J Sign Process Syst
DOI 10.1007/s11265-009-0413-9

DOA Estimation for Multiple Sparse Sources with Arbitrarily Arranged Multiple Sensors

Shoko Araki · Hiroshi Sawada · Ryo Mukai · Shoji Makino

Received: 13 February 2009 / Revised: 16 July 2009 / Accepted: 25 September 2009

Abstract This paper proposes a method for estimating the direction of arrival (DOA) of multiple source signals for an underdetermined situation, where the number of sources N exceeds the number of sensors M ($M < N$). Some DOA estimation methods have already been proposed for underdetermined cases. However, since most of them restrict their microphone array arrangements, their DOA estimation ability is limited to a 2-dimensional plane. To deal with an underdetermined case where sources are distributed arbitrarily, we propose a method that can employ a 2- or 3-dimensional sensor array. Our new method employs the source sparseness assumption to handle an underdetermined case. Our formulation with the sensor coordinate vectors allows us to employ arbitrarily arranged sensors easily. We obtained promising experimental results for 2-dimensionally distributed sensors and sources 3×4, 3×5 (#sensors $\times$ #speech sources), and for 3-dimensional case with 4×5 in a room (reverberation time (RT) of 120 ms). We also investigate the DOA estimation performance under several reverberant conditions.

Keywords Direction of arrival (DOA) · Sparseness · Clustering · Microphone array · Anechoic model

S. Araki (✉) · H. Sawada · R. Mukai
NTT Communication Science Laboratories,
NTT Corporation, 2-4 Hikaridai, Seika-cho,
Soraku-gun, Kyoto 619-0237, Japan
e-mail: shoko@cslab.kecl.ntt.co.jp

H. Sawada
e-mail: sawada@cslab.kecl.ntt.co.jp

R. Mukai
e-mail: ryo@eye.brl.ntt.co.jp

S. Makino
University of Tsukuba, 1-1-1 Tennodai, Tsukuba,
Ibaraki 305-8577, Japan
e-mail: maki@tara.tsukuba.ac.jp

1 Introduction

Direction of arrival (DOA) estimation is an important fundamental technique in the array signal processing field [1–3]. The DOA estimation of speech, which is the focus of this paper, has many applications including teleconference system and robotics applications. Such applications usually have to deal with situations where the active sources outnumber the sensors. In this paper, we propose a new method for estimating the DOAs of sources under such circumstances.

The most widely used DOA estimation methods are subspace based methods, e.g., the MUSIC (MUltiple SIgnal Classification) algorithm [4], and its variants. Because these methods need a noise subspace, they require more sensors than sources $M \geq N + 1$, that is they can be applied only when $M > N$. By contrast, DOA estimation method has been proposed that is based on independent component analysis (ICA) [5–7]. This method estimates DOAs directly from the separation matrix estimated with ICA by utilizing the fact that the separation matrix is related to the source mixing process. Because this method is based on ICA, it can be employed when $M \geq N$. However, it still cannot be used for an underdetermined case where $M < N$.

In order to cope with underdetermined cases where $M < N$, we propose a new DOA estimation method that assumes source sparseness. A sparse source has a

Published online: 09 October 2009

sharp probability density function: the signal is close to zero at most of the time-frequency slots, and has large values on rare occasions (see e.g., [8, 9]). If the signals are assumed to be sufficiently sparse in the time-frequency domain, we can suppose that only one source is dominant in each time-frequency slot. Therefore, the phase difference between sensor observations at each time-frequency slot holds the geometric information of the dominant source at each time-frequency slot. Our method uses this geometric information at each time-frequency slot and clusters the information. As each cluster corresponds to an individual source, we can estimate the DOAs by using the cluster centroids and given sensor location information. We have already proposed a sparseness based blind source separation algorithm with observation vector clustering [10, 11]. In this paper, we show that we can also estimate the DOAs of more sources than sensors by leveraging the observation vector clustering results.

Some DOA estimation methods for underdetermined cases have already been proposed for narrowband signals [12–14], and wideband signals especially for speech signals [15–19]. The method described in [15] is based on source sparseness. It clusters the phase differences of only two sensor observations, and then estimates the DOA. Shamsunder and Giannakis [16] estimates the DOA from the cumulants of observations by assuming the non-gaussianity of the sources, and utilizing a linear sensor array. The authors of [17] also proposed a sparseness method with Laplacian mixture models for a two-sensor setup. These methods limit the DOA estimation ability to a 2-dimensional half-plane. The authors of [18] and [19] utilized a triangular sensor array and clustered the phase differences of each sensor pair by assuming source sparseness. Their method expands the DOA estimation ability to an entire 2-dimensional plane. However, their approach still cannot handle 3-dimensionally distributed sources. Moreover, their formulation assumed a *regular*-triangle sensor array. They have to re-formulate their method to use an array with another arrangement.

On the other hand, as our formulation utilized the sensor coordinate vectors, our newly proposed method is more general. That is, we do not need to re-formulate our method when employing an arbitrary sensor arrangement including a 3-dimensional arrangement. Our method can easily employ a 3-dimensional sensor array, and therefore, estimate the DOAs of 3-dimensionally distributed sources.

In this paper, we show successful results with our new method in estimating the DOAs for $M \times N$ of 3×4, 3×5 (2-dimensional arrangement) and 4×5 (3-dimensional). Neither the MUSIC algorithm nor the ICA based method can be used in such situations. Experimental results also show another advantage of our method compared with the MUSIC algorithm even when there are fewer sources than sensors. When sources are positioned close together, the MUSIC algorithm fails to estimate their DOAs, whereas the proposed method still succeeds. In addition to the results reported in [20], we also show the DOA estimation result with our proposed method under several reverberant conditions [21].

The organization of this paper is as follows. Section 2 describes the problem of DOA estimation and defines DOA. In Section 3, we explain our novel DOA estimation method. Section 4 reports some experimental results obtained with non-linearly arranged sensors in underdetermined scenarios. Even when the sources and sensors are distributed 2- or 3- dimensionally, we can estimate DOAs precisely for each scenario under reverberant (RT = 120 ms) conditions. We also investigated the performance under different reverberation conditions. The final section concludes this paper.

2 Problem Description

2.1 Observation Model

Suppose that sources $\mathrm{s}_1, \ldots, \mathrm{s}_N$ are convolutively mixed and observed at M sensors

$$\mathrm{x}_j(t) = \sum_{i=1}^{N} \sum_l \mathrm{h}_{ji}(l)\, \mathrm{s}_i(t-l), \quad j=1, \ldots, M, \tag{1}$$

where $\mathrm{h}_{ji}(l)$ represents the impulse response from source i to sensor j. In this paper, we particularly consider a situation where the number of sources N can exceed the number of sensors M ($M < N$). Here we assume that the number of sources N is given. Our task is to estimate the DOAs of the N sources from the sensor observations. We will formulate the DOA estimation problem in Section 2.2.

As with most DOA estimation techniques, this paper employs a time-frequency domain approach. Using a short-time Fourier transform (STFT), the convolutive mixtures (1) can be converted to instantaneous mixtures at each frequency f:

$$x_j(f, \tau) = \sum_{i=1}^{N} h_{ji}(f) s_i(f, \tau), \tag{2}$$

or in vector notation,

$$\mathbf{x}(f, \tau) = \sum_{i=1}^{N} \mathbf{h}_i(f) s_i(f, \tau), \tag{3}$$

where $h_{ji}(f)$ is the frequency response from source i to sensor j, $s_i(f, \tau)$ is the STFT of a source signal s_i, and τ is a time index. We call $\mathbf{x} = [x_1, \ldots, x_M]^T$ an

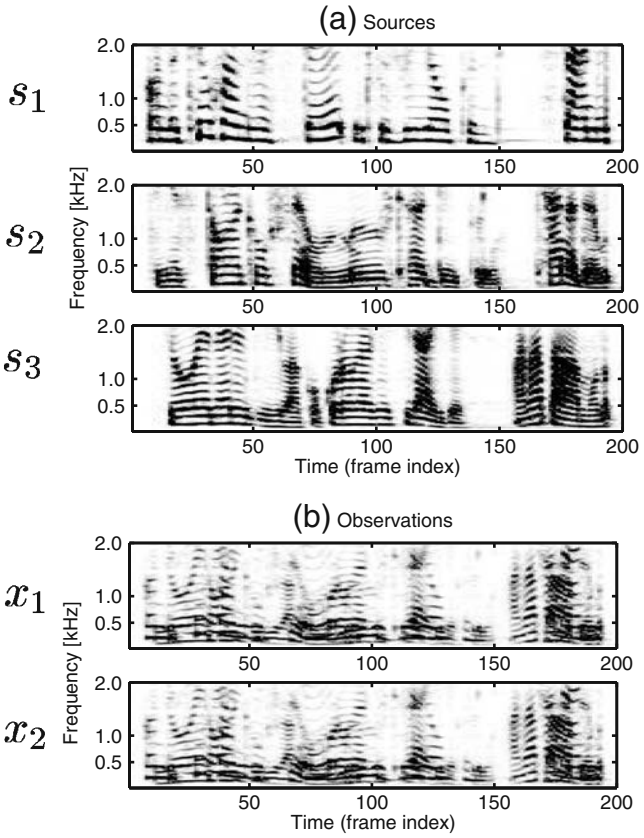


Figure 1 Example spectra of **a** speech sources and **b** observations ($N = 3$, $M = 2$).

observation vector and $\mathbf{h}_i = [h_{1i}, \ldots, h_{Mi}]^T$ is a vector of the frequency responses from source s_i to all sensors.

In the time-frequency domain, the sparseness of a source signal, which is widely used to solve the underdetermined problem [8–11, 15, 18], becomes prominent, if the source is colored and non-stationary such as speech. When the signals are sufficiently sparse, we can assume that the sources rarely overlap at each time-frequency point, and Eq. 3 can be approximated as

$$\mathbf{x}(f,\tau) \approx \mathbf{h}_k(f) s_k(f,\tau), \quad {}^{\exists}k \in \{1, \cdots, N\}, \tag{4}$$

where $s_k(f,\tau)$ is the dominant source at the time-frequency point (f,τ). For instance this is true for speech signals in the time-frequency domain [8, 9, 22]. Figure 1a and b show example spectra of three speech sources and observations, respectively, in which we can see their temporal/frequency sparseness.

2.2 DOAs of Source Signals

Let us define the DOAs of sources in detail. Let $\mathbf{q}_i$ be 3-dimensional vectors of a unit-norm representing the direction of source s_i (Fig. 2a). Here, the location of

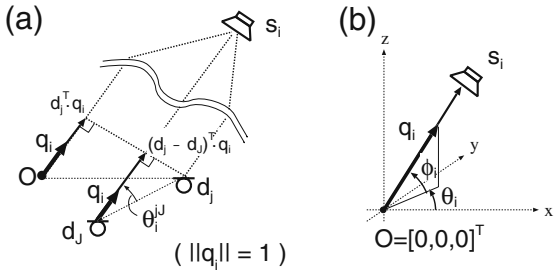


Figure 2 **a** Far-field model, **b** definition of DOA.

sensor j is given by a 3-dimensional vector $\mathbf{d}_j$. The task in this paper is to estimate the DOA $\mathbf{q}_i$ of sources from sensor observations $\mathbf{x}(f,\tau)$ and given sensor locations $\mathbf{d}_j$. Using the azimuth θ_i and elevation ϕ_i (Fig. 2b), the DOA $\mathbf{q}_i$ can be written as

$$\mathbf{q}_i = [\cos\theta_i \cos\phi_i, \sin\theta_i \cos\phi_i, \sin\phi_i]^T. \tag{5}$$

In order to estimate the DOAs of sources, we assume an anechoic model, that is, the frequency response $h_{ji}(f)$ is expressed solely by the time-delay $\tau_{ji} = \mathbf{d}_j^T \mathbf{q}_i / c$ with respect to the origin (see Fig. 2a):

$$h_{ji}(f) \approx \exp[\jmath 2\pi f c^{-1} \mathbf{d}_j^T \mathbf{q}_i], \tag{6}$$

where c is the propagation velocity of the signals. That is, we assume that the frequency response $h_{ji}(f)$ depends only on the path difference $\mathbf{d}_j^T \mathbf{q}_i$ from a source i to origin O and from a source i to a sensor j (Fig. 2a). When we consider the two sensors j and J, we obtain the following expressions:

$$\frac{h_{ji}(f)}{h_{Ji}(f)} \approx \exp[\jmath 2\pi f c^{-1} (\mathbf{d}_j - \mathbf{d}_J)^T \mathbf{q}_i] \tag{7}$$

$$= \exp[\jmath 2\pi f c^{-1} ||\mathbf{d}_j - \mathbf{d}_J|| \cos\theta_i^{jJ}]. \tag{8}$$

These two equations show that we can express the DOA in two ways: the DOA $\mathbf{q}_i$ with respect to a coordinate system, and the angle $\cos\theta_i^{jJ}$ with respect to a sensor pair j-J, (see Fig. 2a).

3 Proposed Method

This section describes our proposed DOA estimation method, which is applicable to an underdetermined case. Figure 3 shows the flow of our method, which utilizes two assumptions, namely a sparseness assumption (4) and an anechoic assumption (6).

When these two assumptions hold, the observation vector $\mathbf{x}(f,\tau)$ is the product of an unknown scalar s_k and the frequency response $\mathbf{h}_k(f)$ (see Eq. 4), where s_k is the dominant source at (f,τ), and $\mathbf{h}_k(f)$ is the

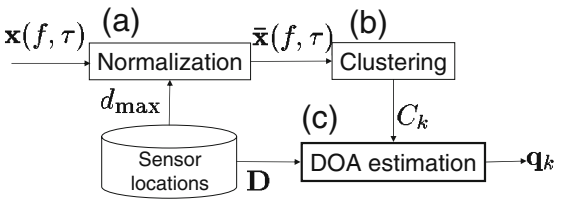


Figure 3 Flow of proposed method.

corresponding frequency response that includes the DOA information $\mathbf{q}_k$ (see Eq. 6). Based on this fact, our method first normalizes the observation vectors $\mathbf{x}(f,\tau)$ so that they are not influenced by a source s_k (Fig. 3a). In addition, by eliminating the frequency dependence in the frequency response $\mathbf{h}_k(f)$ Eq. 6, the geometric information $\mathbf{q}_k$ of the source k becomes more prominent. For such normalizations, we employ the method utilized in our previous blind sparse source separation approach [10, 11].

After being normalized, the observation vectors of all the time-frequency slots (f,τ) can be clustered based on the source geometry $\mathbf{q}_k$. Thus, the second step is the clustering of the normalized observation vectors (Fig. 3b).

Each cluster corresponds to an individual source, and each cluster centroid $\mathbf{c}_k$ contains the DOA information $\mathbf{q}_k$ as shown in Section 3.2. Therefore, in the final step (Fig. 3c), we estimate the DOAs by using the cluster centroids $\mathbf{c}_k$ and given sensor locations $\mathbf{d}_j$.

We explain each step in detail in the following subsections.

3.1 Normalization

In this step, we normalize all observation vectors $\mathbf{x}(f,\tau)$ so that they depend only on the source geometry information. Here we utilize the same observation vector normalization method as in [10, 11], and recall the normalization equation:

$$\bar{x}_j(f,\tau) \leftarrow |x_j(f,\tau)| \exp\left[\jmath \frac{\arg[x_j(f,\tau)/x_J(f,\tau)]}{\alpha_j f}\right] \tag{9}$$

where α_j is a positive constant. We utilize $\alpha_j = \alpha = 4c^{-1}d_{\max}$ in this paper (where $d_{\max}$ is the maximum distance between an arbitrary selected reference sensor J and a sensor $^\forall j \in \{1,\ldots,M\}$). The rationale for the frequency normalization with $\alpha_j = 4c^{-1}d_{\max}$ can be found in the Appendix of [11].

With Eq. 9, the inconstancy of the scalar $s_k(f,\tau)$ found in Eq. 4 is normalized by taking the ratio of two observation components $x_j(f,\tau)/x_J(f,\tau) \approx h_{jk}(f,\tau)/h_{Jk}(f,\tau)$. Note that h_{jk}/h_{Jk} is modeled as Eq. 7 and includes geometry information $\mathbf{q}_k$. The frequency normalization is realized in Eq. 9 by dividing the phase by $\alpha_j f$.

We also employ unit-norm normalization to handle the vectors on a unit-hypersphere,

$$\bar{\mathbf{x}}(f,\tau) \leftarrow \bar{\mathbf{x}}(f,\tau) \,/\, ||\bar{\mathbf{x}}(f,\tau)|| \tag{10}$$

for $\bar{\mathbf{x}}(f,\tau) = [\bar{x}_1(f,\tau),\ldots,\bar{x}_M(f,\tau)]^T$.

3.2 Clustering

If the sparseness Eq. 4 and anechoic Eq. 6 assumptions hold, each component of the normalized observation vector is expressed as

$$\bar{x}_j(f,\tau) = \frac{1}{\sqrt{M}} \exp\left[\jmath \frac{2\pi c^{-1}}{\alpha}(\mathbf{d}_j - \mathbf{d}_J)^T \mathbf{q}_k\right] \tag{11}$$

$$= \frac{1}{\sqrt{M}} \exp\left[\jmath \frac{2\pi c^{-1}}{\alpha}||\mathbf{d}_j - \mathbf{d}_J|| \cos\theta_k^{jJ}\right], \tag{12}$$

by using Eqs. 4, 6, 9, and 10. We can see that the normalized components $\bar{x}_j(f,\tau)$ keep the geometric information of a source $\mathbf{q}_k$, which is dominant at (f,τ). As a result, the normalized vectors $\bar{\mathbf{x}}(f,\tau)$ form clusters based on the source geometry.

Therefore, in the clustering step, normalized vectors $\bar{\mathbf{x}}(f,\tau)$ are clustered into N clusters $C_1,\ldots,C_N$. Note that the normalized vectors $\bar{\mathbf{x}}(f,\tau)$ are M-dimensional complex vectors, and therefore the clustering is carried out in an M-dimensional space. The clustering criterion is to minimize the total sum $\mathcal{J}$ of the squared distances between cluster members and their centroid:

$$\mathcal{J} = \sum_{k=1}^{M} \mathcal{J}_k, \quad \mathcal{J}_k = \sum_{\bar{\mathbf{x}}(f,\tau)\in C_k} ||\bar{\mathbf{x}}(f,\tau) - \mathbf{c}_k||^2. \tag{13}$$

After setting appropriate initial centroids $\mathbf{c}_k$ $(k = 1, \cdots, N)$ (see Appendix), this $\mathcal{J}$ can be minimized by the following iterative updates:

$$C_k = \left\{\bar{\mathbf{x}}(f,\tau) \mid k = \mathrm{argmin}_{k'} ||\bar{\mathbf{x}}(f,\tau) - \mathbf{c}_{k'}||^2\right\} \tag{14}$$

$$\mathbf{c}_k \leftarrow E[\bar{\mathbf{x}}(f,\tau)]_{\bar{\mathbf{x}}\in C_k}, \quad \mathbf{c}_k \leftarrow \mathbf{c}_k/||\mathbf{c}_k||, \tag{15}$$

where $E[\cdot]_{\bar{\mathbf{x}}\in C_k}$ is a mean operator for the members of a cluster C_k. That is the cluster members are determined by Eq. 14 and their centroid is calculated by Eq. 15. This minimization can be performed efficiently with the k-means clustering algorithm [23] with a given source number N.

3.3 DOA Estimation

Because each cluster corresponds to an individual source, the centroid $\mathbf{c}_k$ represents the geometry of the

source s_k. From Eqs. 11, 12 and 15, the j-th component of $\mathbf{c}_k$ is expressed as

$$\{\mathbf{c}_k\}_j \propto \boldsymbol{E}[\bar{x}_j(f,\tau)]_{\bar{\mathbf{x}}\in C_k}$$

$$= \frac{1}{\sqrt{M}} \exp\left[J \frac{2\pi c^{-1}}{\alpha} (\mathbf{d}_j - \mathbf{d}_J)^T \tilde{\mathbf{q}}_k \right] \quad (16)$$

$$= \frac{1}{\sqrt{M}} \exp\left[J \frac{2\pi c^{-1}}{\alpha} ||\mathbf{d}_j - \mathbf{d}_J|| \cos\tilde{\theta}_k^{jJ} \right] \quad (17)$$

where $\tilde{\mathbf{q}}_k$ and $\tilde{\theta}_k^{jJ}$ are the estimated DOAs. We can see that the argument of the centroid $\mathbf{c}_k$ includes the DOA information $\mathbf{q}_k$ and θ_k^{jJ} of a source s_k, e.g.,:

$$\arg\left[\{\mathbf{c}_k\}_j\right] = \frac{2\pi c^{-1}}{\alpha} (\mathbf{d}_j - \mathbf{d}_J)^T \mathbf{q}_k. \quad (18)$$

To estimate the 3-dimensional DOAs of sources $\mathbf{q}_k$, which is our goal, our proposed DOA estimation method combines information from several sensor pairs. Because our observation vector normalization is based on sensor J, we can obtain information from $M-1$ sensor pairs including sensor J. By using all the components of a centroid $\mathbf{c}_k$ and the relationship (18), such $M-1$ sensor pair information is combined as

$$\boldsymbol{r}_k = \frac{2\pi c^{-1}}{\alpha} \boldsymbol{D}\mathbf{q}_k \quad (19)$$

where

$$\boldsymbol{r}_k = [\arg[\{\mathbf{c}_k\}_1], \cdots, \arg[\{\mathbf{c}_k\}_M]]^T,$$

$$\boldsymbol{D} = [\mathbf{d}_1 - \mathbf{d}_J, \cdots, \mathbf{d}_M - \mathbf{d}_J]^T.$$

As there is no exact solution for Eq. 19, in practice we obtain the 3-dimensional DOA $\mathbf{q}_k$ in the least-square sense [24],

$$\mathbf{q}_k = \frac{\alpha}{2\pi c^{-1}} \boldsymbol{D}^+ \boldsymbol{r}_k. \quad (20)$$

where $.^+$ denotes the Moore-Penrose pseudo-inverse. As the calculation of Eq. 20 tends to include some errors, we normalize the norm of $\mathbf{q}_k$

$$\mathbf{q}_k \leftarrow \frac{\mathbf{q}_k}{||\mathbf{q}_k||} \quad (21)$$

so that its norm is unity. If rank$(\boldsymbol{D}) \geq 3$, we can estimate the 3-dimensional DOA.

Using the j-th component of the centroid $\mathbf{c}_k$, we can also estimate the DOA θ_k^{jJ} from Eq. 17 if needed:

$$\cos\tilde{\theta}_k^{jJ} = \frac{\alpha}{2\pi c^{-1}} \frac{\arg\left[\{\mathbf{c}_k\}_j\right]}{||\mathbf{d}_j - \mathbf{d}_J||}. \quad (22)$$

The previous methods with two sensors (e.g., [15]) can estimate only such a 1-dimensional DOA θ_k^{jJ} with regard to a sensor pair.

Note that if we use a cubic sensor array system $[\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4] = \gamma[(0,0,0)^T, (1,0,0)^T, (0,1,0)^T, (0,0,1)^T]$ (γ: a constant) and $J=1$, then $\arg[\{\mathbf{c}_k\}_1] \approx 0$ and $\mathbf{q}_k = \frac{\alpha}{2\pi c^{-1}} \boldsymbol{r}'_k$, where $\boldsymbol{r}'_k = [\arg[\{\mathbf{c}_k\}_2], \arg[\{\mathbf{c}_k\}_3], \arg[\{\mathbf{c}_k\}_4]]^T$. That is, we do not need $\boldsymbol{D}^+$, and DOA $\mathbf{q}_k$ can be obtained simply from $\boldsymbol{r}'_k$. This should be useful as regards a practical implementation.

In this paper, we adopt normalization by Eqs. 9 and 10. This is because we wish to utilize the same approach to normalization as in our previous separation method [10, 11]. That is, in Eq. 9, we maintain the amplitudes at all sensors, although this is not needed for DOA estimation. This information may be useful when we perform source localization, i.e., estimate the positions of sources, using a near-field model. However, for the DOA estimation, we may neglect the amplitude information of the observation vectors and employ another normalization technique, e.g.,

$$\bar{x}_j(f,\tau) \leftarrow \frac{\arg[x_j(f,\tau)/x_J(f,\tau)]}{\alpha_j f}. \quad (23)$$

The DOA can be estimated by Eqs. 20 and 21, where $\boldsymbol{r}_k = [\{\mathbf{c}_k\}_1, \cdots, \{\mathbf{c}_k\}_M]^T$. In this case, the clustering is slightly simplified: Eq. 10 is an M-dimensional complex vector, on the other hand, Eq. 23 is an M-dimensional real vector.

4 Experiments

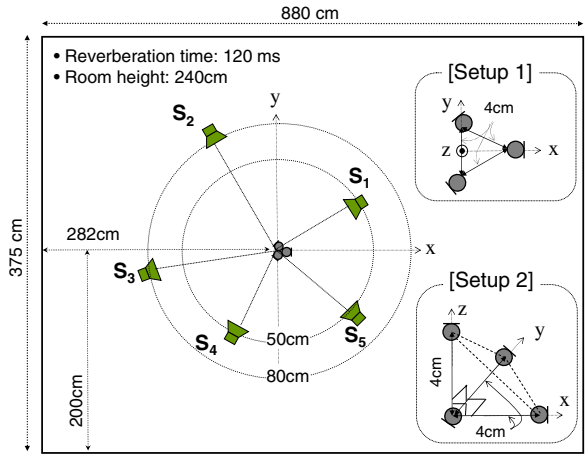
4.1 Experimental Conditions

We performed experiments in a reverberant condition. Observations were made by following Eq. 1 with the impulse responses $h_{ji}(l)$ measured in a room (Fig. 4) and 5-s English speech sources $s_i(t)$ sampled at 8 kHz. The sensor setups are shown in Fig. 4. In order to avoid the spatial aliasing problem, we utilized a sufficiently small sensor spacing (4 cm). The reverberation time of the room was RT = 120 ms. The frame size L for STFT was 512, and the frame shift was 256 ($= L/2$).

We utilized the k-means algorithm for the clustering. The number of sources N was given in this paper. The k-means algorithm is sensitive to the initial values of the centroids especially when N and M become large. How to design the initial values is discussed in the Appendix.

We estimated the DOA $\mathbf{q}_k$ with Eq. 20, and evaluated them with the azimuth θ_k and the elevation ϕ_k (see Eq. 5). The true DOA labels are shown in the tables with the label "true". We investigated eight speaker combinations and averaged the results.

(a)



(b)

[Setup 1] $(M = 3)$

$\mathbf{d}_1 = (0.0, 0.02, 0.0)^T$

$\mathbf{d}_2 = (0.0, -0.02\sqrt{3}, 0.0)^T$

$\mathbf{d}_3 = (0.04, 0.0, 0.0)^T$

[Setup 2] $(M = 4)$

$\mathbf{d}_1 = (0.0, 0.0, 0.0)^T$

$\mathbf{d}_2 = (0.04, 0.0, 0.0)^T$

$\mathbf{d}_3 = (0.0, 0.04, 0.0)^T$

$\mathbf{d}_4 = (0.0, 0.0, 0.04)^T$

Figure 4 Experimental setups with non-linear sensor arrays. **a** Room setup, **b** Sensor coordinates used in Eq. 19 and initial centroid calculations (see Appendix).

4.2 Experimental Results

Table 1 shows the results for four sources with three sensors ($M = 3$, $N = 4$), that were arranged non-linearly (Fig. 4 [Setup 1]). Here, all source heights were the same as the height of the sensor array. Because all elevations ϕ_k are zero, only the results of azimuth θ_k are shown in Table 1. We can see that the DOAs estimated with our proposed method were very close to the true values. Even when we used only 1-s data for the DOA estimation, we still obtained reasonable results as shown in Table 1. Thanks to the frequency normalization, we can handle the all the frequency components together. This allows us to utilize enough data samples and obtain good performance even if we use short observations. This applicability to short data is important e.g., for on-line implementations.

In a more complicated situation where $M = 3$ and $N = 5$, the proposed method estimated the DOAs very accurately as shown in Table 2.

We also applied our method to a 3-dimensional sensor arrangement (Fig. 4 [Setup 2]). To avoid the spatial aliasing problem, we utilized frequency bins up to 3100 Hz in this setup. In this case, the sources had different heights, and therefore, we estimated both azimuths θ_k and elevations ϕ_k. Table 3 shows results for five sources with four sensors ($M = 4$, $N = 5$). As regards azimuths θ_k, although the estimation error was sometimes greater than the result for $M = 3$, $N = 4$ (Table 1), we still obtained reasonable results for such a complicated case. The elevation values ϕ_k were also very close to the true values. We can say that our proposed method can be applied to such a 3-dimensional DOA estimation and that it gives us fairly precise DOAs with a reverberation time RT of 120 ms.

Table 1 Experimental results for $M = 3$, $N = 4$ (Setup 1).

Source	s_1	s_2	s_3	s_4
True	24°	117°	217°	311°
Proposed (5 sec.)	25°	114°	214°	313°
Proposed (1 sec.)	23°	112°	212°	318°

Table 2 Experimental results for $M = 3$, $N = 5$ (Setup 1).

Source	s_1	s_2	s_3	s_4	s_5
True	24°	117°	176°	217°	311°
Proposed (5 sec.)	23°	112°	175°	218°	314°

4.3 Comparison with MUSIC Method in Overdetermined Scenario

To show the effectiveness of our proposed method even for a situation where the MUSIC algorithm can be applied, we performed experiments for two-source three-sensor (Fig. 4 Setup 1, $M = 3$, $N = 2$) cases with both methods. In this paper, the MUSIC method was applied to each frequency bin, where a spatial correlation matrix $E[\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)]$ for the MUSIC was calculated using all the five-second data $E[\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)] = \frac{1}{T}\sum_{\tau=1}^{T}\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)$, where T is the number of frames, and the estimated DOAs at all frequency bins were clustered and averaged. Figure 5 shows the resolution of both methods, and Table 4 shows the estimated DOA θ_k. Figure 5a and c are example MUSIC spectra at a frequency f of 1844 Hz, and (b) and (d) are DOA histograms for members of each cluster (the DOA of

Table 3 Experimental results for $M = 4$, $N = 5$ (Setup 2).

Source		s_1	s_2	s_3	s_4	s_5
True	θ	31°	85°	133°	222°	302°
	ϕ	−26°	6°	30°	39°	−8°
Proposed (5 sec.)	θ	30°	79°	132°	221°	298°
	ϕ	−22°	7°	28°	35°	−9°

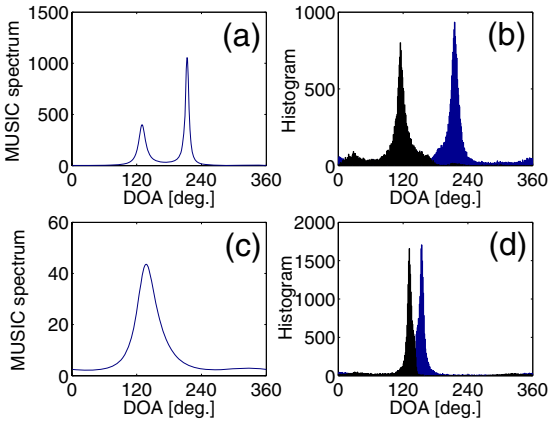


Figure 5 Resolution of MUSIC (**a**),(**c**) at $f = 1844$ Hz and the proposed method (**b**),(**d**). (**a**)(**b**): when sources are far apart ($\theta_1 = 117°$, $\theta_2 = 217°$), (**c**)(**d**): when sources are close together ($\theta_1 = 132°$, $\theta_2 = 154°$).

each member can be calculated with Eq. 20 using $\bar{\mathbf{x}}(f, \tau)$ instead of the centroid $\mathbf{c}_k$). It should be noted that the MUSIC spectra at $f = 1844$ Hz in Fig. 5 are just examples. If we plot the averaged MUSIC spectra over all frequencies, the spectra become duller than that in Fig. 5a and c.

When two sources were placed with a wide spacing (Fig. 5a and b, and Table 4 "far apart"), both MUSIC and the proposed method estimated the directions well enough. In contrast, when the two sources were close to each other (Fig. 5c and d, and Table 4 "close together"), MUSIC failed to estimate the two directions, whereas the proposed method was still successful. We consider that our proposed method with the sparseness assumption has a high resolution, although this resolution depends on the sparseness of the source signals and is affected by the room reverberation condition.

Note that we may be able to use the sparseness for the MUSIC method by using spatial correlation matrix $E[\mathbf{x}(f, \tau)\mathbf{x}^H(f, \tau)] = \mathbf{x}(f, \tau)\mathbf{x}^H(f, \tau)$ at each time-frequency point. However, in the scenario of Fig. 5c and d where two speakers were speaking simultaneously, we could not obtain sufficient resolution due to a large variance of DOA estimate values. This large variance come from the singular value decomposition (SVD) for the instantaneous value of the spatial correlation matrix. Moreover, such a time-frequency MUSIC approach is time-consuming because it requires the SVD for each time-frequency point. When we applied our proposed method, MUSIC in this section, and the time-frequency MUSIC to the same data as Fig. 5, the calculation times were 3.4, 0.42 and 245 (sec.), respectively (Calculation was done by MATLAB7.1 Service Pack 3 with a PC, Intel Xeon 2.66GHz (Quad core) × 2 with a Linux OS).

Table 4 Experimental results for $M = 3$, $N = 2$ (Setup 1).

	Far apart		Close together	
Source	s_1	s_2	s_1	s_2
True	117°	217°	132°	154°
Proposed (5 sec.)	114°	217°	128°	156°
MUSIC	125°	217°	126°	345°

4.4 Performance Under Different Reverberant Conditions

We have shown that our proposed method worked well in weak reverberant conditions of RT = 120 ms. Now we should remember that our proposed method employs the source sparseness Eq. 4 and anechoic Eq. 6 assumptions. In practice, however, these assumptions hardly hold due to reverberation. To study the effects of reverberation, this section investigates the performance of our approach in different reverberant conditions.

We performed experiments under an anechoic condition and some reverberant conditions. For the reverberant tests, observations were simulated by following Eq. 1 with impulse responses h_{jk} measured in a room (Fig. 6). The room reverberation times RTs were 128 and 300 ms. For both RTs, we utilized the same room but changed the wall condition. We also changed the distance R between the sensors and sources. The distances were $R = 50$, 110, and 170 cm (see Fig. 6). Here, we only tested the 3-microphone and 4-source case. We investigated eight speaker combinations and averaged the results.

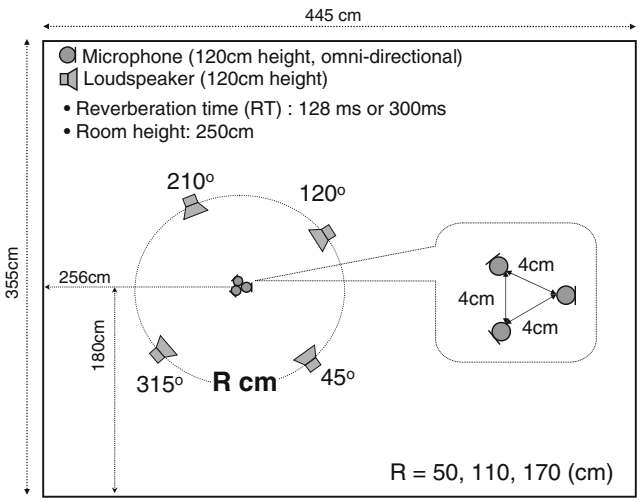


Figure 6 Experimental setup for different reverberations.

First, we investigated how the sparseness Eq. 4 and anechoic Eq. 6 assumptions hold under each condition. To check the sparseness, we evaluated the approximate W-disjoint orthogonality [8, 25] of the reverberant speech signals:

$$r_k(z) = \frac{\sum_{(f,\tau)} |\Phi_{(k,z)}(f,\tau) x_{1k}(f,\tau)|^2}{\sum_{(f,\tau)} |x_{1k}(f,\tau)|^2} \times 100[\%]. \quad (24)$$

In Eq. 24, $x_{1k}(f,\tau)$ means the short-time Fourier transformed observed signal k at sensor 1: $x_{1k}(f,\tau) = \mathrm{STFT}\left[\sum_l \mathrm{h}_{1k}(l)\mathrm{s}_k(t-l)\right]$. Moreover, in Eq. 24, $\Phi_{(k,z)}$ is a time-frequency binary mask that has a parameter z

$$\Phi_{(k,z)}(f,\tau) = \begin{cases} 1 & 20\log_{10}(|x_{1k}(f,\tau)|/|\hat{x}_{1k}(f,\tau)|) > z \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

where $\hat{x}_{1k}(f,\tau)$ is the sum of the interference components at sensor 1: $\hat{x}_{1k}(f,\tau) = \mathrm{STFT}\left[\sum_{i=1, i\neq k}^{N} \mathrm{x}_{1i}(t)\right]$. The approximate WDO $r_k(z)$ indicates the percentage of the energy of source k for time-frequency points where it dominates the other sources by z dB. A larger (smaller) approximate WDO $r_k(z)$ means more (less) sparseness.

For the anechoic measure, we adopted the clarity index [26]:

$$C = 10\log_{10} \frac{\int_0^{80\mathrm{ms}} h^2(t)dt}{\int_{80\mathrm{ms}}^{\infty} h^2(t)dt} [\mathrm{dB}].$$

The clarity index describes the ratio between direct sound and reverberant sound. A small (large) C means the reverberant sound (direct sound) is large. In general, DOA estimation is difficult when the direct sound component is small.

Figure 7 shows the average approximate W-disjoint orthogonality for $z = 10$ [dB] and the average clarity index under each condition. From Fig. 7, we can see that the sparseness decreases when the contribution of the direct sound is small. We can also see that the clarity

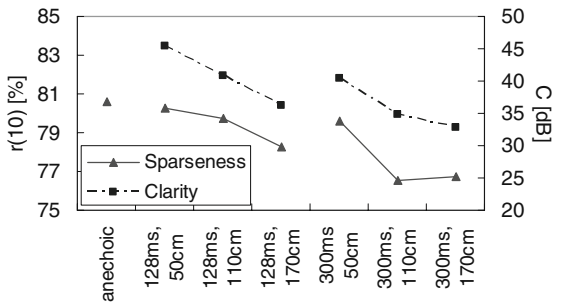


Figure 7 Sparseness measure (W-disjoint orthogonality, *left scale*) and reverberant measure (clarity, *right scale*).

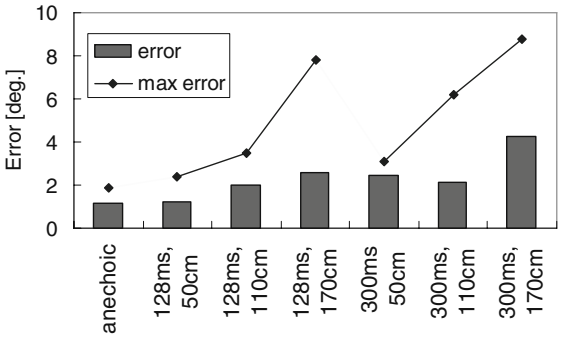


Figure 8 DOA estimation error in degrees.

C becomes small as the reverberation and distance R increase. That is, when the reverberation is long and R is large, the sparseness and anechoic assumptions seem to become corrupted.

Next, we checked the DOA estimation performance in different reverberation conditions. The DOA estimation results for each condition are shown in Fig. 8. The figure plots the estimation error

$$\mathrm{Error}_k = |\theta_k - \hat{\theta}_k|$$

where $\hat{\theta}_k$ represents the true directions (azimuths). The average and maximum errors are shown in Fig. 8. We can see that the DOA estimation error increases when the reverberation and distance R are large. However, the maximum error is still less than 10 degrees.

5 Conclusion

We proposed a new DOA estimation method for underdetermined cases by assuming source sparseness. The method is based on the normalization and clustering of the observation vectors. We obtained promising experimental results for underdetermined cases in a reverberant condition. We also confirmed that our proposed method provides higher resolution when estimating the directions of sources than the MUSIC algorithm.

We also reported the performance under some reverberant conditions, where the sparseness and anechoic assumptions were deteriorating. From the results, we saw that the DOA estimation error is still not very large even under difficult reverberant conditions.

Appendix

This appendix explains how to design the initial values for the k-means algorithm used in the clustering stage.

Because the k-means algorithm is sensitive to the initial values of its centroids, it is preferable to set appropriate initial centroids. In the paper, we designed the initial centroids as follows:

- Set microphone locations $\mathbf{d}_j$ for each setup (see Fig. 4b)
- Calculate the initial directions $\mathbf{q}_i$ so that they were as scattered as possible. Concretely, we utilized Eq. 5 where $\theta_i = \frac{2\pi}{N} \times i \ (i = 1, \cdots, N)$ for $M \geq 3$ and $\frac{\pi}{N} \times i \ (i = 1, \cdots, N)$ for $M = 2$. $\phi_i = 0$ for all i.
- With above $\mathbf{d}_j$ and $\mathbf{q}_i$, we calculated the initial centroid by using Eq. 16.

Note that these initial values of $\mathbf{d}_j$ and $\mathbf{q}_i$ do not have to be exactly the same as the sensor and source setups.

References

1. Pillai, S. U. (1989). *Array signal processing.* New York: Springer.
2. Brandstein, M., & Ward, D. (Eds.) (2001). *Microphone arrays.* New York: Springer.
3. Van Trees, H. L. (Ed.) (2002). *Optimum array processing.* New York: Wiley.
4. Schmidt, R. O. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation 34*, 276–280.
5. Sawada, H., Mukai, R., Araki, S., & Makino, S. (2005). Frequency-domain blind source separation. In Benesty, J., Makino, S., Chen, J. (Eds.), *Speech enhancement* pp. 299–327. New York: Springer.
6. Sawada, H., Mukai, R., Araki, S., & Makino, S. (2005). Multiple source localization using independent component analysis. In *Proc. AP-S/URSI2005.*
7. Lombard, A., Rosenkranz, T., Buchner, H., & Kellermann, W. (2008). Exploiting the self-steering capability of blind source separation to localize two or more sound sources in adverse environments. In *Proc. ITG conference on speech communication.*
8. Yılmaz, Ö., & Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on SP 52*(7), 1830–1847.
9. Bofill, P., & Zibulevsky, M. (2000). Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform. In *Proc. ICA2000* (pp. 87–92).
10. Araki, S., Sawada, H., Mukai, R., & Makino, S. (2005). A novel blind source separation method with observation vector clustering. In *Proc. 2005 international workshop on acoustic echo and noise control (IWAENC 2005)* (pp. 117–120).
11. Araki, S., Sawada, H., Mukai, R., & Makino, S. (2007). Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Processing 87*, 1833–1847.
12. Zhang, Y., Mu, W., & Amin, M. G. (1999). Maximum-likelihood methods for array processing based on time-frequency distributions. In *Proc. SPIE* (Vol. 3807, 502–513).
13. Zhang, Y., Mu, W., & Amin, M. G. (2000). Time-frequency maximum likelihood methods for direction finding. *Journal of the Franklin Institute 337*(4), 483–497.
14. Burintramart, S., Sarkar, T. K., Zhang, Y., & Salazar-Palma, M. (2007). Nonconventional least squares optimization for DOA estimation. *IEEE Transactions on Antennas and Propagation 55*(3), 707–714.
15. Rickard, S., & Dietrich, F (2000). DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET. In: *Proc. SSAP2000* (pp. 311–314).
16. Shamsunder, S., & Giannakis, G. B. (1993). Modeling of non-Gaussian array data using cumulants: DOA estimation of more sources with less sensors. *Signal Processing 30*, 279–297.
17. Mitianoudis, N., & Stathaki, T. (2007). Batch and online underdetermined source separation using Laplacian mixture model. *IEEE Transactions on Audio, Speech, and Language Processing 15*(6), 1818–1832.
18. Matsuo, M., Hioka, Y., & Hamada, N. (2005). Estimating DOA of multiple speech signals by improved histogram mapping method. In *Proc. IWAENC2005* (pp. 129–132).
19. Karbasi, A., Sugiyama, A. (2006). DOA estimation method for an arbitrary triangular microphone arrangement. In *Proc. EUSIPCO2006.*
20. Araki, S., Sawada, H., Mukai, R., & Makino, S. (2006). DOA estimation for multiple sparse sources with normalized observation vector clustering. In *Proc. ICASSP2006* (pp. 33–36).
21. Araki, S., Sawada, H., Mukai, R., & Makino, S. (2006). Performance evaluation of sparse source separation and DOA estimation with observation vector clustering in reverberant environments. In *Proc. IWAENC2006.*
22. Aoki, M., Okamoto, M., Aoki, S., Matsui, H., Sakurai, T., & Kaneda, Y. (2001). Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. *Acoustical Science and Technology 22*, 149–157.
23. Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*, 2nd edn. New York: Wiley Interscience
24. Mukai, R., Sawada, H., Araki, S., & Makino, S. (2004). Frequency domain blind source separation for many speech signals. In: *Proc. ICA 2004. LNCS 3195* (pp. 461–469)
25. Rickard, S., & Yılmaz, Ö (2002). On the approximate W-disjoint orthogonality of speech. In *Proc. ICASSP2002* (Vol. I, pp. 529–532)
26. ISO 3382 (1997). Acoustics—Measurement of the reverberation time of rooms with reference to other acoustical parameters.

Shoko Araki is with NTT Communication Science Laboratories, NTT Corporation, Japan. She received the B.E. and the M.E. degrees from the University of Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D degree from Hokkaido University, Japan in 2007.

Since she joined NTT in 2000, she has been engaged in research on acoustic signal processing, array signal processing, blind source separation applied to speech signals, meeting diarization and auditory scene analysis.

She is a member of the Organizing Committee of the ICA 2003, the Finance Chair of IWAENC 2003, and the Registration Chair of WASPAA 2007. She received the 19th Awaya Prize from Acoustical Society of Japan (ASJ) in 2001, the Best Paper Award of the IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Academic Encouraging Prize from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2006, and the Itakura Prize Innovative Young Researcher Award from (ASJ) in 2008.

She is a member of the IEEE, IEICE, and the ASJ.

Hiroshi Sawada received the B.E., M.E. and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1991, 1993 and 2001, respectively.

He joined NTT in 1993. From 1993 to 2000, he was engaged in research on the computer aided design of digital systems, logic synthesis, and computer architecture. In 2000, he stayed at the Computation Structures Group of MIT for six months. Since 2000, he has been engaged in research on signal processing, microphone array, and blind source separation. He is now the group leader of Learning and Intelligent Systems Research Group at the NTT Communication Science Laboratories.

From 2006 to 2009, he served as an associate editor of the IEEE Transactions on Audio, Speech & Language Processing. He is a member of the Audio and Electroacoustics Technical Committee of the IEEE SP Society. He was a tutorial speaker at ICASSP2007. He served as the publications chairs of the WASPAA~2007 in Mohonk, and served as an organizing committee member for ICA2003 in Nara and the communications chair for IWAENC2003 in Kyoto.

He is the author or co-author of three book chapters, more than 25 journal articles, and more than 100 conference papers. He received the 9th TELECOM System Technology Award for Student from the Telecommunications Advancement Foundation in 1994, and the Best Paper Award of the IEEE Circuit and System Society in 2000.

Dr. Sawada is a senior member of the IEEE, a member of the IEICE and the ASJ.

Ryo Mukai received the B.S. and the M.S. degrees in information science from the University of Tokyo, Japan, in 1990 and 1992, respectively. He joined NTT in 1992. From 1992 to 2000, he was engaged in research and development of processor architecture for network service systems and distributed network systems.

Since 2000, he has been with NTT Communication Science Laboratories, where he was engaged in research of blind source separation. Since 2006, he has been engaged in research and development of audio and video retrieval system. He received the Sato Paper Award of Acoustical Society of Japan (ASJ) in 2005 and the Paper Award of the IEICE in 2005.

He is a senior member of the IEEE, a member of the ACM, ASJ, IEICE, and IPSJ.

Shoji Makino received the B. E., M. E., and Ph. D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively.

He joined NTT in 1981. He is now a Professor at University of Tsukuba. His research interests include adaptive filtering technologies and realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech.

He received the ICA Unsupervised Learning Pioneer Award in 2006, the IEEE MLSP Competition Award in 2007, the TELECOM System Technology Award of the TAF in 2004, the Achievement Award of the IEICE in 1997, the Outstanding Technological Development Award of the ASJ in 1995, the Paper Award of the IEICE in 2005 and 2002, the Paper Award of the ASJ in 2005 and 2002. He is the author or co-author of more than 200 articles in journals and conference proceedings and is responsible for more than 150 patents. He was a Keynote Speaker at ICA2007 and a Tutorial speaker at ICASSP2007.

He is a member of the Award Committee of the IEEE James L. Flanagan Speech & Audio Processing Award. He is a member of the Awards Board and the Conference Board of the IEEE SP Society. He is an Associate Editor of the IEEE Transactions

on Speech and Audio Processing and an Associate Editor of the EURASIP Journal on Applied Signal Processing. He is the Chair of the Technical Committee on Blind Signal Processing of the IEEE CAS Society and a member of the Technical Committee on Audio and Electroacoustics of the IEEE SP Society. He was the Chair of the Technical Committee on Engineering Acoustics of the IEICE and the ASJ. He is a member of the International ICA Steering Committee and a member of the International IWAENC Standing committee.

He is the General Chair of the WASPAA2007, the General Chair of the IWAENC2003, the Organizing Chair of the ICA2003, and is the designated Plenary Chair of ICASSP2012.

He is an IEEE Fellow, an IEICE Fellow, a council member of the ASJ, and a member of the EURASIP.