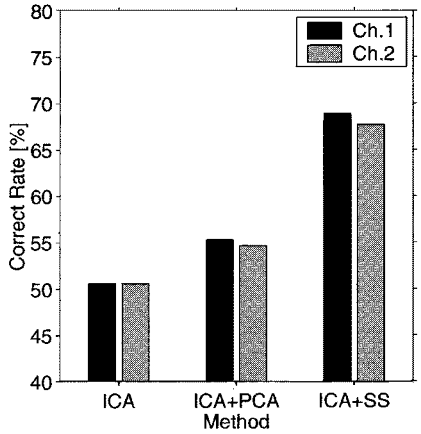


Fig. 10. Comparison of ASR rate for only ICA, ICA + PCA and ICA + Subspace. SS: Subspace.

tors in the adjacent frequencies is minimized. An estimate of the mixing matrix can be obtained as the pseudoinverse of the separation matrix $\mathbf{B}(\omega)$ [17] as

$$\hat{\mathbf{A}}(\omega) = \mathbf{B}^{+}(\omega). \tag{23}$$

Let us denote the mixing matrix multiplied by the arbitrary permutation matrix $\mathbf{P}$ as

$$\overline{\mathbf{A}}^{T}(\omega) = \mathbf{P}\hat{\mathbf{A}}^{T}(\omega). \tag{24}$$

The permutation $\mathbf{P}\hat{\mathbf{A}}^{T}(\omega)$ exchanges the row vectors of $\hat{\mathbf{A}}^{T}(\omega)$ (the column vectors of $\hat{\mathbf{A}}(\omega)$). The column vectors of $\overline{\mathbf{A}}(\omega)$ are denoted as $\overline{\mathbf{A}}(\omega) = [\overline{\mathbf{a}}_1(\omega), \ldots, \overline{\mathbf{a}}_D(\omega)]$. The cosine of the angle θ_n between the two vectors, $\overline{\mathbf{a}}_n(\omega)$ and $\overline{\mathbf{a}}_n(\omega_0)$, is defined as [18]

$$\cos\theta_n = \frac{\overline{\mathbf{a}}_n^H(\omega)\overline{\mathbf{a}}_n(\omega_0)}{\|\overline{\mathbf{a}}_n(\omega)\| \cdot \|\overline{\mathbf{a}}_n^H(\omega_0)\|}. \tag{25}$$

By using this, the permutation matrix is determined as

$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P}} F(\mathbf{P}) \tag{26}$$

where the cost function $F(\mathbf{P})$ is defined as

$$F(\mathbf{P}) = \frac{1}{D}\sum_{n=1}^{D}\cos\theta_n. \tag{27}$$

C. Confidence Measure

The above method assumes that the estimate of the mixing matrix $\hat{\mathbf{A}}(\omega)$ is a good approximation of the true mixing matrix $\mathbf{A}(\omega)$. However, at some frequencies, this assumption may not hold due to the failure of ICA. Since the permutation at frequency ω is determined based on only the information of the two adjacent frequencies, ω and ω_0, and the permutation is solved iteratively with increasing frequency, once the permutation at the certain frequency fails, the permutation in the succeeding frequencies may also fail.

To prevent this, the reference frequency ω_0 is extended to the following frequency range:

$$\omega_0 = \omega - k \cdot \Delta\omega, \qquad \text{for } k = 1, \ldots, K. \tag{28}$$

The cost function (27) is calculated at all K frequencies in this range. Let us denote the value of the cost function at $\omega_0 = \omega - k \cdot \Delta\omega$ as $F(\mathbf{P}, k)$. Next, a confidence measure for $F(\mathbf{P}, k)$ is considered. When the largest value of the cost function $\max F(\mathbf{P}, k)$ is close to $F(\mathbf{P}, k)$ with other permutations, it may be difficult to determine which permutation is correct, and the value of $F(\mathbf{P}, k)$ is not reliable. Based on this, the following confidence measure is defined:

$$C(k) = \max_{\mathbf{P}\in\Omega}[F(\mathbf{P}, k)] - \max_{\mathbf{P}\in\Omega'}[F(\mathbf{P}, k)]. \tag{29}$$

Here, Ω denotes the set of all possible $\mathbf{P}$ while Ω' denotes Ω without $\hat{\mathbf{P}} = \arg\max_{\mathbf{P}\in\Omega}[F(\mathbf{P}, k)]$. The appropriate reference frequency ω_0 is determined as $\omega_0 = \omega - \hat{k} \cdot \Delta\omega$ with

$$\hat{k} = \max_{k} C(k). \tag{30}$$

The permutation is then solved using the information at this reference frequency as

$$\hat{\mathbf{P}} = \arg\max_{\mathbf{P}} F(\mathbf{P}, \hat{k}). \tag{31}$$

VI. Experiment

A. Experimental Conditions

A signal separation experiment was conducted in an ordinary meeting room with a reverberation time of 0.4 s. The configuration of the sound sources (loudspeakers) and the microphones is depicted in Fig. 5. A microphone array with $M = 8$, mounted on a mobile robot (Nomad XR-4000), was used. The microphone array was circular in shape with a diameter of 0.5 m. The impulse responses from the sound sources to the microphones were measured and then convolved with the source signal to generate the input signal $\mathbf{x}(\omega, t)$. For measuring the cross-talk and the performance of the permutation, 50 pairs of Japanese sentences were used. For measuring the ASR rate, 492 pairs of Japanese words were used. As a speech recognizer, HTK software with the phonetic model provided by Japanese Dictation Toolkit [19]

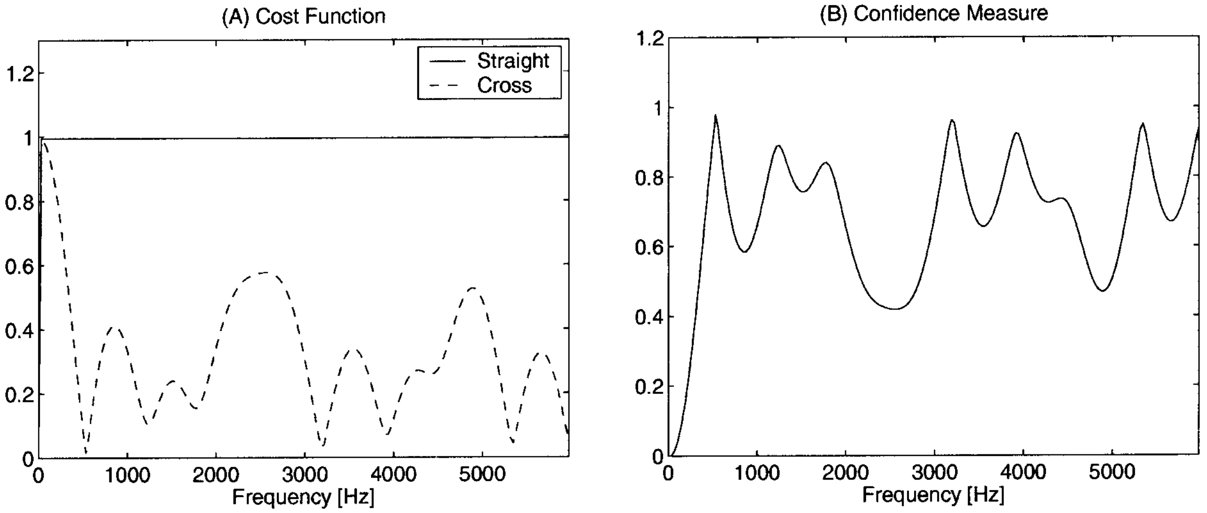


Fig. 11. Theoretical value of the cost function $F(\mathbf{P}, k)$ and the confidence measure $C(k)$ for $k = 1$.

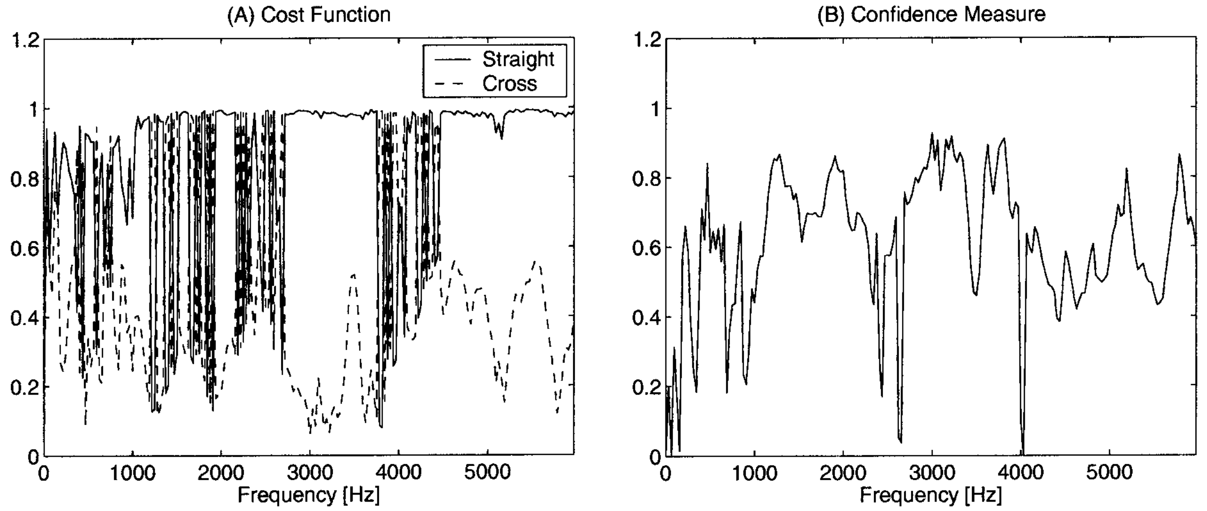


Fig. 12. Measured value of the cost function $F(\mathbf{P}, k)$ and the confidence measure $C(k)$ for $k = 1$.

was employed. The parameters of the ASR system are shown in Table II. The parameters of the BSS system are summarized in Table III.

B. Effect of Subspace Method

For constructing the subspace filter (19), the number of sources, D, is assumed to be known. The permutation in this section is solved by using the cross correlation between the output spectrogram and the source spectrogram (unknown in real situation) as "correct permutation" for evaluating only the effect of the subspace method. This method is denoted as source-output correlation (SOC) hereafter.

Fig. 6 shows the eigenvalue distribution of $\mathbf{R}(\omega)$. For the sake of comparison, the eigenvalue distribution without reflection is also shown. The eigenvalue distribution without reflection was obtained by eliminating the reflections in the impulse response using a window function. By comparing these, it can be seen that the energy of the direct sound is concentrated on the two dominant eigenvalues while the energy of the reflections is spread over the other eigenvalues. Therefore, Properties 1-4 in Section IV-B hold in a practical sense and the subspace method is applicable. However, it should be noted that the eigenvalue distribution for the noise in Fig. 6 was not perfectly flat compared with the ideal case depicted in Fig. 2. This is because the noise is spatially colored to some extent and, thus the assumption, $\mathbf{K} = \mathbf{I}$, did not perfectly hold in the real situation. This mismatch may result in the performance of the subspace filter being lower compared with the case with the spatially white noise.

Fig. 7 shows the directivity pattern of the subspace filter. From this figure, although it is not as clear as that of analytically designed beamformers, it can be seen that two acoustic beams, which are complementary in channel 1 and 2, appear in the directions of the sources, i.e., 0° and 60°. From this, it is understood that the subspace filter works as a self-organizing beamformer.

Fig. 8 shows the spectra of the direct sound and the reflection of Source #1 and #2 at the input/output of the system separately. For ease of viewing, the spectra were normalized by their input spectrum. For comparison, the case of BSS without the subspace method (only ICA) is also shown. In this case, the number of

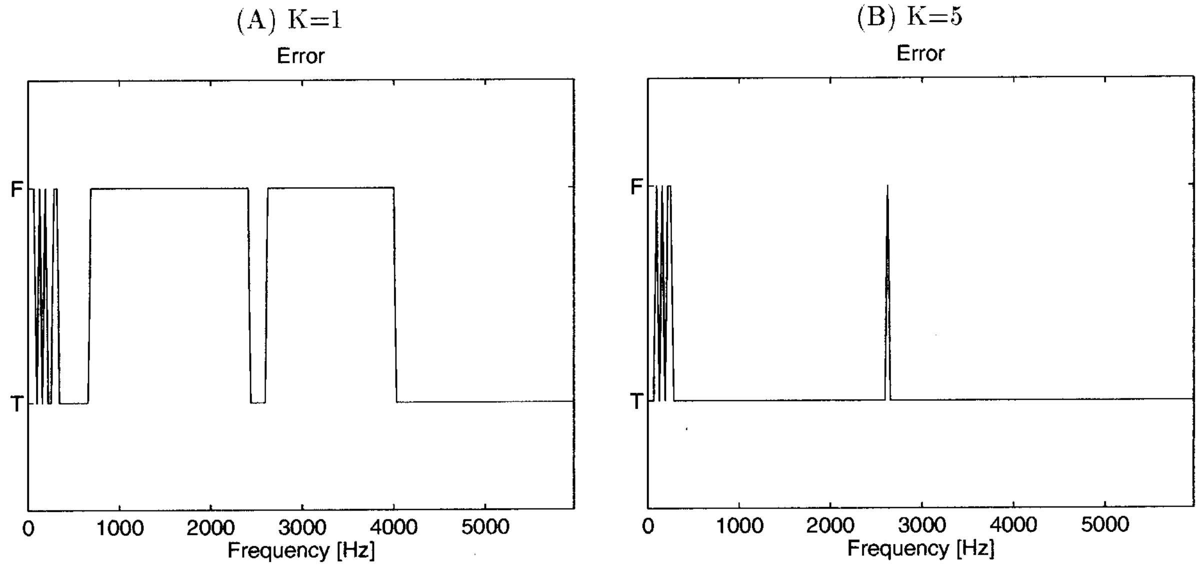


Fig. 13. Permutation error for $K = 1$ and $K = 5$. In the vertical axis, "T" denotes the case when the permutation is the same as that of "correct permutation" while "F" denotes the case when the permutation is different.

microphones is $M = 2(=D)$ and the two microphones closest to the sound sources were used. From Fig. 8(c) and (d), it can be seen that the reflections were reduced by 10-15 dB by the subspace method. On the other hand, Fig. 8(a) shows the effect of cross-talk suppression by ICA.

Fig. 9 shows the overall cross-talk suppression for 50 pairs of sentences shown in ascending order of total performance. The total performance is a simple sum of the cross-talk suppression of channel 1 and 2. From this figure, it can be seen that around 10-15 dB of overall cross-talk suppression for each channel was obtained.

Fig. 10 shows the results of the automatic speech recognition test applied to the output of BSS. As can be seen from this figure, the recognition rate was improved by around 18% by employing the subspace method (denoted as ICA + Subspace) compared to the case without the subspace method (only ICA). For comparison, the case of ICA + PCA was also tested. In this case, the subspace filter, $\mathbf{W} = \mathbf{\Lambda}_s^{-1/2}\mathbf{E}_s^H$, was replaced by the PCA filter, $\mathbf{W}_{PCA} = \mathbf{\Lambda}^{-1/2}\mathbf{E}^H$ [1], in Fig. 1. The number of microphones was $M = 2$ in the same manner as that with only ICA. The effect of PCA is only to orthogonalize the output of PCA (input of ICA). On the other hand, in the case of ICA + Subspace, the subspace method has the effect of both orthogonalizing the output and reducing room reflections. Therefore, from Fig. 10, it is considered that, in the 18% increase in ASR rate, the effect of the orthogonalization accounts for around 5% of the increase and the effect of the reflection reduction accounts for the remaining 13% increase in ASR rate.

C. Permutation

Fig. 11(a) shows the theoretical value of the cost function $F(\mathbf{P}, k)$ with $k = 1$ for the model of the mixing matrix shown in (3). In this figure, "Straight" corresponds to the case when $\mathbf{A}(\omega)$ is unchanged, and "Cross" corresponds to the case when the column vectors of $\mathbf{A}(\omega)$ are exchanged. For the model of the mixing matrix, $\mathbf{A}(\omega)$ does not require any permutation. In this case, therefore, Straight corresponds to the correct permutation, and Cross corresponds to the incorrect permutation. From this figure, it can be seen that Straight shows the value close to one for all frequencies while Cross shows a smaller value at all frequencies except in the very low frequencies. The confidence measure $C(1)$ depicted in Fig. 11(b) shows high values except at the very low frequencies. This means that the proposed cost function can be used for solving permutation at all frequencies except at the very low frequencies. The reason for Cross showing a large value at the very low frequencies is that the phase difference in the column vectors of $\mathbf{A}(\omega)$ is small at the low frequencies.

Fig. 12(a) shows the cost function $F(\mathbf{P}, k)$ with $k = 1$ obtained from the trained filter network $\mathbf{B}(\omega)$ with real data. From this figure, it can be seen that there are many vertical lines. These vertical lines show that it is necessary to exchange the output at those frequencies. In Fig. 12(b), it can be seen that the confidence measure $C(1)$ becomes low at some frequencies.

Fig. 13 shows the permutation error for $K = 1$ and $K = 5$. Permutation error is defined as the case when the result of IFC differs from that of SOC (assumed as correct permutation). When $K = 1$, permutation error "starts" at several frequencies where $C(1)$ is small and "propagates" toward the upper frequencies. On the other hand, when $K = 5$, permutation error is almost completely corrected. This is due to the relations of $\mathbf{A}(\omega)$, ..., $\mathbf{A}(\omega - 5\Delta\omega)$ being taken into account and the unreliable information being ignored by use of the confidence measure $C(k)$.

Fig. 14 shows the error rate. The input was 50 pairs of Japanese sentences. It can be seen that the error rate was small in the frequency range over 300 Hz (the region to the right of the dotted vertical line). On the other hand, below 300 Hz, the error rate increases. However, at these very low frequencies, the performance of ICA is also reduced due to the phase difference in $\mathbf{A}(\omega)$ being small, and the permutation sometimes becomes meaningless.

Table IV shows a comparison of ASR rate when the permutation is solved by SOC and IFC. From this, the ASR rate reduced

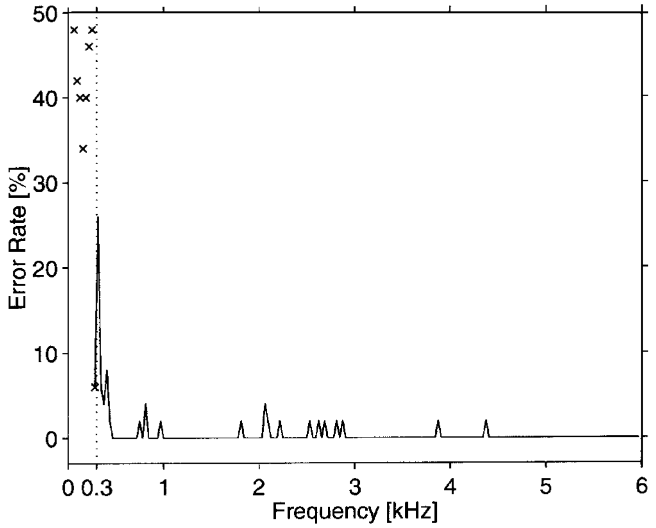


Fig. 14. Error rate in solving permutation. The error rate is plotted with "x" below 300 Hz and with a solid line over 300 Hz.

TABLE IV
ASR RATE FOR DIFFERENT PERMUTATION METHODS [%]

permutation	Ch.1	Ch.2
SOC("correct")	68.5	67.5
IFC	64.4	63.8
difference	-4.1	-3.7

by employing IFC is small (around 4%) compared with the error rate for IFSEC reported in [3] (around 18%). In the application of ASR, the contribution of the lower frequency component is small due to the pre-emphasis [20]. Therefore, the permutation error at very low frequencies is considered to be small.

VII. CONCLUSION

In this paper, an approach combining array processing and ICA for the blind separation of acoustic signals in a reflective environment was proposed. Two array processing techniques were employed for pre- and post-processing of ICA.

As a pre-processor, the subspace method was employed to reduce the effect of room reflections. As shown in this paper, the subspace method functions as a self-organizing beamformer focusing on the target sources and is suitable for the framework of the blind separation. From the results of the experiments, it was shown that the subspace method reduced the power of the reflections by around 10 dB and improved the ASR rate by around 18% for the array and the sound environment used in the experiment.

The performance of the subspace method depends on both the array configuration and the sound environment. Regarding the array configuration, the subspace method is analogous to the conventional DS beamformer since the subspace method has the same noise reduction mechanism as that of the DS beamformer. As for the sound environment, the directivity of reflections is assumed to be small. This assumption holds when reflections are coming from many directions and the coherency of the reflections between the microphones is reduced. The sound environment used in the experiment where the microphone array was placed at some distance from the walls of the room meets this condition. When there is a strong reflection with high directivity such as when the microphone-array is placed close to a hard wall, this assumption may not hold. In this case, some modification may be required for the subspace method [21]. This case must be treated in a future study.

As a post-processor, a new method for solving the permutation problem was proposed. This method utilizes the coherency (continuity) of the mixing matrix at adjacent frequencies, a physical property peculiar to acoustic problems. By employing this method, the permutation error was reduced to 4% in terms of the ASR rate. An advantage of this method is that, unlike IFSEC, the performance of IFC is independent of source spectra. Another advantage over IFSEC is that IFC does not require a large memory space, such as that required for IFSEC, to store the output spectrogram (see Appendix I), a desirable feature for implementation in small-sized hardware such as DSP (digital signal processor).

In this paper, the conventional ICA algorithm was employed to combine the proposed pre- and post-array processing. However, the recent progress of the ICA algorithm will lead to the further improvement of the performance of the proposed system.

APPENDIX I
IFSEC

As indicated in (25) and (27), the cost function of the proposed method for solving permutation is written as

$$F(\mathbf{P}) = \frac{1}{D}\sum_{n=1}^{D} \frac{\bar{\mathbf{a}}_n^H(\omega)\bar{\mathbf{a}}_n(\omega_0)}{\|\bar{\mathbf{a}}_n(\omega)\| \cdot \|\bar{\mathbf{a}}_n(\omega_0)\|}. \tag{32}$$

On the other hand, the cost function of IFSEC is written as

$$F_{IFSEC}(\mathbf{P}) = \frac{1}{D}\sum_{n=1}^{D} \frac{\bar{\mathbf{z}}_n^H(\omega)\bar{\mathbf{z}}_n(\omega_0)}{\|\bar{\mathbf{z}}_n(\omega)\| \cdot \|\bar{\mathbf{z}}_n(\omega_0)\|}. \tag{33}$$

In IFSEC, this cost function is maximized in a manner similar to that of the proposed method to solve the permutation. The vector $\bar{\mathbf{z}}_n$ is the nth column vector of the following matrix in a manner similar to (24)

$$\bar{\mathbf{Z}}^T(\omega) = \mathbf{P}\hat{\mathbf{Z}}^T(\omega). \tag{34}$$

The matrix $\hat{\mathbf{Z}}^T(\omega)$ has the estimated spectral envelope (the output of BSS smoothed by the moving-average) as a column vector as

$$\hat{\mathbf{Z}} = [\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_D] \tag{35}$$

where

$$\hat{\mathbf{z}}_n = [\hat{z}_n(\omega, t_1), \ldots, \hat{z}_n(\omega, t_2)]^T. \tag{36}$$

The symbol $\hat{z}_n(\omega, t)$ denotes the estimated spectral envelope at the nth channel, frequency ω, and the tth time frame. The symbols, $[t_1, t_2]$, denote the period of spectrogram used for solving the permutation.

As indicated in (32) and (33), the essential difference of the proposed IFC and the conventional IFSEC is the vectors used in their cost functions. The dimension of the vector in IFC is always M (8 in this paper), while that of IFSEC is dependent on the length of the spectrogram to be used for solving the permutation. For example, when using 1 s of spectrogram with a 16-point frame shift, the dimension of the vector is 1000 for IFSEC. From this, it can be known that the proposed IFC consumes less memory space and computational load. It should be noted that, for the sake of simplicity in explanation, IFSEC described above was simplified. For further details, see [9].

Appendix II
An Aspect of the Subspace Method as a Self-Organizing Beamformer

According to Properties 1 and 3, the directional component $\mathbf{As}(t)$ can be expanded with the subset of the basis vectors, $\{\mathbf{e}_1, \ldots, \mathbf{e}_D\}$, as

$$\mathbf{As}(t) = \sum_{i=1}^{D} \alpha_i(t)\mathbf{e}_i \tag{37}$$

where $\alpha_i(t)$ is the projection coefficient of $\mathbf{As}(t)$ onto the basis vector $\mathbf{e}_i$. On the other hand, due to Property 2, $\mathbf{n}(t)$ is expanded using all the basis vectors, $\{\mathbf{e}_1, \ldots, \mathbf{e}_M\}$, as

$$\mathbf{n}(t) = \sum_{i=1}^{M} \beta_i(t)\mathbf{e}_i \tag{38}$$

where $\beta_i(t)$ is a projection coefficient of $\mathbf{n}(t)$ onto the basis vector $\mathbf{e}_i$. Equations (37) and (38) can be written in a matrix-vector notation as

$$\mathbf{As}(t) = \mathbf{E}_s\boldsymbol{\alpha}(t) \tag{39}$$

$$\mathbf{n}(t) = \mathbf{E}\boldsymbol{\beta}(t) \tag{40}$$

where $\boldsymbol{\alpha}(t) = [\alpha_1(t), \ldots, \alpha_D(t)]^T$ and $\boldsymbol{\beta}(t) = [\beta_1(t), \ldots, \beta_M(t)]^T$. Equation (40) can be split as

$$\mathbf{n}(t) = \mathbf{n}_s(t) + \mathbf{n}_n(t) \tag{41}$$

where

$$\mathbf{n}_s(t) = \mathbf{E}_s\boldsymbol{\beta}_s(t) \tag{42}$$

$$\mathbf{n}_n(t) = \mathbf{E}_n\boldsymbol{\beta}_n(t) \tag{43}$$

and $\boldsymbol{\beta}_s(t) = [\beta_1(t), \ldots, \beta_D(t)]^T$ and $\boldsymbol{\beta}_n(t) = [\beta_{D+1}(t), \ldots, \beta_M(t)]^T$. From (42) and (43), $\mathbf{n}_s(t) \in \Re(\mathbf{E}_s)$ and $\mathbf{n}_n(t) \in \Re(\mathbf{E}_n)$. Applying the subspace filter to these components in (39), (42) and (43) and using the properties of the eigenvectors, $\mathbf{E}_s^H\mathbf{E}_s = \mathbf{I}$ and $\mathbf{E}_s^H\mathbf{E}_n = 0$, we obtain

$$\mathbf{WAs}(t) = \mathbf{\Lambda}^{-1/2}\mathbf{E}_s^H(\mathbf{E}_s\boldsymbol{\alpha}(t)) = \mathbf{\Lambda}^{-1/2}\boldsymbol{\alpha}(t) \tag{44}$$

$$\mathbf{Wn}_s(t) = \mathbf{\Lambda}^{-1/2}\mathbf{E}_s^H(\mathbf{E}_s\boldsymbol{\beta}_s(t)) = \mathbf{\Lambda}^{-1/2}\boldsymbol{\beta}_s(t) \tag{45}$$

$$\mathbf{Wn}_n(t) = \mathbf{\Lambda}^{-1/2}\mathbf{E}_s^H(\mathbf{E}_n\boldsymbol{\beta}_n(t)) = 0. \tag{46}$$

From these, it can be seen that, by applying the subspace filter, the components in the signal subspace $\mathbf{As}(t)$ and $\mathbf{n}_s(t)$ are preserved while the component in the noise subspace $\mathbf{n}_n(t)$ is cancelled. When the number of microphones M is considerably larger than that of the number of sources D, it is expected that a large portion of $\mathbf{n}(t)$ can be cancelled by this subspace filter.

On the other hand, the DS beamformer in the frequency domain that focuses on the nth target source can be expressed as [22]

$$\mathbf{y}_{DS}(t) = \mathbf{w}_{DS}\mathbf{x}(t) \tag{47}$$

where

$$\mathbf{w}_{DS} = \frac{1}{M}[e^{j\omega\tau_{1,n}}, \ldots, e^{j\omega\tau_{M,n}}]. \tag{48}$$

For the sake of simplicity, it is assumed that $H_{1,n} = \cdots = H_{M,n} = 1$ in (3). By using the vector notation, (48) can be written as

$$\mathbf{w}_{DS} = \frac{\mathbf{a}_n^H}{\mathbf{a}_n^H\mathbf{a}_n} \tag{49}$$

where $\mathbf{a}_n$ denotes the nth column vector of $\mathbf{A}$. The denominator, $\mathbf{a}_n^H\mathbf{a}_n$, is employed as a normalization factor. By extending (47) and (49) so that the D target sources are focused, the DS beamformer becomes

$$\mathbf{y}_{DS}(t) = \mathbf{W}_{DS}\mathbf{x}(t) \tag{50}$$

where

$$\mathbf{W}_{DS} = \frac{\mathbf{A}^H}{\mathbf{A}^H\mathbf{A}}. \tag{51}$$

Applying the DS beamformer $\mathbf{W}_{DS}$ to $\mathbf{n}_n(t)$

$$\mathbf{W}_{DS}\mathbf{n}_n(t) = \frac{\mathbf{A}^H}{\mathbf{A}^H\mathbf{A}}\mathbf{E}_n\boldsymbol{\beta}_n(t) = 0. \tag{52}$$

This is because, due to Property 4, $\mathbf{A}^H\mathbf{E}_n = 0$.

According to the above discussion, the subspace filter $\mathbf{W}$ and the DS beamformer $\mathbf{W}_{DS}$ have the same noise reduction mechanism, i.e., a mechanism which cancels the component in the noise subspace, $\mathbf{n}_n(t)$. The essential difference in the subspace method and the DS beamformer is that, in the DS beamformer, knowledge of the mixing matrix $\mathbf{A}$ is required in the design of the beamformer as shown in (51) while, in the subspace method, no previous knowledge is required. In this sense, the subspace filter can be considered as a self-organizing beamformer.

References

[1] T.-W. Lee, *Independent Component Analysis*. Norwell, MA: Kluwer, 1998.

[2] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, pp. 2677–2684, Oct. 1999.

[3] F. Asano and S. Ikeda, "Evaluation and real-time implementation of blind source separation system using time-delayed decorrelation," in *Proc. ICA2000*, June 2000, pp. 411–415.

[4] M. Kawamoto, A. Barros, K. Matsuoka, and N. Ohnishi, "A method of real-world blind separation implementation in frequency domain," in *Proc. ICA 2000*, June 2000, pp. 267–272.

[5] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[6] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech, Audio Processing*, vol. 8, pp. 497–507, Sept. 2000.

[7] F. Asano, Y. Motomura, H. Asoh, and T. Matsui, "Effect of pca filter in blind source separation," in *Proc. ICA2000*, June 2000, pp. 57–62.

[8] S. Ikeda and N. Murata, "A method of ica in time-frequency domain," in *Proc. ICA'99*, Jan. 1999, pp. 365–371.

[9] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomput.*, vol. 41, pp. 1–24, 2001.

[10] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[11] S. Amari, T. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, 1997.

[12] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Proc. Int. Workshop on Independence and Artificial Neural Networks*, 1998.

[13] R. Roy and T. Kailath, "Esprit—Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 984–995, July 1989.

[14] A. Cantoni and L. C. Godara, "Resolving the direction of sources in a correlated field incident on an array," *J. Acoust. Soc. Amer.*, vol. 67, pp. 1247–1255, 1981.

[15] H. Hung and M. Kaveh, "Focussing matrices for coherent signal-subspace processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 1272–1281, 1988.

[16] ——, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 4, pp. 823–831, 1985.

[17] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Processing*, vol. 45, pp. 434–443, Feb. 1997.

[18] G. Strang, *Linear Algebra and Its Application*. Orlando, FL: Harcourt Brace Jovanovich, 1988.

[19] T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Japanese dictation toolkit: Plug-and-play framework for speech recognition r&d," in *Proc. ICASSP'99*, Mar. 1999, pp. I–393.

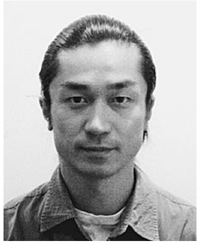
[20] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[21] T. Shan, M. wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 806–811, Aug. 1985.

[22] S. U. Pillai, *Array Signal Processing*. New York: Springer-Verlag, 1989.

Futoshi Asano (M'95) received the B.S. degree in electrical engineering, and the M.S. and Ph.D. degrees in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1986, 1988 and 1991, respectively.

From 1991 to 1995, he was a Research Associate at R.I.E.C., Tohoku University. From 1993 to 1994, he was a Visiting Researcher at A.R.L., Pennsylvania State University. From 1995 to 2001, he was with the Electrotechnical Laboratory, Tsukuba, Japan. Currently, he is a Group Leader in National Institute of Advanced Industrial Science and Technology, Tsukuba. His research interests include array signal processing, adaptive signal processing, neural network, statistical signal processing, and speech recognition.



Shiro Ikeda (M'95) received the B.Eng., M.Eng., and Dr.Eng. in mathematical engineering and information physics from University of Tokyo, Japan, in 1991, 1993, and 1995, respectively.

He has been an Associate Professor with the Brain Science and Engineering Department at the Kyushu Institute of Technology, Japan, since April 2001. His research interests include information geometry, signal processing, and machine learning.

Michiaki Ogawa was born in Fukui, Japan, on February 26, 1977. He received the Bachelor of Library and Information Science degree from University of Library and Information Science, Japan, in 1999 and the M.E. degree from University of Tsukuba, Japan, in 2001.

Hideki Asoh (M'90) received the M.Eng. degree in information engineering from the University of Tokyo, Japan, in 1983.

In April 1983, he joined the Electrotechnical Laboratory as a Researcher. From 1990 to 2001, he was a Senior Researcher. From 1990 to 1991, he was with the German National Research Center for Information Technology as a Visiting Research Scientist. Since April 2001, he has been a Senior Research Scientist of Information Technology Institute, National Institute of Advanced Industrial Science and Technology, Japan. His research interests are in machine learning and man–machine interaction.

Nobuhiko Kitawaki (M'87) was born in Aichi, Japan, on September 27, 1946. He received the B.E., M.E., and Dr.Eng. degrees from Tohoku University, Japan, in 1969, 1971, and 1981, respectively.

From 1971 to 1997, he was engaged in research on speech and acoustics information processing at the Laboratories of Nippon Telegraph and Telephone Corporation, Japan. From 1993 to 1997, he was Executive Manager of the Speech and Acoustics Laboratory. He currently serves as Professor of the Institute of Information Sciences and Electronics, and Dean of the College of International Studies, University of Tsukuba, Japan. He contributed to ITU-T Study Group 12 from 1981 to 2000, and has served as a Rapporteur since 1985.

Dr. Kitawaki is a Fellow of the Institute of Electronics, Information, and Communication Engineers of Japan (IEICEJ) and a member of the Acoustical Society of Japan, and the Information Processing Society of Japan. He received Paper Awards from IEICEJ in 1979 and 1984.