Noise Adaptive Optimization of Matrix Initialization for Frequency-Domain Independent Component Analysis

Makoto Yamada
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Kyoto, 619-0237, Japan

Gordon Wichern
School of Arts, Media + Engineering
Arizona State University, Tempe, AZ 85281, USA

Kazunobu Kondo
Yamaha Corporation
203 Matsunokijima, Iwata, Shizuoka, 438-0192, Japan

Masashi Sugiyama
Department of Computer Science, Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

Hiroshi Sawada
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Kyoto, 619-0237, Japan

Abstract

Initializing an unmixing matrix is an important problem in source separation since an objective function to be optimized is typically non-convex. In this paper, we consider the problem of two-source signal separation from a two-microphone array located on a mobile device, where a point source such as a speech signal is placed in front of the array, while no information is available about another interference signal. We propose a simple and computationally efficient method for estimating the geometry and source type (a point or diffuse) of the interference signal, which allows us to adaptively choose a suitable unmixing matrix initialization scheme. Our proposed method, *noise adaptive optimization of matrix initialization (NAOMI)*, is shown to be effective through source separation simulations.

Keywords

Mobile source separation device, covariance fitting, source type classification, frequency-domain independent component analysis

1 Introduction

Real-time implementation of frequency-domain independent component analysis (FDICA) [1, 2, 3] into mobile devices has recently attracted a great deal of attention from the audio industry [4, 5, 6], because of its various potential applications such as speech enhancement and speaker separation. However, since surrounding environments, source positions, and source mixing rates are constantly changing in mobile device applications, it is highly challenging to achieve a good performance in real-world environments.

Many effective approaches have been proposed for improving the performance of FDICA by exploiting knowledge of room and sensor geometry [7], geometric information of sound sources [2, 8, 9, 10, 11], a sophisticated prior model of speech signals [12], higher-order frequency dependencies [13, 14], and post-processing with non-stationary spectral subtraction [15]. However, these approaches implicitly assume knowledge of the sound source geometry and the source type (a point source, a diffuse source, etc.), and thus, they are valid only in limited cases. In addition, since the cost function of FDICA is typically non-convex, FDICA is not guaranteed to converge to the global optimal solution, when the initial unmixing matrix is incorrectly chosen [3, 16, 17]. Thus, unmixing matrix initialization is a key factor for successfully implementing FDICA in mobile devices.

A popular unmixing matrix initialization technique is the combination of *delay-and-sum (DS)* and *null beamformers* (NBF) [3, 16, 17], which is known to be robust to the FDICA permutation problem [3]. However, beamformer-based initialization heavily depends on the sound source geometry and the source type. Thus, beamformer-based initialization itself is not suited for mobile use, without a reasonable estimator of the source geometry and the source type.

In this paper, we propose an algorithm called *noise adaptive optimization of matrix initialization* (NAOMI) for estimating the source geometry and the source type. We consider the problem of two-source separation from a two-microphone array, where a point source such as a speech signal is placed in front of the array, while another *interfering* source should be separated and removed using FDICA. The interfering source is either another point source that is not located directly in front of the microphones (e.g., a speech signal that is not intended to be captured by the microphones) or a diffuse source (e.g., loud background music or airplane engine rumble). To identify the type of the interfering source, we first estimate its direction of arrival (DOA) at each frequency bin using *covariance fitting* [18], and then use the variance of the estimated DOAs to classify the interfering source. The initial unmixing matrix is then selected based on the estimated source type. The effectiveness of the proposed method for speech de-noising is evaluated through source separation simulations.

2 Problem Formulation

In this section, we formulate FDICA and review beamformer-based unmixing matrix initialization methods.

2.1 Frequency-Domain Independent Component Analysis (FDICA)

The K observed signals by N microphones in real environments can be modeled as convolutive mixtures:

$$x_j(t) := \sum_{i=1}^{K} \sum_{k=1}^{P} a_{ji}(k) s_i(t-k+1),\ j = 1, \ldots, N, \tag{1}$$

where s_i is the signal from source i, x_j is the observed signal at microphone j, and a_{ji} is the P-tap impulse response from source i to microphone j. Converting the time-domain convolutive mixtures into the frequency domain by the Short-Time Fourier Transform (STFT), we can express the convolutive mixture at frequency f of the τ-th windowed frame as

$$\boldsymbol{x}(f,\tau) := \boldsymbol{A}(f)\boldsymbol{s}(f,\tau), \tag{2}$$

where $\boldsymbol{x}(f,\tau) = [x_1(f,\tau), \ldots x_N(f,\tau)]^\top \in \mathbb{C}^N$ is the observed signal vector at the microphones, $\boldsymbol{A}(f) = [\boldsymbol{a}_1(f), \ldots, \boldsymbol{a}_K(f)] \in \mathbb{C}^{N \times K}$ is the mixing matrix, $\mathbf{s}(f,\tau) = [s_1(f,\tau), \ldots, s_K(f,\tau)]^\top \in \mathbb{C}^K$ is the source signal, and $^\top$ denotes a transpose of matrix. In this paper, we assume that the numbers of observed signals and microphones are $N = K = 2$.

Let us define the separated signals $\boldsymbol{y}(f,\tau) = [y_1(f,\tau), \cdots, y_N(f,\tau)]^\top \in \mathbb{C}^N$ as

$$\boldsymbol{y}(f,\tau) := \boldsymbol{W}(f)\boldsymbol{x}(f,\tau), \tag{3}$$

where $\boldsymbol{W}(f) \in \mathbb{C}^{N \times N}$ is called the unmixing matrix. Then, the goal of FDICA is to estimate the unmixing matrix $\boldsymbol{W}(f)$ such that $\boldsymbol{W}(f)\boldsymbol{A}(f) = \boldsymbol{I}$ up to scaling and permutation, where $\boldsymbol{I}$ is the identity matrix. In this paper, we employ an information-theoretic approach [1, 19]:

$$\boldsymbol{W}^*(f) := \underset{\boldsymbol{W}(f)}{\operatorname{argmin}} \int p(\boldsymbol{y}(f)) \log \frac{p(\boldsymbol{y}(f))}{\prod_{k=1}^{K} p_k(y_k(f))} \mathrm{d}\boldsymbol{y}(f), \tag{4}$$

where $p(\boldsymbol{y}(f))$ is the probability density function (pdf) of $\boldsymbol{y}(f)$, $\boldsymbol{y}(f) = [y_1(f), \ldots, y_K(f)]^\top$, $p_k(y_k(f))$ is the marginal pdf of $y_k(f)$. Note that $\boldsymbol{y}(f,\tau)$ is the sample-approximated signal of $\boldsymbol{y}(f)$. The unmixing matrix $\boldsymbol{W}(f)$ can be obtained by using the following iterative formula [1]:

$$\boldsymbol{W}^{(l+1)}(f) = \boldsymbol{W}^{(l)}(f) + \eta\big\{\mathrm{I} - E_\tau[\boldsymbol{\phi}(f,\tau)\boldsymbol{y}^H(f,\tau)]\boldsymbol{W}^{(l)}(f)\big\}, \tag{5}$$

where ^{H} is the Hermitian transpose of a matrix, $\boldsymbol{W}^{(l)}(f)$ is l-th iteration of $\boldsymbol{W}(f)$, η is a step-size parameter, and $E_\tau[\boldsymbol{y}(f,\tau)]$ is the expectation of $\boldsymbol{y}(f,\tau)$ with respect to τ. The nonlinear function $\boldsymbol{\phi}(\cdot)$ is defined by

$$\boldsymbol{\phi}(f,\tau) := [\phi_1(f,\tau) \cdots \phi_K(f,\tau)]^\top, \tag{6}$$

$$\phi_k(f,\tau) := \mathrm{sgn}(\mathrm{Re}\{y_k(f,\tau)\}) + j\mathrm{sgn}(\mathrm{Im}\{y_k(f,\tau)\}), \tag{7}$$

where sgn$(\cdot)$ is the sign function and Re$\{\cdot\}$ and Im$\{\cdot\}$ denote the real and imaginary parts of a complex number, respectively.

Since the cost function (4) is scale- and permutation-invariant with respect to $\boldsymbol{W}(f)$, removing scaling and permutation ambiguity is an important issue in FDICA. In this paper, we employ the projection-back method for removing scaling ambiguity [2]:

$$\widehat{\boldsymbol{W}}(f) \longleftarrow \mathrm{diag}(\widehat{\boldsymbol{W}}^{-1}(f))\widehat{\boldsymbol{W}}(f), \tag{8}$$

where diag$(\boldsymbol{W})$ denotes the diagonal matrix with diagonal elements $W_{1,1}, W_{2,2}, \ldots, W_{K,K}$. To alleviate the permutation ambiguity, we may use post-processing methods based on correlation [2] or the direction of arrival (DOA) [10].

Finally, we can obtain a separated time-domain signal of k-th channel by applying the *inverse Fourier transform* to separated frequency-domain signal $\widehat{\boldsymbol{y}}_k = [\widehat{y}_k(1), \widehat{y}_k(2), \ldots, \widehat{y}_k(N_f)]^\top$, where N_f is the number of frequency bins.

Since the cost function of FDICA (4) is *non-convex* with respect to $\boldsymbol{W}(f)$ [19], $W^{(\infty)}(f)$ may not converge to the global optimal solution when the initial unmixing matrix $\boldsymbol{W}^{(0)}(f)$ is set inappropriately. Thus, it would be more promising to initialize the unmixing matrix adaptively depending on the source positions or source types.

2.2 Beamformer-Based Unmixing Matrix Initialization

We assume that one of the two signal sources is a *point source* located in front of the microphone array. In such a case, the possible combinations of sound source types are *point source* + *point source* or *point source* + *diffuse source*, where we define a *point source* as a speech signal located near the microphone array, while a *diffuse source* is defined as a widely spread source located far from the microphone array. Below, we review two popular beamformer-based unmixing matrix initialization techniques for the *point source* + *point source* and *point source* + *diffuse source* cases, respectively.

2.2.1 Initialization with Null Beamformer (NBF)

NBF-based initialization is often used for separating mixtures of two point sources, i.e., speech + speech, and is given as follows [3]:

$$\boldsymbol{W}^{(0)}(f) := \begin{bmatrix} 1 & -e^{i2\pi f d \sin(\theta_2)/V_c} \\ 1 & -e^{i2\pi f d \sin(\theta_1)/V_c} \end{bmatrix}, \tag{9}$$

where θ_1 and θ_2 are the DOAs of the point sources, d is the distance between the two microphones, and V_c is the speed of sound. The first row of Eq.(9) cancels the signal from direction θ_2 and enhances the signal from direction θ_1, while the second row cancels the signal from direction θ_1 and enhances the signal from direction θ_2. By using the same θ_1 and θ_2 values to initialize the unmixing matrix for every frequency, NBF-based initialization is considered robust to the permutation problem. However, if the DOAs of point sources θ_1 and θ_2 are incorrectly set, $\boldsymbol{W}^{(\infty)}(f)$ might fail to converge to an acceptable solution. In addition, if the observed signal is *point source + diffuse source*, NBF-based initialization is not a reasonable choice, since the DOA of a diffuse source is usually not well-defined.

2.2.2 Initialization with Delay-and-Sum and Null Beamformer (DS-NBF)

When the source type is *point source + diffuse source*, we can not estimate the DOA of the diffuse source. In such a case, the following initial unmixing matrix is often used [16, 17]:

$$\boldsymbol{W}^{(0)}(f) := \begin{bmatrix} \frac{1}{2} & \frac{1}{2}e^{i2\pi f d \sin(\theta_1)/V_c} \\ 1 & -e^{i2\pi f d \sin(\theta_1)/V_c} \end{bmatrix}, \tag{10}$$

where the first row of Eq.(10) enhances the signal from direction θ_1, while the second row cancels the signal from direction θ_1, i.e., enhances the signal from all other directions. DS-NBF initialization does not require the DOA of the diffuse source. Thus, DS-NBF initialization is well suited for the *point source + diffuse source* separation problem. However, it may not work well for the separation of two point sources, necessitating an automatic method for classifying the type of interfering source.

3 Noise Adaptive Optimization of Matrix Initialization (NAOMI)

In this section, we propose the *noise adaptive optimization of matrix initialization* (NAOMI) algorithm for the estimation of the source geometry and source type, where we assume that the geometry of a target source is known. NAOMI consists of two components: Estimating the geometry of the interfering sound source (i.e., θ_2) and classifying the type of the interfering source based on the estimated geometry. Then we choose the NBF or DS-NBF initialization scheme adaptively depending on the type of interfering source.

3.1 Estimating the Geometry of the Interfering Sound Source

Let $\boldsymbol{d}_{\mathrm{f}}(\theta) \in \mathbb{C}^N$ denotes the response of the microphone array to a plane wave of unit amplitude arriving from direction θ at frequency f; we will refer to $\boldsymbol{d}_{\mathrm{f}}(\cdot)$ as an *array manifold* or *steering vector*. If we assume that narrow-band sources $s_1(f,\tau)$ and $s_2(f,\tau)$ are impinging on the microphone array at angles θ_1 and θ_2, respectively, then the vector array input $\boldsymbol{x}(f,\tau) \in \mathbb{C}^N$ can be represented as

$$\boldsymbol{x}(f,\tau) = s_1(f,\tau)\boldsymbol{d}_{\mathrm{f}}(\theta_1) + s_2(f,\tau)\boldsymbol{d}_{\mathrm{f}}(\theta_2), \tag{11}$$

where we assumed that $\boldsymbol{a}_1(f,\tau) = \boldsymbol{d}_{\mathrm{f}}(\theta_1)$ and $\boldsymbol{a}_2(f,\tau) = \boldsymbol{d}_{\mathrm{f}}(\theta_2)$. Here, we assume that the direction of the target source θ_1 is given in advance (i.e., $\boldsymbol{d}_{\mathrm{f}}(\theta_1)$ is computed using microphone array structure information). In particular, when a target source is located in front of a microphone array, we have $\boldsymbol{d}_{\mathrm{f}}(0) = [\frac{1}{2}\ \frac{1}{2}]^{\top}$.

If $s_1(f,\tau)$ and $s_2(f,\tau)$ are independent and zero-mean signals, then we can express the covariance matrix of $\boldsymbol{x}(f,\tau)$ as

$$\boldsymbol{R}_{\mathrm{xx}}(f) = \boldsymbol{R}_{\mathrm{s_1s_1}}(f) + \boldsymbol{R}_{\mathrm{s_2s_2}}(f), \tag{12}$$

where $\sigma_1^2(f) = E_\tau[s_1^2(f,\tau)]$, $\sigma_2^2(f) = E_\tau[s_2^2(f,\tau)]$, $E_\tau[s_1^2(f,\tau)]$ is the expectation of $s_1^2(f,\tau)$ with respect to τ, $\boldsymbol{R}_{\mathrm{s_1s_1}}(f) = \sigma_1^2(f)\boldsymbol{d}_{\mathrm{f}}(\theta_1)\boldsymbol{d}_{\mathrm{f}}(\theta_1)^H$ is the covariance matrix of a *known* target signal, and $\boldsymbol{R}_{\mathrm{s_2s_2}}(f) = \sigma_2^2(f)\boldsymbol{d}_{\mathrm{f}}(\theta_2)\boldsymbol{d}_{\mathrm{f}}(\theta_2)^H$ is the covariance matrix of an *unknown* interference signal.

Since the interference covariance matrix $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ is not available in practice, we need to estimate $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ from the observed signals. From Eq.(12), $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ can be written as

$$\boldsymbol{R}_{\mathrm{s_2s_2}}(f) = \boldsymbol{R}_{\mathrm{xx}}(f) - \sigma_1^2(f)\boldsymbol{d}_{\mathrm{f}}(\theta_1)\boldsymbol{d}_{\mathrm{f}}^H(\theta_1), \tag{13}$$

where $\boldsymbol{d}_{\mathrm{f}}(\theta_1)$ is assumed to be known and $\sigma_1^2(f)$ is the unknown parameter. Since $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ is independent of the target signal x_1, we want to remove the $\boldsymbol{d}_{\mathrm{f}}(\theta_1)$-component from $\boldsymbol{R}_{\mathrm{xx}}(f)$ as much as possible. That is, estimation of $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ can be achieved by maximizing $\sigma_1^2(f)$. Moreover, since the covariance matrix is positive semidefinite, $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ should fulfill the positive semi-definiteness. Then, the estimation problem of $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ can be formulated as

$$\max \sigma_1^2(f) \quad \text{s.t. } \boldsymbol{R}_{\mathrm{xx}}(f) - \sigma_1^2(f)\boldsymbol{d}_{\mathrm{f}}(\theta_1)\boldsymbol{d}_{\mathrm{f}}^H(\theta_1) \succeq 0, \tag{14}$$

where $\succeq 0$ means the positive semi-definiteness of a matrix. This formulation is known as *covariance fitting*, and is often used for robust beamformer estimation [20, 18]. The optimal $\sigma^2(f)$ is given as (see Eqs.(8) and (9) in [20] for the detail derivation)

$$\widehat{\sigma}_1^2(f) = \frac{1}{\boldsymbol{d}_{\mathrm{f}}^H(\theta_1)\boldsymbol{R}_{\mathrm{xx}}^{-1}(f)\boldsymbol{d}_{\mathrm{f}}(\theta_1)}, \tag{15}$$

while an estimate of $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ is given by

$$\widehat{\boldsymbol{R}}_{\mathrm{s_2s_2}}(f) = \boldsymbol{R}_{\mathrm{xx}}(f) - \widehat{\sigma}_1^2(f)\boldsymbol{d}_{\mathrm{f}}(\theta_1)\boldsymbol{d}_{\mathrm{f}}^H(\theta_1). \tag{16}$$

Since $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ is assumed to be a rank one matrix, $\boldsymbol{d}_{\mathrm{f}}(\theta_2)$ is estimated from the normalized eigenvector of $\boldsymbol{R}_{\mathrm{s_2s_2}}(f)$ associated with a leading eigenvalue, which we denote by $\widehat{\boldsymbol{d}}_f$. Thus, we estimate the DOA of the *interfering* point source as

$$\widehat{\theta}_2(f) = \arg\max_{\theta} |\widehat{\boldsymbol{d}}^H(f)\boldsymbol{d}_{\mathrm{f}}(\theta)|, \quad f = 1, \dots, N_f, \tag{17}$$

where N_f is the number of frequency bins.

The goal here is to estimate the DOA of an interference signal from N_f DOAs, and the simplest way would be to average N_f DOAs (i.e., $\widehat{\theta}_2 = \frac{1}{N_f}\sum_{f=1}^{N_f}\widehat{\theta}_2(f)$). However, it is in practice difficult to estimate $\theta_2(f)$ accurately at low frequencies due to the small time-difference between the observed signals at the two microphones, and spatial aliasing occurs if $f > \frac{V_c}{2d}$. Thus, simple averaging tends to perform poorly. For this reason, we instead use DOAs from only a certain range of frequencies to estimate θ_2 as

$$\widehat{\theta}_2 = \frac{1}{f_e - f_s} \sum_{f=f_s}^{f_e} \widehat{\theta}_2(f), \tag{18}$$

where f_s is the low-frequency cutoff and the high-frequency cutoff is $f_e \leq \frac{V_c}{2d}$.

Note that the computational cost of the above method is almost equivalent to that of a single update iteration of Eq.(5) for all frequency bins. Thus, the proposed method is computationally very efficient.

3.2 Source Type Classification

In the case of *point source + point source*, the estimated DOA of the interfering source at each frequency bin is close to the true DOA. On the other hand, since a diffuse source consists of the reverberation or mixture of many sound sources, the estimated DOAs of a diffuse source tend to be spread over various frequency bins. Here, we propose using the variance of estimated DOAs to decide whether the source mixture type is *point source + point source* or *point source + diffuse source.* The variance of estimated DOAs is given by

$$\widehat{\sigma}^2 = \frac{1}{f_e - f_s} \sum_{f=f_s}^{f_e} (\theta_2(f) - \widehat{\theta}_2)^2. \tag{19}$$

Finally, we select the initial unmixing matrix using NBF if $\widehat{\sigma}^2 < \rho$ (point source + point source) or DS-NBF if $\widehat{\sigma}^2 \geq \rho$ (point source + diffuse source). Furthermore, if the estimated DOA $\widehat{\theta}_2$ is close to θ_1, the separation performance is degraded. To mitigate this problem, we heuristically choose the DS-NBF beamformer when $|\widehat{\theta_2 - \theta_1}| < \epsilon$, where ϵ is a threshold parameter.

A pseudo code of the proposed algorithm is described in Algorithm 1. Note that line 9 is for the decorrelation and normalization of the unmixing matrix [21]. Usually, this step should be executed at every iteration. However, it is computationally rather demanding for mobile devices, so we decided to execute this only once. As shown later, the proposed method still performs well in experiments.

4 Experiments

In this section, we assess the effectiveness of the proposed method NAOMI through experiments.

Algorithm 1 Noise Adaptive Optimization of Matrix Initialization (NAOMI)

```
$\theta_1 = 0$;
compute $\widehat{\theta}_2$;
for $f = 0$; $f \leq \frac{FFTSize}{2} + 1$; $f$++ do
    if $\widehat{\sigma}^2 < \rho$ or $|\widehat{\theta}_2 - \theta_1| \geq \epsilon$ then
        $\boldsymbol{W}^{(0)}(f) = \begin{bmatrix} 1 & -e^{i2\pi fd\sin(\widehat{\theta}_2)/V_c} \\ 1 & -e^{i2\pi fd\sin(\theta_1)/V_c} \end{bmatrix}$;
    else
        $\boldsymbol{W}^{(0)}(f) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}e^{i2\pi fd\sin(\theta_1)/V_c} \\ 1 & -e^{i2\pi fd\sin(\theta_1)/V_c} \end{bmatrix}$;
    end if
    $\boldsymbol{W}^{(0)}(f) = (\boldsymbol{W}^{(0)}(f)\boldsymbol{R}_{\mathrm{xx}}(f)\boldsymbol{W}^{(0)}(f)^H)^{-\frac{1}{2}}\boldsymbol{W}^{(0)}(f)$;
    for $l = 0$; $l < L$; $l$++ do
        $\boldsymbol{W}^{(l+1)}(f) = \boldsymbol{W}^{(l)}(f) + \eta\{\boldsymbol{I} - E_\tau[\boldsymbol{\phi}(f,\tau)\boldsymbol{y}^H(f,\tau)]\boldsymbol{W}^{(l)}(f)\}$;
    end for
end for
```

4.1 Setup

We use two microphones separated by 5.4cm, signals sampled at 8kHz, and 1024-sample frame size with 50% overlap. The detailed recording settings are described in Table 1.

For the proposed method, the parameters ρ, η, and ϵ are experimentally set to 0.7, 0.01, and 20, respectively, and the number of FDICA iterations (Eq.(5)) is fixed to 100 in Sections 4.2–4.4. For conventional methods, we fix the number of FDICA iterations to 100.

To evaluate the robustness to permutation of the separated sources, we do not explicitly solve the permutation problem via post-processing methods. We use the average noise reduction rate (NRR) as a performance measure [3].

4.2 *Point Source* + *Point Source* Separation in Anechoic Chamber

In this experiment, we use speech signals from 2 male and 4 female speakers, recorded in an anechoic chamber. By choosing one speaker as the point source located in front of the array and another speaker as the interfering source, 30 speaker combinations are used in the experiments. The interfering point source is placed at either -90, -45, 45, or 90 degrees, respectively, while the target point source is placed at 0 degree. Figure 1 shows the placement of sound sources and microphones.

We compare the proposed method NAOMI to NBF-based initialization with nulls at 0 and 90 degrees, that with nulls at 0 and -90 degrees, and DS-NBF combination consisting of DS at 0 degree and NBF at 0 degree [3, 16, 17]. We use the first three seconds of each observed signal to estimate an unmixing matrix and then evaluate its estimation accuracy using the rest of signals.

Table 1: Recording settings.

Sampling Rate	8 [kHz]
FFT Size	1024 [sample]
FFT Shift	512 [sample]
Microphone Type	OMNI, SHURE SM93
Number of Microphones	2
Interval of Microphones	5.4 [cm]

Figure 2 shows the average NRR for 30 combinations of speakers at each unknown point-source angle as a function of the FDICA iteration number in Eq.(5). As can be seen, the proposed method gives a high NRR for every interfering point-source position, while the other initialization methods work only if the DOA used to initialize the unmixing matrix matches that of the interfering point source. The results also show that the NRR of the proposed method is close to the ideal unmixing matrix initialization.

4.3 *Point Source* + *Point Source* Separation in Reverberant Room

Here, we use speech signals from 2 male and 4 female speakers, which are recorded in a reverberant room with 400ms reverberation time. The interfering point source is placed at either -90, -45, 45, or 90 degrees respectively, while the target point source is placed at 0 degree. Figure 3 shows the placement of sound sources and microphones in the reverberant room. We compare the proposed method NAOMI to NBF-based initialization with nulls at 0 and 90 degrees, that with nulls at 0 and -90 degrees, and DS-NBF combination consisting of DS at 0 degree and NBF at 0 degree [3, 16, 17].

Figure 4 shows the averaged NRR for 30 combinations of speakers at each unknown point source angle. It clearly shows that the proposed method provides a high NRR for every point-source positions compared to existing methods, even in the presence of heavy reverberation. Note, however, that the performance of the proposed method slightly inferior to FDICA with ideal unmixing matrix initialization. A possible reason for this is that DOA estimation tends to be inaccurate under heavy reverberation and thus an estimated initial unmixing matrix can be poor. However, since we observed that FDICA can still learn a reasonably good unmixing matrix from a slightly degraded initial matrix, the performance degradation of FDICA caused by poor DOA estimation is limited.

4.4 *Point Source* + *Diffuse Source* Separation in Reverberant Room

In this experiment, we use a speech signal from 2 male and 4 female speakers as a target point source located in front of the microphone array and the ambient sound of a *shinkansen* (bullet train) as an interfering diffuse source. Thus, six total sound mixtures are used in this experiment. Figure 5 shows the placement of sound source and microphones.

Figure 6 shows the average NRR for six speaker combinations as a function of the FDICA iteration number given in Eq.(5). The proposed method gives a high NRR since it choose DS-NBF as an initial projection matrix, while NBF-based initialization performs poorly. This result clearly shows that selecting an initial unmixing matrix with respect to the correct sound source is useful in FDICA. In addition, since we observed that FDICA does not outperform the initial DS-NBF in this experiment, we can use this prior knowledge to improve the source separation performance, i.e., it is possible to set the number of FDICA iterations in *point source + diffuse source* small in the proposed approach.

4.5 Environmental Adaptation in Reverberant Room

Finally, we evaluate the proposed system in a reverberant room with changing environmental conditions. The total duration of signals used in this experiment is 30s, where the source signal consists of three parts: speech (0 deg) + speech (-45 deg) (0s–10s), speech (0 deg) + speech (45 deg) (10s–20s), and speech (0 deg) + ambient noise (20s–30s).

We use a two-second block of signals for estimating the unmixing matrix, and then evaluate its estimation accuracy using the next non-overlapping two-second block. In addition, we initialize an unmixing initial matrix in each block so that the proposed method can deal with sound sources and types that change quickly. For the proposed method, the parameters ρ, η, and ϵ are experimentally set to 0.7, 0.01, and 20, respectively, and the number of FDICA iterations (Eq.(5)) is fixed to 100 for the two-point source separation case and 10 for the point source + diffuse source case.

We compare the proposed method to NBF with nulls at 0 and 45 degrees, NBF with nulls at 0 and -45 degrees, and DS-NBF. NRR as a function of time is compared among the four methods in Figure 7. As can be seen, the proposed method gives a high NRR even if the source mixture types are changed (with a time lag equal to the block size), while the other initialization methods work only if DOA used to initialize the unmixing matrix matches that of the interfering point source. Note that NRR of the proposed method in the *point source + diffuse source* section (i.e., time (t) = 20–30) is higher than that of DS-NBF. This is because we know a source type in the proposed method by Eq.(19) and we can set the number of iterations in the *point source + diffuse source* case small (as discussed in Section 4.4). This is a clear advantage of the proposed method over existing methods.

5 Conclusions

In this paper, we proposed a simple algorithm for initialization of the FDICA unmixing matrix called noise adaptive optimization of matrix initialization (NAOMI). The experimental results showed the effectiveness of the proposed method in a realistic environment

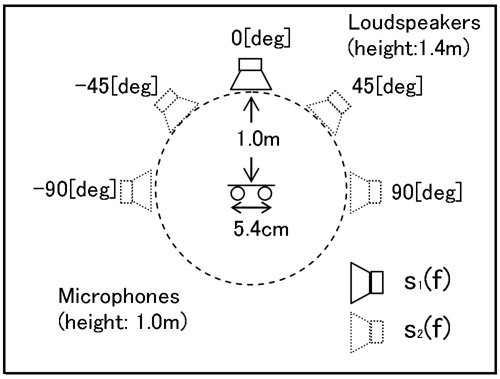
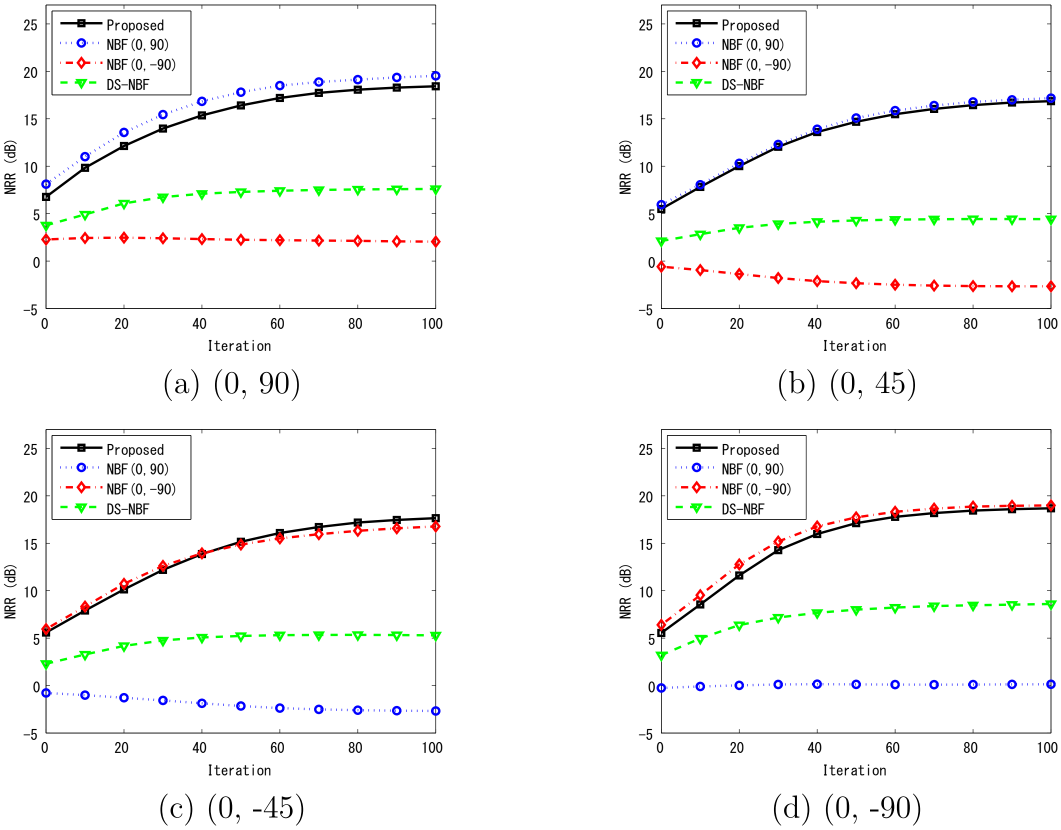


Figure 1: Recording environment in anechoic chamber. Interference source is located either −90, −45, 45, or 90 degrees, while target point source is placed at 0 degree.



(a) (0, 90)

(b) (0, 45)

(c) (0, -45)

(d) (0, -90)

Figure 2: Noise reduction rate (NRR) as a function of FDICA iteration in anechoic chamber. DOA of true sources are shown in the bracket (θ_1, θ_2).

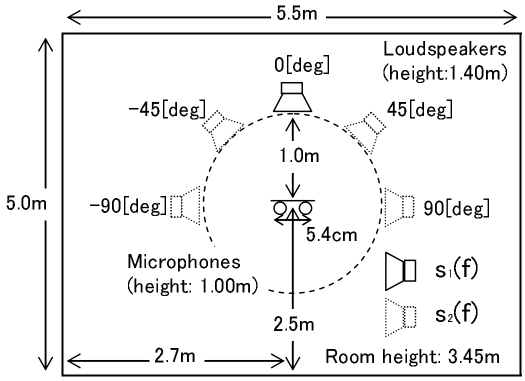
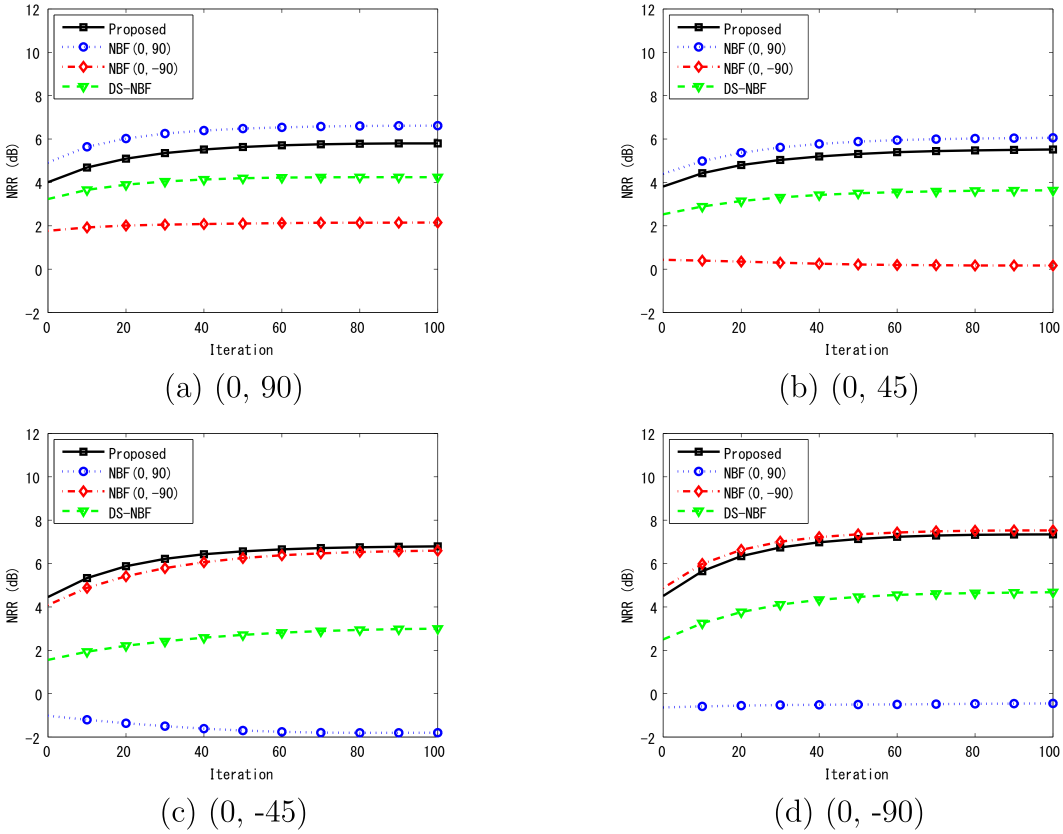


Figure 3: Recording environment in reverberant room. Interference source is located either −90, −45, 45, or 90 degrees, while target point source is placed at 0 degree.



(a) (0, 90)

(b) (0, 45)

(c) (0, -45)

(d) (0, -90)

Figure 4: Noise reduction rate (NRR) as a function of FDICA iteration in reverberant room. DOA of true sources are shown in the bracket (θ_1, θ_2).

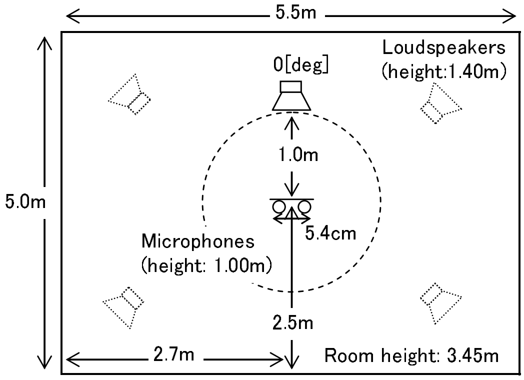


Figure 5: Recording environment in diffuse interfering-noise case.

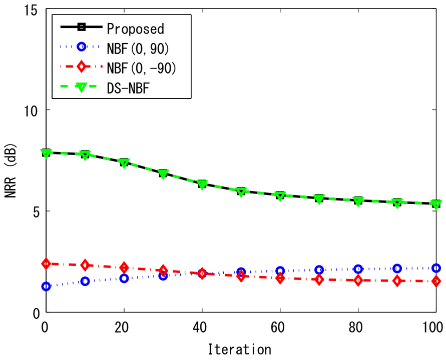


Figure 6: Source separation result in interfering diffuse-source case.

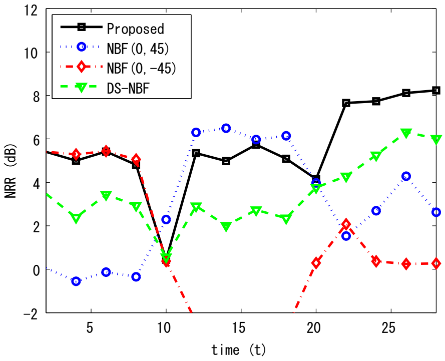


Figure 7: Source separation results in varying environmental conditions.

when compared with conventional beamformer-based initialization methods.

To extend the current algorithm for many source signals is an important future work. A possible approach for this extension is to increase the number of microphones. However, this straightforward extension also increases the computational cost and memory size. Thus, to deal with many microphone in a mobile device, developing computationally efficient FDICA with less memory consumption is necessary.

Acknowledgements

The authors would like to thank Prof. Hiroshi Saruwatari and Dr. Yu Takahashi for their helpful comments. MS was supported by MEXT KAKENHI 23300069.

References

[1] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1–3):21–34, 1998.

[2] S. Ikeda and N. Murata. A method of ICA in time-frequency domain. In *Proceedings of International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 365–371, 1999.

[3] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano. Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, 2003(11):1135–1146, 2003.

[4] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, Y. Ikeda, H. Hashimoto, and T. Morita. Blind separation of acoustic signals combining simo-model-based independent component analysis and binary masking. *EURASIP Journal on Applied Signal Processing*, 2006:1–17, 2006.

[5] Y. Mori, H. Saruwatari, T. Takatani, K. Shikano, T. Hiekata, and T. Morita. ICA and binary-mask-based blind source separation with small directional microphones. In *Proceedings of ICA*, pages 649–657, 2006.

[6] T. Hiekata, T. Morita, Y. Ikeda, H. Hashimoto, R. Zhang, Y. Takahashi, H. Saruwatari, and K. Shikano. Multiple ICA-based real-time blind source extraction applied to handy size microphone. In *Proceedings of the IEEE International Conference on Audio Speech and Signal Processing*, pages 121–124, 2009.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino. *Frequency domain blind source separation.* Springer, 2005.

[8] L. Parra and C. Alvino. Geometric source separation: merging convolutive source separation with geometric beamforming. In *Proceedings of the IEEE Signal Processing Society Workshop*, pages 273–282, 2001.

[9] G. W. Taylor, M. L. Seltzer, and A. Acero. Maximum a posteriori ICA:applying prior knowledge to the separation of acoustic sources. In *Proceedings of the IEEE International Conference on Audio Speech and Signal Processing*, pages 1821–1824, 2008.

[10] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, 12(5):530–538, 2004.

[11] M. Ogasawara, T. Nishino, and K. Takeda. Blind source separation using dodecahedral microphone array under reverberant conditions. *IEICE Transactions*, 94-A(3):897–906, 2011.

[12] H. Attias. *Source separation with a sensor array using graphical models and subband filtering.* MIT Press, Cambridge, MA, 2003.

[13] T. Kim, H.-T. Attias, S-Y. Lee, and T-W. Lee. Blind source separation exploiting higher-order frequency dependencies. *IEEE Transactions on Audio, Speech & Language Processing*, 15(1):70–79, 2007.

[14] A. Hiroe. Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. In *Proceedings of ICA*, pages 601–608, 2006.

[15] R. Mukai, H. Sawada, S. Araki, and S. Makino. Robust real-time blind source separation for moving speakers in a room. In *Proceedings of the IEEE International Conference on Audio Speech and Signal Processing*, pages 469–472, 2003.

[16] Y. Takahashi, T. Takatani, H. Saruwatari, and K. Shikano. Blind spatial subtraction array with independent component analysis for hands-free speech recognition. In *Proceedings of the IEEE International Workshop on Acoustic Echo and Noise Control*, Paris, France, 2006.

[17] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano. Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Transactions on Audio, Speech & Language Processing*, 17(4):650–664, 2009.

[18] J. Li, P. Stoica, and Z. Wang. On robust Capon beamforming and diagonal loading. *IEEE Transactions on Signal Processing*, 51(7):1702–1715, 2003.

[19] S. Amari, A. Chichocki, and H. H. Yang. *A new learning algorithm for blind signal separation.* MIT Press, Cambridge, MA, 1996.

[20] P. Stoica, Z. Wang, and J. Li. Robust Capon beamforming. *IEEE Transactions on Signal Processing Letters*, 10(6):172–175, 2003.

[21] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley Interscience, 2001.