Acoust. Sci. & Tech. **26**, 6 (2005)

PAPER

Direction of arrival estimation using matching pursuit and its application to source separation for convolved mixtures

Yasuhiro Oikawa* and Yoshio Yamasaki†

Global Information and Telecommunication Institute, Waseda University, 1011 Okuboyama, Nishi-Tomida, Honjo, 367–0035, Japan

(*Received 2 August 2004, Accepted for publication 14 March 2005*)

Abstract: In this paper, we describe a new source separation method in which uses spatial information derived from the direction of arrival (DOA) estimates of each direct and reflected sound issued. The method we propose has the following steps: (1) each DOA is estimated using matching pursuit and reoptimized after each new DOA is estimated, (2) using these DOA estimates, the mixing matrix is also estimated and the inverse of the mixing matrix is used to separate the mixture signals. In our experiments, we obtained a better signal separation with the new method than with the conventional frequency-domain independent component analysis (ICA)-based source separation method.

Keywords: DOA estimation, Source separation, Delay-and-Sum array, Matching pursuit algorithm, DOA reoptimization

PACS number: 43.38.Hz, 43.60.Gk, 43.60.Pt [DOI: 10.1250/ast.26.486]

1. INTRODUCTION

Source separation refers to signal processing techniques aimed at recovering sources from observations of unknown mixtures of unknown sources. It has applications in hands-free communication devices and hearing aids. Consequently, source separation has recently been studied by many researchers. Many methods based on independent component analysis (ICA) have been proposed [1–8]. ICA is used to estimate the unmixing matrix and to separate mixture signals into independent components, assuming that the source signals are independent. In [7,8] spatial information, that is, the direction of arrival (DOA) of direct sounds, was included in the frequency-domain ICA to avoid permutations of the separated frequency components. This was further exploited in [9], where the spatial information (source locations) was used to form constraints on the possible inverse systems that result in an independent source.

Non-ICA-based source separation methods have also been proposed [10,11]. Sparse decomposition with matching pursuit and its application to source separation was proposed by Gribonval [11] who computed the sparse decomposition of stereo audio signals with a matching pursuit algorithm and found that the parameters of the atoms of decomposition were clustered. Estimates of sources were then recovered by partial reconstruction using only the appropriate atoms of decomposition. For instantaneous mixtures or for convolved mixtures consisting of short impulse responses, these methods are very effective in separating sources. However, in the case where the mixture is a convolved mixture and impulse responses are long, which is common in the real world, they perform poorly and the separation is insufficient [12].

A source separation method based purely on spatial information is conventional beam former (see, e.g., [13]), which suppresses signals that are impinging on a microphone array from undesirable directions. The conventional beam former is often insufficient for separating signals, since, in a real environment, the transfer function between source and microphone includes many reflective pathways. However, if we can estimate all reflected sounds, it should be possible to estimate the complete impulse response and to separate sounds. Because it is difficult to estimate all reflected sounds, it would be more effective to consider the spatial information and estimate direct and main reflected sounds to establish a source separation system.

In this paper, we propose a source separation method in which spatial information derived from the results of DOA estimates for not only direct sounds but also early reflected sounds is used. We need to find many DOAs to estimate the mixing system. However it is impossible to find true DOAs

*e-mail: yoikawa@waseda.jp
†e-mail: y-yamasaki@waseda.jp

using conventional beam forming techniques when the number of sources exceeds that of microphones. We suggest a new DOA estimation technique, which is to use a matching pursuit algorithm, by which it becomes possible to find true DOAs even if the number of sources exceeds that of microphones. The basic outline of our algorithm is as follows. We first find the normalized power of the array output, $P(\theta)$, as a function of the DOA, θ. Then for the DOA estimation of the direct and indirect (reflected) signals, we apply a matching pursuit algorithm that includes a reoptimization of the directions of arrival at each iteration step. The main purpose of the matching pursuit step is to find the filter coefficients of our FIR model of the channel. By including the reoptimization step in the matching pursuit, we are able to avoid false DOA estimates, which are a problem when using the conventional beam former in an environment of multiple sources. The sounds coming from different DOAs are then classified into a small set of sources. From the classified DOAs, we form estimates of the impulse responses for each source and microphone combination. The separated source signals are obtained by filtering the observations with the inverse of the mixing matrix estimate.

We compared our method with the conventional frequency-domain ICA-based source separation method [7] using two sources, two microphones, and a convolved mixture. In our experiments, we obtained better signal separation for the new method than that for conventional frequency-domain ICA-based source separation.

2. DOA ESTIMATION

Our DOA estimation consists of the following steps. We first separately calculate the normalized power of the array output, $P(\theta)$, for each frequency bin using the Delay-and-Sum method [14]. We then average $P(\theta)$ over all frequency bins. Finally, we perform peak picking using a matching pursuit algorithm to estimate the DOA over all frequency bands. The matching pursuit algorithm includes, after each iteration step, a reoptimization of all DOAs found thus far. Its main characteristic is that it is possible to find true DOAs when the number of sources exceeds that of microphones. We will discuss these steps in more detail in the following subsections.

2.1. Calculation of Power of Array Output

The power of the Delay-and-Sum array output is calculated as

$$P(\theta) = \boldsymbol{d}(\theta)^{\mathrm{H}} \boldsymbol{R} \boldsymbol{d}(\theta), \tag{1}$$

where $\boldsymbol{d}(\theta)$ is the steering vector:

$$\boldsymbol{d}(\theta) = [1, \exp(-j\omega\tau), \cdots, \exp(-j\omega(M-1)\tau)]^{\mathrm{T}}. \tag{2}$$

Here, ω is the angular frequency, M is the number of microphones, $\tau = d \sin\theta / c$, d is the distance between microphones, c is the velocity of sound, and $\boldsymbol{R}$ is the covariance matrix of the array outputs $\boldsymbol{x}(t)$ i.e.:

$$\boldsymbol{R} = E[\boldsymbol{x}(t)\boldsymbol{x}(t)^{\mathrm{H}}]. \tag{3}$$

For K sounds (i.e., K different DOAs) and two microphones, the observed signals are

$$\boldsymbol{X}(\omega, t) = \begin{bmatrix} X_1(\omega, t) \\ X_2(\omega, t) \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{K} H_{1k}(\omega) S_k(\omega, t) \\ \sum_{k=1}^{K} H_{2k}(\omega) S_k(\omega, t) \end{bmatrix}, \tag{4}$$

where t is the time, H_{1k} and H_{2k} are the respective transfer functions between the kth sound and each microphone, and $S_k(\omega, t)$ is the kth sound. The covariance matrix is

$$\boldsymbol{R}(\omega) = E[\boldsymbol{X}(\omega, t)\boldsymbol{X}(\omega, t)^{\mathrm{H}}] = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}, \tag{5}$$

$$r_{11} = E\left[\left|\sum_{k=1}^{K} H_{1k} S_k\right|^2\right], \tag{6}$$

$$r_{12} = E\left[\sum_{k=1}^{K} H_{1k} H_{2k}^* |S_k|^2 + \sum_{l \neq k} H_{1k} H_{2l}^* S_k S_l^*\right], \tag{7}$$

$$r_{21} = r_{12}^*, \text{ and} \tag{8}$$

$$r_{22} = E\left[\left|\sum_{k=1}^{K} H_{2k} S_k\right|^2\right], \tag{9}$$

and the steering vector is

$$\boldsymbol{d}(\theta, \omega) = [1, \exp(-j\omega\tau)]^{\mathrm{T}}. \tag{10}$$

The power of array output is then

$$\begin{aligned} P(\theta, \omega) &= \boldsymbol{d}(\theta, \omega)^{\mathrm{H}} \boldsymbol{R}(\omega) \boldsymbol{d}(\theta, \omega) \\ &= r_{11} + r_{22} + 2 \cdot \Re\{r_{12} \exp(-j\omega\tau)\}, \end{aligned} \tag{11}$$

where $\Re$ indicates the real component.

The first and second terms of Eq. (11) do not depend on θ and we only need to consider the third term. The third term of Eq. (11), $\hat{P}(\theta, \omega)$, is

$$\begin{aligned} \hat{P}(\theta, \omega) &= P(\theta, \omega) - E[|X_1|^2] - E[|X_2|^2] \\ &= 2\Re\left[E\left[\sum_{k=1}^{K} H_{1k} H_{2k}^* |S_k|^2 \right.\right. \\ &\qquad \left.\left. + \sum_{l \neq k} H_{1k} H_{2l}^* S_k S_l^*\right] \exp(-j\omega\tau)\right]. \end{aligned} \tag{12}$$

Therefore, the average of $\hat{P}(\theta, \omega)$ over the frequency bins is

$$\hat{P}_{\mathrm{avrg}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \hat{P}(\theta, \omega_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{K} 2\cdot E[|S_k(\omega_i)|^2] \cdot \Re\{H_{1k}(\omega_i)H_{2k}(\omega_i)^* \exp(-j\omega_i\tau)\}\right] + \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{l\neq k} 2\cdot \Re\{H_{1k}(\omega_i)H_{2l}(\omega_i)^* \cdot E[S_k(\omega_i)S_l(\omega_i)^*]\exp(-j\omega_i\tau)\}\right], \quad (13)$$

where N is the number of frequency bins. Since $E[S_k(\omega)S_l(\omega)^*]$ is generally smaller for $k \neq l$ than for $k = l$, we have assumed that the second term in Eq. (13) can be set to zero. We can then rewrite Eq. (13) as

$$\hat{\boldsymbol{P}}_{\text{avrg}}(\theta) \approx \sum_{k=1}^{K} \hat{\boldsymbol{P}}_{\text{avrg}}(\theta|\theta_k), \quad (14)$$

where $\hat{\boldsymbol{P}}_{\text{avrg}}(\theta|\theta_k)$ is the frequency average of the θ-dependent component of the array output power from the kth sound, i.e., the θ-dependent component of the Delay-and-Sum array output is approximately the sum of that for each source.

As we are only interested in finding the DOAs at this point, we let $H_{1k}(\omega) = \exp(-j\omega\tau_{1k})$ and $H_{2k}(\omega) = \exp(-j\omega\tau_{2k})$. Thus, the frequency average from the kth sound, $\hat{\boldsymbol{P}}_{\text{avrg}}(\theta|\theta_k)$, becomes

$$\hat{\boldsymbol{P}}_{\text{avrg}}(\theta|\theta_k) = \frac{2E[|S_k(\omega_i)|^2]}{N} \cdot \sum_{i=1}^{N} \Re\{\exp(-j\omega_i(\tau_{1k} - \tau_{2k})) \cdot \exp(-j\omega_i\tau)\}, \quad (15)$$

$$\tau_{1k} - \tau_{2k} = \frac{d\sin\theta_k}{c}, \text{ and} \quad (16)$$

$$\tau = \frac{d\sin\theta}{c}, \quad (17)$$

where θ_k is the true direction of the kth sound position and θ is the steering direction.

2.2. Matching Pursuit to Estimate DOA

A matching pursuit algorithm was introduced to decompose any signal into a linear expansion of waveforms [15]. We used a modified matching pursuit algorithm that includes a reoptimization step [16] to decompose the signal into a set of direct and reflected sounds. We define the vector of the angles of i DOAs that is estimated during i iterations as

$$\Theta_i = [\hat{\theta}_1, \cdots, \hat{\theta}_i]^{\mathrm{T}}, \quad (18)$$

where Θ_0 is a vector without any elements. The matching pursuit algorithm for DOA estimation consists of the following steps.

Step 1) Define a dictionary as

$$\mathcal{D} = \{\hat{\boldsymbol{P}}_{\text{avrgn}}(\theta|\theta_k)\}_{-\pi/2<\theta_k<\pi/2}, \quad (19)$$

i.e., an element of family $\mathcal{D}$ is defined as Eq. (15) normalized by its norm:

$$\hat{\boldsymbol{P}}_{\text{avrgn}}(\theta|\theta_k) = \frac{\hat{\boldsymbol{P}}_{\text{avrg}}(\theta|\theta_k)}{\sqrt{\frac{1}{\pi}\int_{-\pi/2}^{\pi/2} |\hat{\boldsymbol{P}}_{\text{avrg}}(\theta|\theta_k)|^2 d\theta}}. \quad (20)$$

Step 2) Initialization:

$$e_0(\theta) = \hat{\boldsymbol{P}}_{\text{observed}}(\theta) \text{ and} \quad (21)$$

$$i = 1. \quad (22)$$

Step 3) Calculate the residual for all θ_k:

$$e_i(\theta|\theta_k) = e_{i-1}(\theta) - a_{i-1}(\theta_k)\hat{\boldsymbol{P}}_{\text{avrgn}}(\theta|\theta_k), \quad (23)$$

where $a_{i-1}(\theta_k)$ denotes the inner product of $e_{i-1}(\theta)$ and $\hat{\boldsymbol{P}}_{\text{avrgn}}(\theta|\theta_k)$.

Step 4) Select θ_k (estimate DOA $\hat{\theta}_i$):

$$\hat{\theta}_i = \underset{\theta_k}{\text{argmin}} \sum |e_i(\theta|\theta_k)|^2. \quad (24)$$

Step 5) Reoptimize Θ_i (all DOAs) and calculate the residual $e_i(\theta)$:

$$e_i(\theta) = e_0(\theta) - \sum_{l=1}^{i} \hat{a}(\hat{\theta}_l)\hat{\boldsymbol{P}}_{\text{avrgn}}(\theta|\hat{\theta}_l), \quad (25)$$

where $\hat{a}(\hat{\theta}_l)$ is computed using Eq. (32).

Step 6) If

$$10\log\frac{\int {e_0}^2(\theta)d\theta}{\int {e_i}^2(\theta)d\theta} < \delta, \quad (26)$$

where δ is the stopping criterion, and

$$i = i + 1, \quad (27)$$

return to Step 3), or else end the procedure.

2.3. Reoptimization of DOAs

A high-quality, consistent analysis-synthesis method with reoptimization of amplitude and frequency parameters in sinusoidal coding has been described by Vos *et al.* [16]. They presented techniques for the optimization of sinusoidal parameters based on the squared difference between the input signal and reconstruction. Here, we use a similar method in order to reoptimize the DOAs with a gradient

algorithm. We describe optimization techniques of DOAs based on the squared difference between the array output and reconstruction of components for estimated DOAs.

We define the vector of the angles of L DOAs as

$$\Theta = [\theta_1, \cdots, \theta_L]^{\mathrm{T}}. \tag{28}$$

The basis vectors and the observed vector are defined as

$$\hat{\boldsymbol{p}}_{\mathrm{avrgn}}(\theta_k) = \left[\hat{P}_{\mathrm{avrgn}}\left(-\frac{\pi}{2} \,\middle|\, \theta_k\right), \cdots, \hat{P}_{\mathrm{avrgn}}\left(\frac{\pi}{2} \,\middle|\, \theta_k\right)\right]^{\mathrm{T}} \text{ and} \tag{29}$$

$$\boldsymbol{e}_0 = \left[e_0\left(-\frac{\pi}{2}\right), \cdots, e_0\left(\frac{\pi}{2}\right)\right]^{\mathrm{T}}, \tag{30}$$

where we discretized the normalized frequency average of the power of array output of the kth sound as a function of the continuous steering direction variable θ.

For a given set of DOAs, the analysis matrix containing the basis vectors is constructed according to

$$\hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta} = [\hat{\boldsymbol{p}}_{\mathrm{avrgn}}(\theta_1), \cdots, \hat{\boldsymbol{p}}_{\mathrm{avrgn}}(\theta_L)]. \tag{31}$$

The projection of $\boldsymbol{e}_0$ onto a space that is defined by bases $\hat{\boldsymbol{p}}_{\mathrm{avrgn}}(\theta_1), \cdots$, and $\hat{\boldsymbol{p}}_{\mathrm{avrgn}}(\theta_L)$ is

$$\hat{\boldsymbol{a}} = (\hat{\boldsymbol{P}}^{\mathrm{T}}_{\mathrm{avrgn}\Theta} \cdot \hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta})^{-1} \cdot \hat{\boldsymbol{P}}^{\mathrm{T}}_{\mathrm{avrgn}\Theta} \cdot \boldsymbol{e}_0, \tag{32}$$

which is from the least-squares residual. Optimum DOAs are those for which the energy of the projection of $\boldsymbol{e}_0$ onto the column space of $\hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta}$ is maximized:

$$\begin{aligned} &\underset{\Theta}{\mathrm{argmax}}\{\hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta} \cdot \hat{\boldsymbol{a}}\}^{\mathrm{T}} \cdot \{\hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta} \cdot \hat{\boldsymbol{a}}\} \\ &\quad = \underset{\Theta}{\mathrm{argmax}}\, \boldsymbol{e}_0^{\mathrm{T}} \boldsymbol{P}_{\Theta} \boldsymbol{e}_0, \end{aligned} \tag{33}$$

where we define the projection matrix as

$$\boldsymbol{P}_{\Theta} = \hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta} \cdot (\hat{\boldsymbol{P}}^{\mathrm{T}}_{\mathrm{avrgn}\Theta} \cdot \hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta})^{-1} \cdot \hat{\boldsymbol{P}}^{\mathrm{T}}_{\mathrm{avrgn}\Theta}. \tag{34}$$

We used a gradient search based on Newton's method to find the local maximum of $\boldsymbol{e}_0^{\mathrm{T}} \boldsymbol{P}_{\Theta} \boldsymbol{e}_0$ in the neighborhood of a set of initial DOAs. At iteration m, Newton's method defines the updated DOAs as

$$\Theta^{(m)} = \Theta^{(m-1)} + \boldsymbol{H}^{-1}_{\Theta^{(m-1)}} \boldsymbol{g}_{\Theta^{(m-1)}}, \tag{35}$$

where $\boldsymbol{g}_{\Theta}$ and $\boldsymbol{H}_{\Theta}$ represent the gradient vector and the Hessian matrix of the energy of the projection:

$$\boldsymbol{g}_{\Theta} = \frac{\partial}{\partial \Theta} \boldsymbol{e}_0^{\mathrm{T}} \boldsymbol{P}_{\Theta} \boldsymbol{e}_0 \text{ and} \tag{36}$$

$$\boldsymbol{H}_{\Theta} = \frac{\partial^2}{\partial \Theta \partial \Theta^{\mathrm{T}}} \boldsymbol{e}_0^{\mathrm{T}} \boldsymbol{P}_{\Theta} \boldsymbol{e}_0. \tag{37}$$

We now present the equations for $\boldsymbol{g}_{\Theta}$ and $\boldsymbol{H}_{\Theta}$. They can be obtained from the formulas of the first and second derivatives of a projection matrix $\boldsymbol{P}_{\Theta}$ with respect to the elements in Θ [17]. Since $\hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta}$ is of full rank,

$$\begin{aligned} \boldsymbol{P}_{\Theta} &= \hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta} \cdot (\hat{\boldsymbol{P}}^{*}_{\mathrm{avrgn}\Theta} \cdot \hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta})^{-1} \cdot \hat{\boldsymbol{P}}^{*}_{\mathrm{avrgn}\Theta} \\ &= \hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta} \cdot \hat{\boldsymbol{P}}^{+}_{\mathrm{avrgn}\Theta}, \end{aligned} \tag{38}$$

where $(\cdot)^*$ means the Hermitian transpose and $(\cdot)^+$ means the pseudoinverse. The element of $\boldsymbol{g}_{\Theta}$ is

$$\begin{aligned} \frac{\partial}{\partial \theta_\eta} \boldsymbol{e}_0^{\mathrm{T}} \boldsymbol{P}_{\Theta} \boldsymbol{e}_0 &= \boldsymbol{e}_0^{\mathrm{T}} \frac{\partial \boldsymbol{P}_{\Theta}}{\partial \theta_\eta} \boldsymbol{e}_0 \\ &= \boldsymbol{e}_0^{\mathrm{T}} \boldsymbol{P}_{\Theta\eta} \boldsymbol{e}_0 \\ &= \boldsymbol{e}_0^{\mathrm{T}} (\hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} + \hat{\boldsymbol{P}} \hat{\boldsymbol{P}}^{+}_\eta) \boldsymbol{e}_0^{\mathrm{T}}, \end{aligned} \tag{39}$$

where we write $\hat{\boldsymbol{P}}$ instead of $\hat{\boldsymbol{P}}_{\mathrm{avrgn}\Theta}$ for the ease of notation. After some algebraic manipulations, the following is obtained for the pseudoinverse:

$$\hat{\boldsymbol{P}}^{+}_\eta = (\hat{\boldsymbol{P}}^{*} \hat{\boldsymbol{P}})^{-1} \hat{\boldsymbol{P}}^{*}_\eta \boldsymbol{P}^{\perp}_{\Theta} - \hat{\boldsymbol{P}}^{+} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+}, \tag{40}$$

where $\boldsymbol{P}^{\perp}_{\Theta} = \boldsymbol{I} - \boldsymbol{P}_{\Theta}$. Combining Eqs. (39) and (40) gives

$$\frac{\partial}{\partial \theta_\eta} \boldsymbol{e}_0^{\mathrm{T}} \boldsymbol{P}_{\Theta} \boldsymbol{e}_0 = \boldsymbol{e}_0^{\mathrm{T}} \{\boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} + (\boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+})^{*}\} \boldsymbol{e}_0. \tag{41}$$

The second derivative is given by

$$\begin{aligned} &\frac{\partial^2}{\partial \theta_\eta \partial \theta_\xi} \boldsymbol{e}_0^{\mathrm{T}} \boldsymbol{P}_{\Theta} \boldsymbol{e}_0 \\ &\quad = \boldsymbol{e}_0^{\mathrm{T}} \{\boldsymbol{P}^{\perp}_{\Theta\xi} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} + \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_{\eta\xi} \hat{\boldsymbol{P}}^{+} + \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+}_\xi \\ &\qquad + (\boldsymbol{P}^{\perp}_{\Theta\xi} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} + \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_{\eta\xi} \hat{\boldsymbol{P}}^{+} + \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+}_\xi)^{*}\} \boldsymbol{e}_0. \end{aligned} \tag{42}$$

Using Eq. (40) and $\boldsymbol{P}^{\perp}_{\Theta\xi} = -\boldsymbol{P}_{\Theta\xi}$ gives

$$\begin{aligned} &\frac{\partial^2}{\partial \theta_\eta \partial \theta_\xi} \boldsymbol{e}_0^{\mathrm{T}} \boldsymbol{P}_{\Theta} \boldsymbol{e}_0 \\ &\quad = \boldsymbol{e}_0^{\mathrm{T}} \{-\boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\xi \hat{\boldsymbol{P}}^{+} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} - \hat{\boldsymbol{P}}^{+*} \hat{\boldsymbol{P}}^{*}_\xi \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} \\ &\qquad + \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_{\eta\xi} \hat{\boldsymbol{P}}^{+} + \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta (\hat{\boldsymbol{P}}^{*} \hat{\boldsymbol{P}})^{-1} \hat{\boldsymbol{P}}^{*}_\xi \boldsymbol{P}^{\perp}_{\Theta} \\ &\qquad - \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} \hat{\boldsymbol{P}}_\xi \hat{\boldsymbol{P}}^{+} \\ &\qquad + \big(- \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\xi \hat{\boldsymbol{P}}^{+} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} - \hat{\boldsymbol{P}}^{+*} \hat{\boldsymbol{P}}^{*}_\xi \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} \\ &\qquad + \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_{\eta\xi} \hat{\boldsymbol{P}}^{+} + \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta (\hat{\boldsymbol{P}}^{*} \hat{\boldsymbol{P}})^{-1} \hat{\boldsymbol{P}}^{*}_\xi \boldsymbol{P}^{\perp}_{\Theta} \\ &\qquad - \boldsymbol{P}^{\perp}_{\Theta} \hat{\boldsymbol{P}}_\eta \hat{\boldsymbol{P}}^{+} \hat{\boldsymbol{P}}_\xi \hat{\boldsymbol{P}}^{+} \big)^{*}\} \boldsymbol{e}_0. \end{aligned} \tag{43}$$

2.4. DOA Estimation Experiments

We compared the DOAs estimation performance without and with reoptimization of DOA for artificial data. The experiment conditions are shown in Fig. 1. The two speech sources and two walls were arranged as shown in Fig. 1. The speech sounds were assumed to arrive at microphones from each direct and each first-reflected sound direction. This is equivalent to a mixture with one true source and one image source for each independent signal.

We used 64 mixtures with 8 English and 8 French speeches, which were spoken by one female and one male

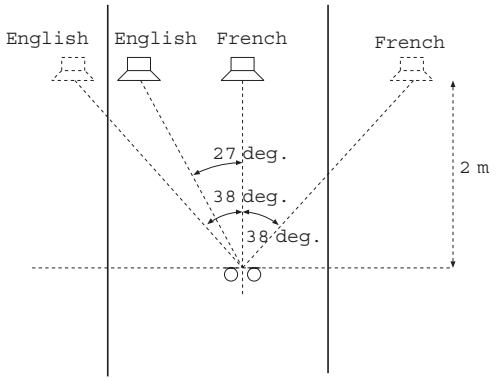


Fig. 1 Artificial mixture conditions.

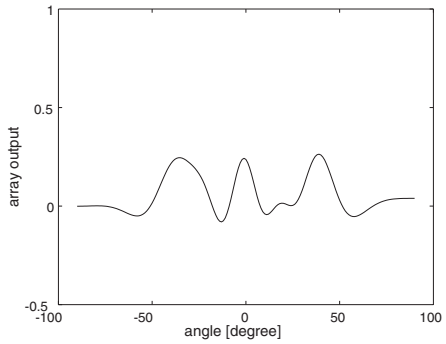


Fig. 2 Observed power.

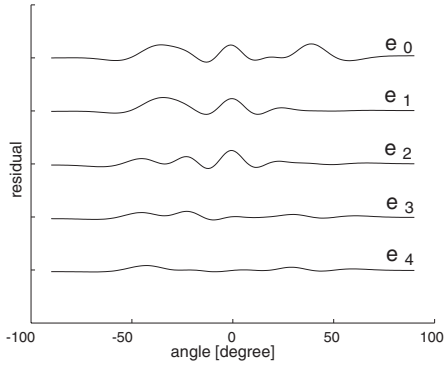


Fig. 3 Matching pursuit iterations without reoptimization.

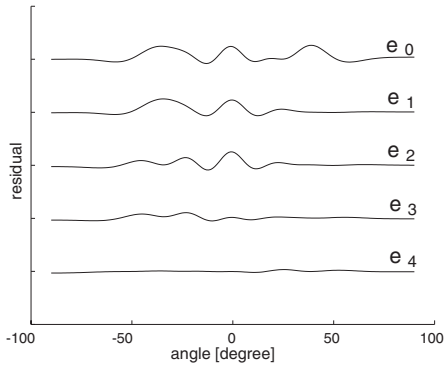


Fig. 4 Matching pursuit iterations with reoptimization.

for each language. The average length of speech is 5 s. We estimated DOAs for each sentence and averaged 64 results. The distance between microphones was 5 cm. We used a data sampling frequency of 44.1 kHz, a frame length of 46 ms, and a frame update of 23 ms.

Figure 2 shows the array output. Figure 3 has the matching pursuit iteration without reoptimization. Figure 4 has that with reoptimization. The x-axis is the direction of arrival, and the y-axis is the power of the arriving sound. The top curve is the observed power e_0. The second curve is the residual e_1 after the first DOA is estimated. The third, fourth, and last curves are the residual e_2, e_3, and e_4. Figure 5 shows the estimated DOA without reoptimization. Figure 6 shows that with reoptimization.

Table 1 lists DOA estimation results. When the reoptimization process was included, it had much better performance of DOA estimation than that without reoptimization.

3. IMPULSE RESPONSE ESTIMATION AND UNMIXING

If we find a DOA whose amplitude is a_1, time lag from the time origin is D_1, and time delay between microphones is τ_1, the impulse responses for the source and each microphone are

$$\hat{A}_{11}(t) = a_1 \cdot \delta\left(t - \left(D_1 + \frac{\tau_1}{2}\right)\right), \text{ and} \tag{44}$$

Table 1 DOA estimation.

Method	DOA [degree]
True	−38.0, −27.0, 0.0, 38.0
Without reoptimization	−34.0, −23.0, −1.0, 39.0
With reoptimization	−38.5, −27.7, −0.3, 38.2

$$\hat{A}_{21}(t) = a_1 \cdot \delta\left(t - \left(D_1 - \frac{\tau_1}{2}\right)\right). \tag{45}$$

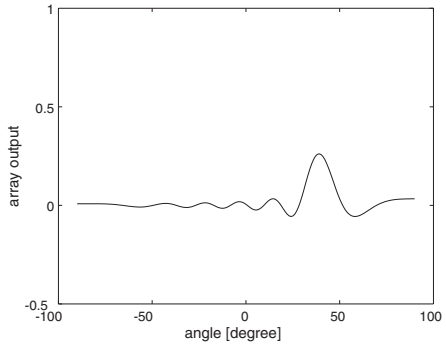
If we assume that the estimated L DOAs originate from two different sources, we can classify L different sounds into two categories based on their cross-correlation. If L_1 sources are classified into category 1 and L_2 sources are classified into category 2, the mixing matrix can be estimated in the time domain as

$$\hat{\boldsymbol{A}} = \begin{bmatrix} \hat{A}_{11}(t) & \hat{A}_{12}(t) \\ \hat{A}_{21}(t) & \hat{A}_{22}(t) \end{bmatrix}, \tag{46}$$

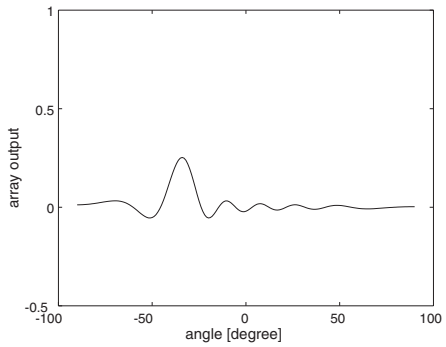
$$\hat{A}_{11}(t) = \sum_{i=1}^{L_1} a_{1i} \cdot \delta\left(t - \left(D_0 + \frac{\tau_{1i}}{2} + D_{1i}\right)\right), \tag{47}$$

$$\hat{A}_{12}(t) = \sum_{i=1}^{L_2} a_{2i} \cdot \delta\left(t - \left(D_0 + \frac{\tau_{2i}}{2} + D_{2i}\right)\right), \tag{48}$$

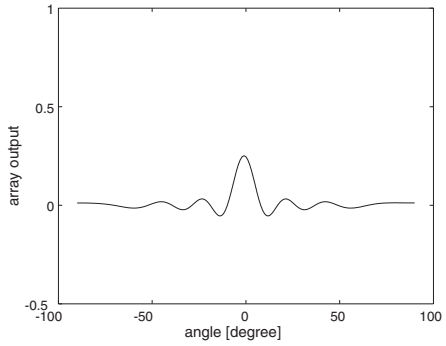
$$\hat{A}_{21}(t) = \sum_{i=1}^{L_1} a_{1i} \cdot \delta\left(t - \left(D_0 - \frac{\tau_{1i}}{2} + D_{1i}\right)\right), \text{ and} \tag{49}$$



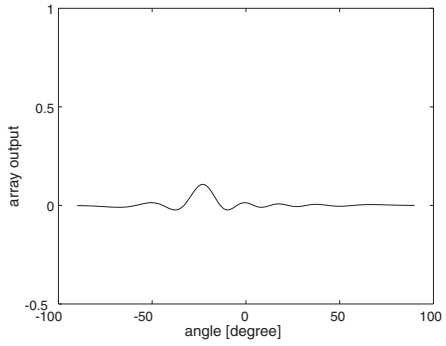
(a) First DOA.



(b) Second DOA.

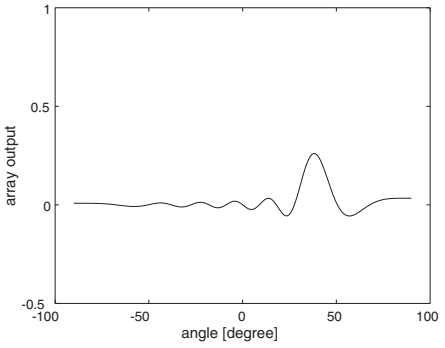


(c) Third DOA.

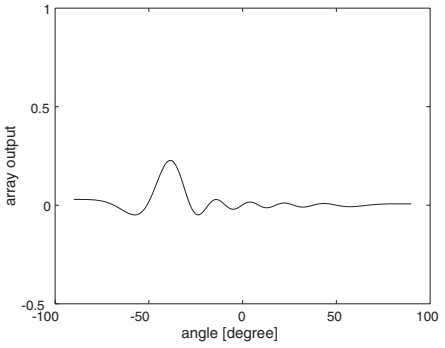


(d) Fourth DOA.

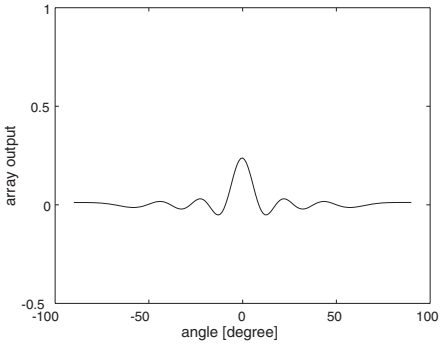
Fig. 5 Estimated DOAs without reoptimization.



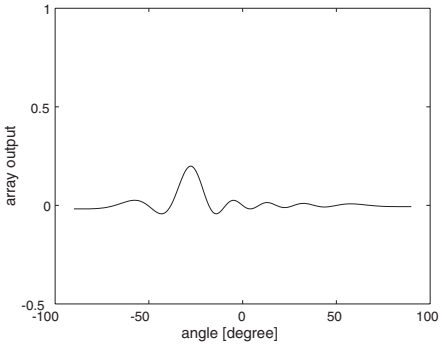
(a) First DOA.



(b) Second DOA.



(c) Third DOA.



(d) Fourth DOA.

Fig. 6 Estimated DOAs with reoptimization.

$$\hat{A}_{22}(t) = \sum_{i=1}^{L_2} a_{2i} \cdot \delta\left(t - \left(D_0 - \frac{\tau_{2i}}{2} + D_{2i}\right)\right), \tag{50}$$

where τ_{1i} and τ_{2i} are the time delays between microphones for each DOA i, D_0 is an initial delay (which is arbitrary), D_{1i} and D_{2i} are the time lags from the first sounds (which are obtained from calculating cross-correlations of sounds and the first sound), and a_{1i} and a_{2i} are the amplitudes of sounds obtained from DOA estimation results ($\hat{a}(\hat{\theta}_l)$ in Eq. (25)).

We computed unmixing matrix $\boldsymbol{B}$ as the inverse of the mixing matrix

$$\boldsymbol{B} = \hat{\boldsymbol{A}}^{-1}. \tag{51}$$

4. SOURCE SEPARATION EXPERIMENTS

We separated mixture sources for both artificial and real-world data and compared our results with those of the ICA method obtained by Kurita *et al.* [7]. We set the number of iterations to 1000 and the step-size parameter to 0.001 for the conventional ICA-based source separation. The stopping criterion was set to 30 dB for the new method. To evaluate the performance, we used the noise reduction rate (*NRR*), which is defined as the output signal-to-noise ratio (*SNR*) in dB minus the input *SNR* in dB [12]:

$$NRR = \frac{NRR_1 + NRR_2}{2}, \tag{52}$$

$$NRR_i = SNR_{Oi} - SNR_{Ii}, \tag{53}$$

$$SNR_{Oi} = 10\log\frac{\sum_\omega |C_{ii}(\omega)S_i(\omega)|^2}{\sum_\omega |C_{ij}(\omega)S_j(\omega)|^2}, \text{ and} \tag{54}$$

$$SNR_{Ii} = 10\log\frac{\sum_\omega |A_{ii}(\omega)S_i(\omega)|^2}{\sum_\omega |A_{ij}(\omega)S_j(\omega)|^2}, \tag{55}$$

where $\boldsymbol{C} = \boldsymbol{BA}$ and $i \neq j$.

The distance between microphones was 5 cm. We used a data sampling frequency of 44.1 kHz, a frame length of 46 ms, and a frame update of 23 ms.

4.1. Artificial Experiments

The artificial conditions and sources were the same as those described in Sect. 2.4. Table 2 lists the experimental results.

Table 2 *NRR* values for artificial data.

Method	*NRR* [dB]
Conventional ICA	18.7
Proposed method	25.5

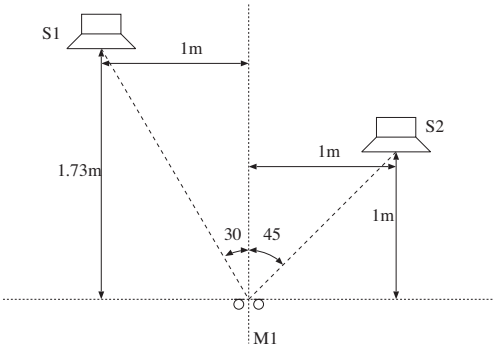


Fig. 7 Real-world mixture conditions.

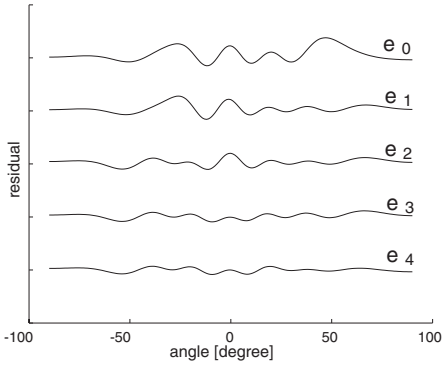


Fig. 8 Matching pursuit iterations in room 1.

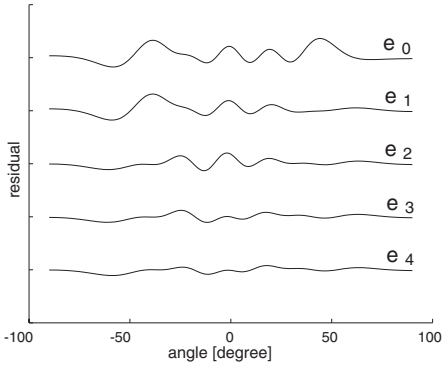


Fig. 9 Matching pursuit iterations in room 2.

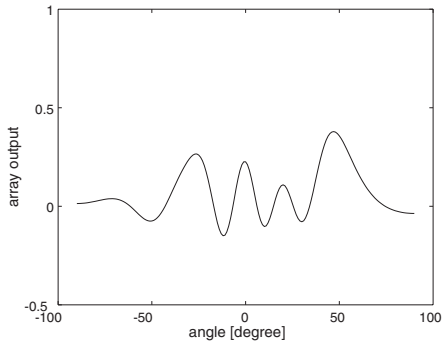


Fig. 10 Observed power in room 1.

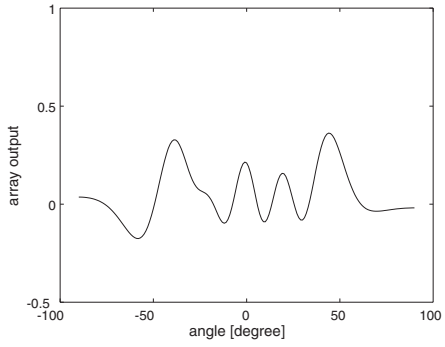
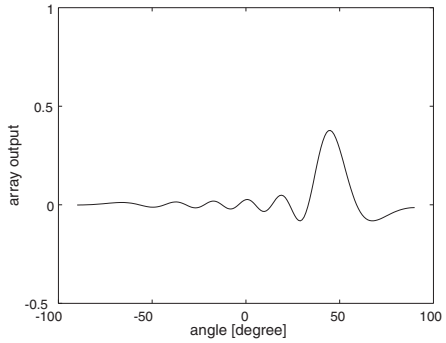
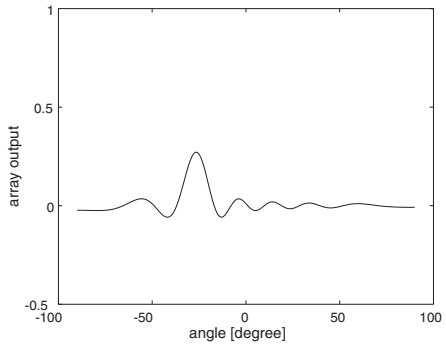


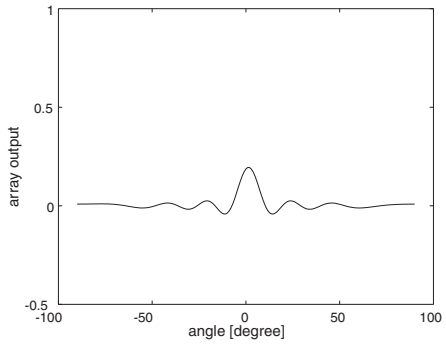
Fig. 11 Observed power in room 2.



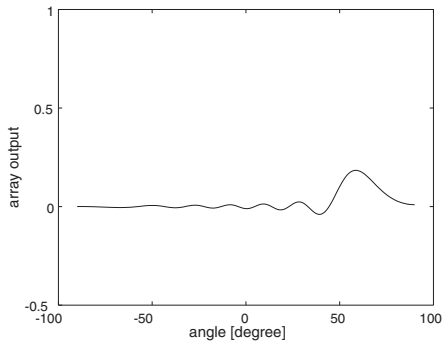
(a) First DOA.



(b) Second DOA.

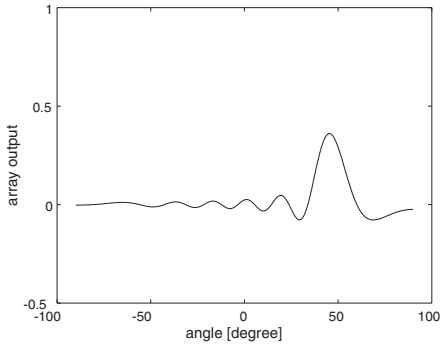


(c) Third DOA.

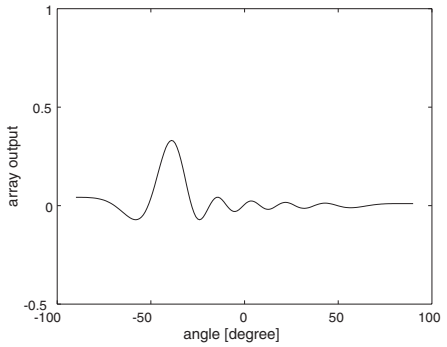


(d) Fourth DOA.

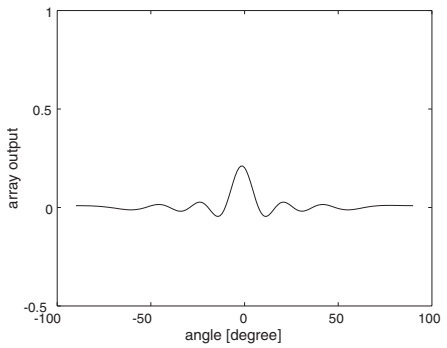
Fig. 12 DOA estimation in room 1.



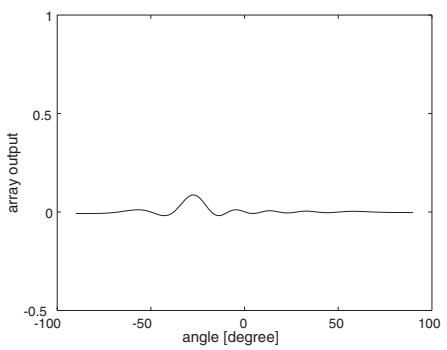
(a) First DOA.



(b) Second DOA.



(c) Third DOA.



(d) Fourth DOA.

Fig. 13 DOA estimation in room 2.

4.2. Real-World Experiments

The real-world data were recorded in two rooms. A reverberation time was 0.36 s in room 1 and 0.94 s in room 2. Figure 7 shows the real-world experiment conditions. Sources were located at -30 degrees and 45 degrees. The sources are the same as those described in Sect. 2.4.

Figures 8 and 9 show the matching pursuit iterations in room 1 and room 2, respectively. Figures 10 and 11 show

Table 3 *NRR* values for real-world data.

Method	*NRR* [dB] (room 1)	*NRR* [dB] (room 2)
Conventional ICA	4.51	3.98
Proposed method	6.62	5.15

the array output in each room. Figures 12 and 13 show the estimated DOA in each room. Table 3 lists the experimental results. We found ten sounds in the real-world data that reached the stopping criterion of 30 dB. The residual e_{10} was almost flat.

Separation by the proposed method was superior to that by conventional ICA-based source separation for both artificial and real-world data. The improvement of separation in the real-world data experiment is less than that in the artificial experiment, since the estimation of reflected sounds was insufficient.

5. CONCLUSION

In this paper, we proposed a new source separation method in which spatial information derived from the results of DOA estimates for each direct and reflected sound obtained by beam forming is used. Its main advantage is that we can estimate the mixing system for direct and early reflected sounds and separate sounds by a suitable technique with sound source separation in the real world. A matching pursuit algorithm that includes a reoptimization step for each iteration was used for DOA estimation; it could estimate these correctly.

Source separation by our proposed method was better than that by ICA in both artificial and real-world data experiments. It improved the noise reduction rate by about 7 dB in an artificial data experiment and by about 2.0 dB and 1.0 dB in a real-world data experiment which has reverberation time of about 0.4 s and 1.0 s, respectively.

We expect that additional improvement can be achieved in real-world cases if the accuracy of estimation of reflected sounds is increased. This could for instance be achieved by narrowing the beam in DOA estimation. We are currently working on this.

ACKNOWLEDGMENT

The authors express their gratitude to Professor Bastiaan Kleijn and Professor Arne Leijon, KTH, for many fruitful discussions and suggestions.

REFERENCES

[1] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Comput.*, **7**, 1129–1159 (1995).
[2] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, **86**, 2009–2025 (1998).
[3] S. Amari and A. Cichocki, "Adaptive blind signal processing—Neural network applications," *Proc. IEEE*, **86**, 2026–2048 (1998).
[4] T.-W. Lee, *Independent Component Analysis. Theory and Applications* (Kluwer Academic Publishers, Boston, 1998).
[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, **22**, 21–34 (1998).
[6] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," *Proc. Int. Workshop Independent Comp. Analysis Blind Signal Separation* (*ICA'99*), pp. 365–371 (1999).
[7] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3140–3143 (2000).
[8] H. Saruwatari, T. Kawamura and K. Shikano, "Blind source separation for speech based on fast-convergence algorithm with ICA and beamforming," *Proc. Eurospeech*, pp. 2603–2606 (2001).
[9] L. C. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, **10**, 352–362 (2002).
[10] A. Jourjine, S. Rickard and Ö. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 2985–2988 (2000).
[11] R. Gribonval, "Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3057–3060 (2002).
[12] S. Araki, S. Makino, T. Nishikawa and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 2737–2740 (2001).
[13] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Acoust. Speech Signal Process. Mag.*, **5**, 4–24 (1988).
[14] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques* (Prentice-Hall, Englewood Cliffs, 1993).
[15] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, **41**, 3397–3415 (1993).
[16] K. Vos, R. Vafin, R. Heusdens and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," *Proc. 1999 Audio Eng. Soc. 17th Conf. "High Quality Audio Coding*," pp. 244–250 (1999).
[17] M. Viberg and B. Ottersten, "Sensor array processing based on subspace fitting," *IEEE Trans. Signal Process.*, **39**, 1110–1121 (1991).