Blind Source Separation for Robot Audition using fixed HRTF beamforming

Mounira Maazaoui, Karim Abed-Meraim, Yves Grenier

EURASIP Journal on Advances in Signal Processing

SpringerOpen

Blind source separation for robot audition using fixed HRTF beamforming

EURASIP Journal on Advances in Signal Processing 2012,
2012:58 doi:10.1186/1687-6180-2012-58

Mounira Maazaoui (maazaoui@telecom-paristech.fr)
Karim Abed-Meraim (abed@telecom-paristech.fr)
Yves Grenier (yves.grenier@telecom-paristech.fr)

ISSN 1687-6180

Article type Research

Submission date 15 June 2011

Acceptance date 6 March 2012

Publication date 6 March 2012

Article URL http://asp.eurasipjournals.com/content/2012/1/58

Blind source separation for robot audition using fixed HRTF beamforming

Mounira Maazaoui, Karim Abed-Meraim and Yves Grenier

Institute TELECOM, TELECOM ParisTech, CNRS-LTCI 37/39,

rue Dareau, 75014, Paris, France

Email addresses:

maazaoui@telecom-paristech.fr

abed@telecom-paristech.fr

yves.grenier@telecom-paristech.fr

Abstract

In this article, we present a two-stage blind source separation (BSS) algorithm for robot audition. The first stage consists in a fixed beamforming preprocessing to reduce the reverberation and the environmental noise. Since we are in a robot audition context, the manifold of the sensor array in this case is hard to model due to

the presence of the head of the robot, so we use pre-measured head related transfer functions (HRTFs) to estimate the beamforming filters. The use of the HRTF to estimate the beamformers allows to capture the effect of the head on the manifold of the microphone array. The second stage is a BSS algorithm based on a sparsity criterion which is the minimization of the l_1 norm of the sources. We present different configuration of our algorithm and we show that it has promising results and that the fixed beamforming preprocessing improves the separation results.

1 Introduction

Robot audition consists in the aptitude of an humanoid to understand its acoustic environment, separate and localize sources, identify speakers and recognize their emotions. This complex task is one of the target points of the ROMEO project[a] that we work on. This project aims to build an humanoid (ROMEO) that can act as a comprehensive assistant for persons suffering from loss of autonomy. Our task in this project is focused on the blind source separation (BSS) topic using a microphone array (more than two sensors). Source separation is a very important step for human-robot interaction: it allows latter tasks like speakers identification, speech and motion recognition and environmental sound analysis to be achieved properly. In a BSS task, the separation should be done from the

received microphone signals without prior knowledge of the mixing process. The only knowledge is limited to the array geometry.

The problem of BSS has been studied by many authors [1], and we present here some of the state-of-the-art methods related to robot audition. Tamai et al. [2] performed sound source localization by a delay and sum beamforming and source separation in a real environment with frequency band selection using a microphone array located on three rings with 32 microphones. Yamamoto et al. [3] proposed a source separation technique based on geometric constraints as a preprocessing for the speech recognition module in their robot audition system. This system was implemented in the humanoids SIG2 and Honda ASIMO with an eight sensors microphone array, as a part of a more complete system for robot audition named HARK [4]. Saruwatari et al. [5] proposed a two-stage binaural BSS system for an humanoid. They combined a single-input multiple-output model based on independent component analysis (ICA) and a binary mask processing.

One of the main challenges of BSS remains to obtain good BSS performance in a real reverberant environments. A beamforming preprocessing can be a solution to improve BSS performance in a reverberant room. Beamforming consists in estimating a spatial filter that operates on the outputs of a microphone array in order to form a beam with a desired directivity pattern [6]. It is useful for

many purposes, particularly for enhancing a desired signal from its measurement corrupted by noise, competing sources and reverberation [6]. Beamforming filters can be estimated in a fixed or in an adaptive way. A fixed beamforming, contrarily to an adaptive one, does not depend on the sensors data, the beamformer is built for a set of fixed desired directions. In this article, we propose a two-stage BSS technique where a fixed beamforming is used in a preprocessing step.

Ding et al. propose to use a beamforming preprocessing where the steering directions are the directions of arrival (DOA) of the sources. In this case, the DOA of the sources are supposed to be known *a priori* [7]. The authors evaluate their method in a determined case with 2 and 4 sources and a circular microphone array. Saruwatari et al. present a combined ICA [8] and beamforming method: first the authors perform a subband ICA and estimate the direction of arrivals (DOA) of the sources using the directivity patterns in each frequency bin, second they use the estimated DOA to build a null beamforming, and third they integrate the subband ICA and the null beamforming by selecting the most suitable separation matrix in each frequency [9]. In this article, we propose to use a fixed beamforming preprocessing with fixed steering directions, independently from the direction of arrival of the sources, and we compare this preprocessing method to the one proposed by Wang et al. We are interested in studying the effect of the beamform-

ing as a preprocessing tool so we are not going to include the algorithm of [9] in our evaluation (the authors of [9] use the beamforming as a separation method alternatively with ICA).

However, in a beamforming task, we need to know the manifold of the sensor array, which is hard to model for the robot audition case because the head of the robot alters the acoustic near field. To overcome the problem of the array geometry modeling and take into account the influence of the robot's head on the received signals, we propose to use the head related transfer functions (HRTFs) of the robot's head as steering vectors to build the fixed beamformer. The main advantages of our method are its reduced computational cost (as compared to the one based on adaptive beamforming), its improved separation quality and its relatively fast convergence rate. Its weaknesses consist in the lack of theoretical analysis or proofs that guarantee the convergence to the desired solution and in the case where source localization is needed, our method provides only a rough estimation of the direction of arrival.

This article is organized as follows: in Section 2, we present the signal model used in the BSS task, Sections 4 and 3 are dedicated respectively for the beamforming using HRTF step and for the presentation of the BSS using sparsity criterion step, and we assess the algorithms performances in Section 5, while Section

6 provides some concluding remarks.

2 Signal model

Assume N sound sources $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ and an array of M microphones with outputs denoted by $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$, where t is the time index. We assume that we are in an overdetermined case with $M > N$ and that the number of sources N is known *a priori*. In Section 3.3 however, we propose a method of source number estimation in the robot audition case. As we are in a real environment context, the output signals in the time domain are modeled as the sum of the convolution between the sound sources and the impulse responses of the different propagation paths between the sources and the sensors, truncated at the length of $L+1$:

$$\mathbf{x}(t) = \sum_{l=0}^{L} \mathbf{h}(l)\,\mathbf{s}(t-l) + \mathbf{n}(t) \tag{1}$$

where $\mathbf{h}(l)$ is the l^{th} matrix of impulse response and $\mathbf{n}(t)$ is a noise vector. We consider a spatially decorrelated diffuse noise which energy is supposed to be negligible comparing to the punctual sources ones. If the noise is punctual, it will be considered as a sound source. This scenario corresponds to our experimental

and real life application setups.

In the frequency domain, when the length of the analysis window N_f of the short time fourier transform (STFT) is longer than twice the length of the mixing filter L, the output signals at the time-frequency bin (f,k) can be approximated as:

$$\mathbf{X}(f,k) \simeq \mathbf{H}(f)\,\mathbf{S}(f,k) \tag{2}$$

where $\mathbf{X}(f,k) = [X_1(f,k),\ldots,X_M(f,k)]^H$ $\left(\text{respectively } \mathbf{S}(f,k) = [S_1(f,k),\ldots,S_N(f,k)]^H\right)$ is the STFT of $\{\mathbf{x}(t)\}_{1\leq t\leq T}$ $\left(\text{respectively } \{\mathbf{s}(t)\}_{1\leq t\leq T}\right)$ in the frequency bin $f \in \left[1, \frac{N_f}{2}+1\right]$ and the time bin $k \in [1, N_T]$, and $\mathbf{H}$ is the Fourier transform of the mixing filters $\{\mathbf{h}(l)\}_{0\leq l\leq L}$. Using an appropriate separation criterion, our objective is to find for each frequency bin a separation matrix $\mathbf{F}(f)$ that leads to an estimation of the original sources in the time-frequency domain:

$$\mathbf{Y}(f,k) = \mathbf{F}(f)\,\mathbf{X}(f,k) \tag{3}$$

The inverse STFT of the estimated sources in the frequency domain $\mathbf{Y}$ allows the recovery of the estimated sources $\mathbf{y}(t) = [y_1(t),\ldots,y_N(t)]^T$ in the time domain.

Separating the sources for each frequency bin introduces the permutation problem: the order of the estimated sources is not the same from one frequency to

another. To solve the permutation problem, we use the method proposed by Weihua and Fenggang and described in [10]. This method is based on the signals correlation between two adjacent frequencies. In this article, we are not going to investigate the permutation problem and we use the cited method for all the proposed algorithm.

The separation matrix $\mathbf{F}(f)$ is estimated using a two-step blind separation algorithm: a fixed beamforming preprocessing step and a BSS step (cf. Figure 1). $\mathbf{F}(f)$ is written as the combination of the results of those two steps:

$$\mathbf{F}(f) = \mathbf{W}(f)\mathbf{B}(f) \tag{4}$$

where $\mathbf{W}(f)$ is the separation matrix estimated using a sparsity criterion and $\mathbf{B}(f)$ is a fixed beamforming filter. More details are presented in the following subsections (cf. Algorithm 1).

2.1 Beamforming preprocessing

The role of the beamformer is essentially to reduce the reverberation and the interferences coming from directions other than the looked up ones. Once the reverberation is reduced, Equation (2) is better satisfied which leads to an improved BSS quality.

We consider $\{\mathbf{B}(f)\}_{1\leq f\leq \frac{N_f}{2}+1}$ a set of fixed beamforming filters of size $K \times M$, where K is the number of the desired beams, $K \geq N$. Those filters are calculated beforehand (before the beginning of the processing) and used in the beamforming preprocessing step (cf. Section 3). The outputs of the beamformers at each frequency f are:

$$\mathbf{Z}(f,k) = \mathbf{B}(f)\,\mathbf{X}(f,k) \tag{5}$$

2.2 Blind source separation

The BSS step consists in estimating a separation matrix $\mathbf{W}(f)$ that leads to separated sources at each frequency bin f. The separation matrix $\mathbf{W}(f)$ is estimated by minimizing, with respect to $\mathbf{W}(f)$, a cost function ψ based on a sparsity criterion, under a unit norm constraints for $\mathbf{W}(f)$. The chosen optimization technique is the natural gradient (cf. Section 4). The separation matrix is estimated from the output signals of the beamformers $\mathbf{Z}(f,k)$ and the estimated sources are then written as:

$$\mathbf{Y}(f,k) = \mathbf{W}(f)\,\mathbf{Z}(f,k) \tag{6}$$

3 Fixed beamforming using HRTF

In the case of robot audition, the geometry of the microphone array is fixed once for all. To build the fixed beamformers, we need to determine the "desired" steering directions and the characteristics of the beam pattern (the beamwidth, the amplitude of the sidelobes and the position of nulls). The beamformers are estimated only once for all scenarii using these spatial information and independently of the measured mixture in the sensors.

The least-square (LS) technique is used [6] to estimate the beamformer filters that will achieve the desired beam pattern according to a desired direction response. To accomplish this beamformers estimation, we need to calculate the steering vectors which represent the phase delays of a plane wave evaluated at the microphone array elements.

In the free field, the steering vector of an M elements array at a frequency f and for a steering direction θ is known. For example, for a linear array, we have:

$$\mathbf{a}(f,\theta) = \left[1, e^{-j2\pi f \frac{d}{c}\sin\theta}, \ldots, e^{-j2\pi f \frac{d}{c}(M-1)\sin\theta}\right]^T \tag{7}$$

where d is the distance between two sensors and c is the speed of sound.

In the case of robot audition, the microphones are often fixed in the head of the robot (cf. Figure 2). The free field model of the steering vectors presented in

Equation (7) does not take into account the influence of the head on the surrounding acoustic fields, and in this case, the microphone array manifold is not modeled (unknown).

For a human hearing, there is a spectral filtering of the sound source by the head and the pinna, and thus a transfer function between the source and each ear is defined and refered to as: the HRTF. The HRTF takes into account the interaural time difference[b] (ITD), the interaural intensity difference[c] (IID) and the shape of the head and the pinna. It defines how a sound emitted from a specific location and altered by the head and the pinna is received at an ear. The notion of HRTF remains the same if we replace the human head by a dummy head and the ears by two microphones. We extend the usual concept of binaural HRTF to the context of robot audition where the humanoid is equipped with a microphone array. In our case, a HRTF $h_m(f,\theta)$ at frequency f characterizes how a signal emitted from a specific direction θ is received at the mth sensor fixed in a head.

We propose to use the HRTFs as steering vectors for the beamformer filters calculation (cf. figure 3) and replace the unknown array manifold by a discrete distribution of HRTFs on a group of N_S a priori chosen steering directions $\Theta = \{\theta_1, \ldots, \theta_{N_S}\}$. The HRTFs are measured in an anechoïc room as explained in Section 5.

Let $h_m(f,\theta)$ be the HRTF at frequency f from the emission point located at θ to the mth sensor. The steering vector is then:

$$\mathbf{a}(f,\theta) = [h_1(f,\theta),\ldots,h_M(f,\theta)]^T \tag{8}$$

Given Equation (8), one can express the normalized LS beamformer for a desired direction θ_i as [6]:

$$\mathbf{b}(f,\theta_i) = \frac{\mathbf{R}_{\mathbf{aa}}^{-1}(f)\,\mathbf{a}(f,\theta_i)}{\mathbf{a}^H(f,\theta_i)\,\mathbf{R}_{\mathbf{aa}}^{-1}(f)\,\mathbf{a}(f,\theta_i)} \tag{9}$$

where $\mathbf{R}_{\mathbf{aa}}(f) = \frac{1}{N_S}\sum_{\theta\in\Theta}\mathbf{a}(f,\theta)\,\mathbf{a}^H(f,\theta)$. Given K desired steering directions $\theta_1,\ldots,\theta_K$, the beamforming matrix $\mathbf{B}(f)$ is:

$$\mathbf{B}(f) = [\mathbf{b}(f,\theta_1),\ldots,\mathbf{b}(f,\theta_K)]^T \tag{10}$$

In the following, we present the different configurations of the combined beamforming-BSS algorithm.

3.1 Beamforming with known DOA

If the direction-of-arrivals (DOAs) of the sources are known *a priori*, mainly by a source localization method, the beamforming filters are estimated using this spatial information of the sources location (cf. Figure 4). Therefore, the desired directions are the DOAs of the sources and we select the corresponding HRTFs

to build the desired response vectors $\mathbf{a}(f,\theta)$. This is an ideal method to compare our results with. Indeed, we consider that source localization is beyond the scope of this article (in [7] where the beamforming with known DOAs was proposed for a circular microphone array, the authors have assumed that the DOAs are known *a priori*).

3.2 Beamforming with fixed DOA

Estimating the DOAs of the sources to build the beamformers is time consuming and not always accurate in the reverberant environments. So we propose to build K fixed beams with arbitrary desired directions chosen such as they cover all the useful space directions (cf. Figure 5). We use the output of all the beamformers directly in the BSS algorithm. In this case, we still have an overdetermined separation problem with N sources and K mixtures.

3.3 Beamforming with beams selection

In this configuration, we still have K fixed beams with arbitrary desired directions, but we are not going to use all the outputs of those beamformers (cf. figure 6). We select the N beamformer outputs with the highest energy, corresponding to the beams that are the closest to the sources (we suppose that the energies of the

sources are quite close to each other). In this case, after beamforming, we are in a determined separation problem with N sources and $K = N$ mixtures (cf. Algorithm 2).

Fixed beamforming with beams selection can be derived and used for the source number as well as a rough DOAs estimation. We fix a maximum number of sources $N_{\max} < K$. In each frequency bin, after the beamforming filtering (5), we select the $N_{\max}$ beams with the highest energies (instead of selecting N beams as in the previous paragraph). Then, we build over all the selected steering directions a histogram that corresponds to their overall number of occurrence (cf. Figure 7). After a thresholding, we select the beams corresponding to the peaks (a peak corresponds to a local maximum point associated to the number of selected beams over all the frequencies). The filters that correspond to those beams are our final beamforming filters, the number of peaks correspond to the number of sources and the corresponding steering directions provide us with a rough estimation of the DOAs.

4 BSS using sparsity criterion

In the BSS step, we estimate the separation matrix $\mathbf{W}(f)$ by minimizing, with respect to the separation matrix $\mathbf{W}(f)$, a cost function ψ based on a sparsity criterion, under a unit norm constraint for $\mathbf{W}(f)$:

$$\min_{\mathbf{W}} \psi(\mathbf{W}(f)) \text{ such that } \|\mathbf{W}(f)\| = 1 \tag{11}$$

The optimization technique used to update the separation matrix $\mathbf{W}(f)$ is the natural gradient. Section 4.1 summarizes the natural gradient algorithm [11], Section 4.2 shows how we use this optimization algorithm in our cost function.

4.1 Natural gradient algorithm

The natural gradient is an optimization method proposed by Amari et al. [11]. In this modified gradient search method, the standard gradient search direction is altered according to the local Riemannien structure of the parameter space. This guarantees the invariance of the natural gradient search direction to the statistical relationship between the parameters of the model and leads to a statistically efficient learning performance [12].

Assume that we want to update a separation matrix $\mathbf{W}$ according to a loss

function $\psi(\mathbf{W})$. The gradient update of this matrix is given by:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu \nabla \psi(\mathbf{W}_t) \tag{12}$$

where $\nabla \psi(\mathbf{W})$ is the gradient of the function $\psi(\mathbf{W})$ and t refers to the iteration (or time) index. From [12], the natural gradient of a loss function $\psi(\mathbf{W})$, noted $\tilde{\nabla} \psi(\mathbf{W})$, is given by:

$$\tilde{\nabla} \psi(\mathbf{W}) = \nabla \psi(\mathbf{W}) \mathbf{W}^H \mathbf{W} \tag{13}$$

The natural gradient update of the separation matrix $\mathbf{W}$ is then:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu \nabla \psi(\mathbf{W}_t) \mathbf{W}_t^H \mathbf{W}_t \tag{14}$$

4.2 Sparsity separation criterion

Speech signal is known to be sparse in the time-frequency domain: the number of time-frequency points where the speech signal is active (i.e., of non negligeable energy) is small comparing to the total number of time-frequency points (cf. Figure 8).

We consider a separation criterion based on the sparsity of the signals in the time-frequency domain. For every frequency bin, we look for a separation matrix $\mathbf{W}(f)$ that leads to the sparsest estimated sources $\mathbf{Y}(f,:) = [\mathbf{Y}(f,1), \ldots, \mathbf{Y}(f,N_T)]$.

In the same manner, we define the mixture matrix in each frequency bin $\mathbf{X}(f,:) = [\mathbf{X}(f,1),\ldots,\mathbf{X}(f,N_T)]$.

To measure the sparsity of a signal, the l_1 norm is the most used sparsity measure thanks to its convexity [13]. The smaller is the l_1 norm of a signal, the sparser it is. However, the l_1 norm is not the only measure of sparsity [13]. We presented recently a parameterized l_p norm algorithm for BSS, where we made the sparsity constraint harder through the iterations of the optimization process [14]. In this article, we use the l_1 norm to measure the sparsity of signal $\mathbf{Y}(f,:)$, and hence the cost function is:

$$\psi(\mathbf{W}(f)) = \sum_{i=1}^{N}\sum_{k=1}^{N_T} |Y_i(f,k)| \tag{15}$$

To have the sparsest estimated sources, we should minimize $\psi(\mathbf{W}(f))$ and we use the natural gradient search technique to find the optimum separation matrix $\mathbf{W}(f)$:

$$\mathbf{W}_{t+1}(f) = \mathbf{W}_t(f) - \mu\nabla\psi(\mathbf{W}_t(f))\,\mathbf{W}_t^H(f)\,\mathbf{W}_t(f) \tag{16}$$

The differential of $\psi(\mathbf{W}(f))$ is:

$$\mathrm{d}\psi(\mathbf{W}(f)) = \mathbf{f}(\mathbf{Y}(f,:))\,\mathrm{d}\mathbf{Y}^H(f,:) \tag{17}$$

where $\mathbf{f}(\mathbf{Y}(f,:)) = \text{sign}(\mathbf{Y}(f,:))$ is a matrix with the same size as $\mathbf{Y}(f,:)$ in which the (i,j)th entry is $\text{sign}(Y_i(f,j))$.[d] Thus, the gradient of $\psi(\mathbf{W})$ is expressed as:

$$\nabla\psi(\mathbf{W}(f)) = \mathbf{f}(\mathbf{Y}(f,:))\,\mathbf{X}^H(f,:) \tag{18}$$

which gives the expression of the natural gradient of $\psi(\mathbf{W}_t(f))$:

$$\begin{aligned}\tilde{\nabla}\psi(\mathbf{W}_t(f)) &= \nabla\psi(\mathbf{W}_t(f))\,\mathbf{W}_t^H(f)\,\mathbf{W}_t(f)\\ &= \mathbf{f}(\mathbf{Y}_t(f,:))\,\mathbf{Y}_t^H(f,:)\,\mathbf{W}_t(f)\end{aligned} \tag{19}$$

The update equation of $\mathbf{W}_t(f)$ for a frequency bin f is then:

$$\mathbf{W}_{t+1}(f) = \mathbf{W}_t(f) - \mu\mathbf{G}_t(f)\,\mathbf{W}_t(f) \tag{20}$$

with $\mathbf{G}_t(f) = \mathbf{f}(\mathbf{Y}_t(f,:))\,\mathbf{Y}_t^H(f,:)$.

The convergence of the natural gradient is conditioned both by the initial coefficients $\mathbf{W}_0(f)$ of the separation matrix and the step size of the update and it is quite difficult to choose the parameters that allow fast convergence without risking divergence. Douglas and Gupta [15] proposed to impose a scaling constraint to the separation matrix $\mathbf{W}_t(f)$ to maintain a constant gradient magnitude along the algorithm iterations. They assert that with this scaling and a fixed step size μ,

the algorithm has fast convergence and excellent performance independently of the magnitude of $\mathbf{X}(f,:)$ and $\mathbf{W}_0(f)$. Applying this scaling constraint, our update function becomes:

$$\mathbf{W}_{t+1}(f) = c_t(f)\,\mathbf{W}_t(f) - \mu c_t^2(f)\,\mathbf{G}_t(f)\,\mathbf{W}_t(f) \tag{21}$$

with $c_t(f) = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_t^{ij}(f)\right|}$ and $g_t^{ij}(f) = [\mathbf{G}_t(f)]_{ij}$.

4.3 Initialization

When we are in an overdetermined case, we use a whitening process for the initialization of the separation matrix $\mathbf{W}_0$. The whitening is an important preprocessing in an overdetermined BSS algorithm as it allows to focus the energy of the received signals in the useful signal space. The separation matrix is initialized as follow:

$$\mathbf{W}_0 = \sqrt{\mathbf{D}_M^{-1}}\mathbf{E}_{:M}^H$$

where $\mathbf{D}_M$ is a matrix containing the first M rows and M columns of the matrix $\mathbf{D}$ and $\mathbf{E}_{:M}$ is the matrix containing the first M columns of the matrix $\mathbf{E}$. $\mathbf{D}$ and $\mathbf{E}$ are respectively the diagonal matrix and the unitary matrix of the singular value

decomposition of the autocorrelation matrix of the received data $\mathbf{X}(f,:)$ or the filtered data after beamforming $\mathbf{Z}(f,:)$.

If we are in a determined case, in particular when we select the beams with the highest energy after the beamforming filtering or when the steering directions correspond to the direction of arrivals of the sources, the initialization of the separation matrix is done with the identity matrix:

$$\mathbf{W}_0 = \mathbf{I}_N$$

5 Experimental results

5.1 Experimental database

To evaluate the proposed BSS techniques, we built two databases: a HRTFs database and a speech database.

5.1.1 HRTF database

We recorded the HRTF database in the anechoic room of Telecom ParisTech (cf. Figure 2) using the Golay codes process [16]. As we are in a robot audition context, we model the future robot by a child size dummy ($1m20$) for the sound acquisition process, with 16 sensors fixed in its head (cf. Figure 9).

We measured 504 HRTF for each microphone as follow:

- 72 azimuth angles from 0° to 355° with a 5° step
- 7 elevation angles: $-40°$, $-27°$, 0°, 20°, 45°, 60° and 90°

To measure the HRTFs, the dummy was fixed on a turntable in the center of the loudspeaker arc in the anechoic room (cf. Figure 2). For each azimuth angle, a sequence of complementary Golay codes is emitted sequentially from each loudspeaker (this is to vary the elevation) and recorded with the 16 sensors array. This operation was repeated for all the azimuth angles. The Golay complementary sequences have the useful property that their autocorrelation functions have complementary sidelobes: the sum of the autocorrelation sequences is exactly zero everywhere except at the origin. Using this property and the recorded complementary Golay codes, the HRTF are calculated as in [16].

Details about the experimental process of HRTF calculation as well as the HRTF databases at the sampling frequencies of 48 and 16 KHz are available at `http://www.tsi.telecom-paristech.fr/aao/?p=347`.

5.1.2 Test database

The test signals were recorded in a moderately reverberant room where the reverberation time is $\mathrm{RT}_{30} = 300\,\mathrm{ms}$ (cf. Figure 10). Figure 11 shows the different

positions of the sources in the room. We chose to evaluate the proposed algorithm on a separation of two sources: the first source is always the one placed at 0° and the second source is chosen from 20° to 90°.

The output signals $\mathbf{x}(t)$ are the convolutions of 40 pairs of speech sources (male and female speaking French and English) by two of the impulse responses $\{\mathbf{h}(l)\}_{0\leq l\leq L}$ measured for the direction of arrivals presented in Figure 11.

The characteristics of the signals and the BSS algorithms are summarized in Table 1.

5.2 Evaluation results

In this section, we evaluate different configurations of the presented algorithm[e]:

(1) The beamforming stage only: beamforming of 37 lobes from $-90°$ to $90°$ with a step angle of 5° (BF[5°])

(2) The BSS algorithm only

   (a) with minimization of the l_1 norm (BSS-l_1)

   (b) with ICA from [15] (ICA)

(3) The two-stage algorithm, BSS and the beamforming preprocessing:

(a) beamforming of N lobes in the DOA of the sources (BF[DOA]+BSS-l_1)

(b) beamforming of 7 lobes from $-90°$ to $90°$ with a step angle of $30°$ (BF[$30°$] + BSS-l_1 when the l_1 norm minimization is used in the BSS step and BF[$30°$] + ICA when ICA is used in the BSS step)

(c) beamforming of 13 lobes from $-90°$ to $90°$ with a step angle of $15°$ (BF[$15°$] + BSS-l_1)

(d) beamforming of 19 lobes from $-90°$ to $90°$ with a step angle of $10°$ (BF[$10°$] + BSS-l_1)

(e) beamforming of 37 lobes from $-90°$ to $90°$ with a step angle of $5°$ (BF[$5°$] + BSS-l_1)

(f) beamforming of 7 lobes from $-90°$ to $90°$ with a step angle of $30°$ with selection of the N beams containing the highest energy before proceeding the BSS (BF[$30°$] + BS + BSS-l_1)

(g) beamforming of 37 lobes from $-90°$ to $90°$ with a step angle of $5°$ with selection of the N beams containing the highest energy before proceeding the BSS (BF[$5°$] + BS + BSS-l_1)

We evaluate the proposed two-stage algorithm by the signal-to-interference ratio (SIR) and the ratio (SDR) estimated using the BSS-eval toolbox [17]. All the

presented curves are the average result of the 40 pairs of speech.

5.2.1 Influence of the beamforming preprocessing

Figures 12 and 13 show that the SIR and SDR of the two-stage algorithm with the fixed beamforming preprocessing BF[5°] + BSS-l_1 and BF[30°] + BSS-l_1 are better than the SIR and SDR of the separation algorithm with l_1 norm alone BSS-l_1 and much better than the ones we obtain by the fixed beamforming BF[5°] only. The SIR and SDR of the received signals in microphones 1 and 2 (labeled as *sensors data* in the figures) is taken as reference to illustrate the performance gain of our method. However this increase in the SIR and SDR by the fixed beamforming preprocessing is limited and do not reach the performance of the beamforming preprocessing with known DOA BF[DOA] + BSS-l_1 as shown in Figures 14 and 15. But we can overcome this limitation by the beam selection as shown in the sequel.

Figures 16 and 17 show the SIR and SDR obtained with different inter-beam angle of the beamforming preprocessing, the steering directions vary from $-90°$ to 90°: beamforming with 7 beams with a step angle of 30° (BF[30°] + BSS-l_1), beamforming of 13 beams with a step angle of 15° (BF[15°] + BSS-l_1), beamforming of 19 beams with a step angle of 10° (BF[10°] + BSS-l_1) and beamform-

ing with 37 beams with a step angle of 5°. The results show that when we increase the number of the beams, the SIR and especially the SDR increases. For BF[15°] + BSS-l_1, BF[10°] + BSS-l_1 and BF[5°] + BSS-l_1, the beamforming preprocessing increases the SDR of the estimated sources comparing with the single stage BSS-l_1 algorithm. The SIR with a beamforming preprocessing is also better than the single stage BSS-l_1 algorithm, and this for all the tested configurations of the fixed steering direction beamforming prepossessing.

Influence of the beams selection

As we can observe from Figures 12, 13, 14, and 15, the beamforming preprocessing with beams selection (BF[30°] + BS + BSS-l_1 and BF[5°] + BS + BSS-l_1) and the beamforming preprocessing with known direction of arrivals (BF[DOA] + BSS-l_1) have close results in terms of SIR (cf. Figures 12 and 14) and SDR (cf. Figures 13 and 15). However, if we are in a reverberant environment where the direction of arrivals can not be estimated accurately, the beamforming preprocessing with beams selection would be a good solution to improve the SIR and the SDR of the estimated sources comparing to the use of the BSS algorithm only (BSS-l_1).

Comparing BF[5°] + BS + BSS-l_1 in Figure 12 and BF[30°] + BS + BSS-l_1 in Figure 14 show that the impact of the inter-beam angle is quite weak with respect

to the separation gain. However, the beamforming preprocessing with beams selection of 5° inter-beam angle step allows us to estimate correctly the DOA of the sources with a step of 5° as shown in Figure 18. The latter represents the selected beam directions for all considered experiments (i.e., the 40 experiments) and for different source locations.

5.2.2 Comparison between BSS-l_1 and ICA

Independent component analysis and the l_1 norm minimization have quite close results with or without the preprocessing step. However, we believe that replacing BSS-l_1 by BSS-l_p with $p < 1$ or with varying p value might lead to a significant improvement of the separation quality. This observation is based on the preliminary results we obtained in [14] and would be the focus of future investigations.

5.2.3 Convergence analysis

We procceed to the analysis of the convergence of the proposed algorithm by observing the convergence rates through the iterations and for the considered DOA (cf. Figure 19). Each curve represents the average of cost function (15) averaged for all the frequencies. As we can see in Figure 19b, our iterative algorithm converges quite quickly (typically 10 to 20 iterations) towards its steady state. We

notice also that the convergence rate of the proposed two stage method with beam selection is better than the convergence of BSS-l_1. Indded, in this context, the separation algorithm BSS-l_1 converges to its steady state after 30 to 40 iterations. Moreover, the cost function of the two stage algorithm reaches lower values than the separation algorithm only and thus, the beamforming preprocessing helps for better convergence.

6 Conclusion

In this article, we present a two-stage BSS algorithm for robot audition. The first stage is a preprocessing step with fixed beamforming. To deal with the effect of the head of the robot in the acoustic near field and model the manifold of the sensors array, we used HRTFs as steering vectors in the beamformers estimation step. The second stage is a BSS algorithm exploiting the sparsity of the sources in the time-frequency domain.

We tested different configurations of this algorithm with steering directions of the beams equal to the direction of arrivals of the sources and with fixed steering directions. We also varied the step angle between the beams. The beamforming preprocessing improves the separation performance as it reduces the reverberation

and noise effects. The maximum gain is obtained when we select the beams with the highest energies and use the corresponding filters as beamformers or when the sources DOAs are known. The beamforming preprocessing with fixed steering directions has also good performance and does not use an estimation of the DOAs or beam selection, which represent a gain in the processing time. Using the 5° step beamforming preprocessing with beams selection, we can also have a rough estimation of the direction of arrivals of the sources.

Acknowledgement

This work is funded by the Ile-de-France region, the General Directorate for Competitiveness, Industry and Services (DGCIS) and the City of Paris, as a part of the ROMEO project.

Competing interests

The authors declare that they have no competing interests.

Endnotes

[a]Romeo project: www.projetromeo.com. [b]The ITD is the difference in arrival times of a sound wavefront at the left and right ears. [c]The IID is the amplitude difference of a sound that reaches the right and left ears. [d]For a complex number z, sign(z)$= \frac{z}{|z|}$. [e]The names of the algorithms that we are going to use in the legends of the figures are between brackets.

References

[1] Pierre Comon and Christian Jutten, *Handbook of Blind Source Separation , Independent Component Analysis and Applications*, Elsevier, 2010.

[2] Y. Tamai, Y. Sasaki, S. Kagami, and H. Mizoguchi, "Three ring microphone array for 3d sound localization and separation for mobile robot audition," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4172–4177, Aug. 2005.

[3] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J.-M. Valin, K. Komatani, T. Ogata, and H.G. Okuno, "Design and implementation of a robot audition system for automatic speech recognition of simultaneous speech," *IEEE Workshop on Automatic Speech Recognition Understanding*, pp. 111–116, 2007.

[4] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "High performance sound source separation adaptable to environmental changes for robot audition," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2165–2171, Sept. 2008.

[5] H. Saruwatari, Y. Mori, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and

T. Morita, “Two-stage blind source separation based on ica and binary masking for real-time robot audition system,” *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2303–2308, 2005.

[6] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone Array Signal Processing, Chapter 3: Conventional beamforming techniques*, Springer, 1rst edition, 2008.

[7] Heping Ding Lin Wang and Fuliang Yin, “Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010.

[8] Pierre Comon, “Independent component analysis, a new concept?,” *Signal Processing*, 1994.

[9] 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*, pp. 1135–1146, 2003.

[10] Wang Weihua and Huang Fenggang, "Improved method for solving permutation problem of frequency domain blind source separation," *6th IEEE International Conference on Industrial Informatics*, pp. 703–706, July 2008.

[11] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," *Advances in Neural Information Processing Systems*, pp. 757–763, 1996.

[12] Shun-Ichi Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, 1998.

[13] Hurley Niall and Rickard Scott, "Comparing measures of sparsity," *IEEE Workshop on Machine Learning for Signal Processing*, vol. 55, pp. 4723–4741, October 2009.

[14] M Maazaoui, Y Grenier, and K Abed-Meraim, "Frequency domain blind source separation for robot audition using a parameterized sparsity criterion," *19th European Signal Processing Conference, EUSIPCO*, 2011.

[15] S.C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 637–640, Apr. 2007.

[16] S. Foster, "Impulse response measurement using golay codes," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '86*, Apr. 1986, vol. 11, pp. 929 – 932.

[17] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462 –1469, July 2006.

Algorithm 1 Combined beamforming and BSS algorithm

1. **Input:**

   (a) The output of the microphone array $\mathbf{x} = [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_T)]$

   (b) The beamforming pre-calculated filters $\{\mathbf{B}(f)\}_{1 \leq f \leq \frac{N_f}{2}+1}$

2. $\{\mathbf{X}(f,k)\}_{1 \leq f \leq N_f, 1 \leq k \leq N_T} = \text{STFT}(\mathbf{x})$

3. for each frequency bin f

   (a) beamforming preprocessing step: $\mathbf{Z}(f,:) = \mathbf{B}(f)\mathbf{X}(f,:)$

   (b) initialization step: $\mathbf{W}(f) = \mathbf{W}_0(f)$

   (c) $\mathbf{Y}_0(f,:) = \mathbf{W}_0(f)\mathbf{Z}(f,:)$

   (d) for each iteration t:

   blind source separation step to estimate $\mathbf{W}(f)$

4. Permutation problem solving

5. **Output:** the estimated sources $\mathbf{y} = \text{ISTFT}\left(\{\mathbf{Y}(f,k)\}_{1 \leq f \leq N_f, 1 \leq k \leq N_K}\right)$

Algorithm 2 Beams selection algorithm

1. SelectedBeams = Ø

2. for each frequency bin f:

   (a) Form K beams (beamformer outputs) $\mathbf{Z}(f,:) = \mathbf{B}(f)\mathbf{X}(f,:)$, $\mathbf{Z}(f,:) = [\mathbf{z}_1(f,:),\ldots,\mathbf{z}_K(f,:)]^T$

   (b) Compute the energy of the beamformer outputs: $\mathbf{E}(f) = [e_1(f),\ldots,e_K(f)]$ with $e_i(f) = \frac{1}{N_T}\sum_{k=1}^{N_T}|\mathbf{z}_i(f,k)|^2$

   (c) Decreasing order sort of $\mathbf{E}(f)$, Beams are the beams corresponding to the sorted energies: Beams = $sort\,(\mathbf{E}(f))$

   (d) Select the N highest energies, the indexes are stored in B.

   (e) SelectedBeams = SelectedBeams $\cup$ B

3. Compute the frequency of appearance of each beam and store the occurrences in I.

4. Select the N beams with the highest occurrence

Table 1: Parameters of the blind source separation algorithms

Sampling frequency	16 KHz
Analysis window	Hanning
Analysis window length	2048
Shift length	1,024
μ	0.2
Signals length	5 s
Number of iterations	100

Figure 1: The processing scheme of the combined beamforming-BSS algorithm.

Figure 2: The dummy in the anechoïc room (left) and the microphone array of 16 sensors (right).

Figure 3: Example of a beam pattern using HRTFs for $\theta_i = 50^\circ$ (in dB).

Figure 4: Beamforming with known DOAs.

Figure 5: Beamforming with fixed steering directions (fixed lobes).

Figure 6: Beamforming with fixed steering directions and beams selection.

Figure 7: Estimation of the source number and DOAs using fixed beamforming: DOAs = 0° and 40°: we used $N_{\max} = 5$, 1024 frequency bins and an inter-beam angle for the fixed beamformers equal to 10°.

Figure 8: Sparsity of the speech signal in the time-frequency domain comparing to the time domain .**(a)** Speech sentence in the time domain **(b)** Time-frequency representation of the speech sentence

Figure 9: The detailed configuration of the microphone array.

Figure 10: Energy decay curve of the room used for the reverberant recording.

Figure 11: The position of the sources and their directions of arrival in the reverberant room.

Figure 12: SIR comparison in a real environment: source 1° is at 0° and source 2 varies from 20° to 90°— effect of the beamforming preprocessing on the SIR of the estimated sources.

Figure 13: SDR comparison in a real environment: source 1 is at 0° and source 2 varies from 20° to 90°—effect of the beamforming preprocessing on the SDR of the estimated sources.

Figure 14: SIR comparison in a real environment: source 1 is at 0° and source 2 varies from 20° to 90°.

Figure 15: SDR comparison in a real environment: source 1 is at 0° and source 2 varies from 20° to 90°.

Figure 16: SIR of different configuration of the beamforming preprocessing with fixed steering direction: inter-beams angles are 30°, 15°, 10°, and 5°, respectively.

Figure 17: SDR of different configuration of the beamforming preprocessing with fixed steering direction: inter-beams angles are 30°, 15°, 10°, and 5°, respectively.

Figure 18: DOA estimation using the BF[5°]+BS algorithm for the 40 experiments.

Figure 19: Convergence rates: the value of the cost function through the iterations and for different DOA. (a) BSS-l_1 **(b)** BF[5°]+BS+BSS-l_1

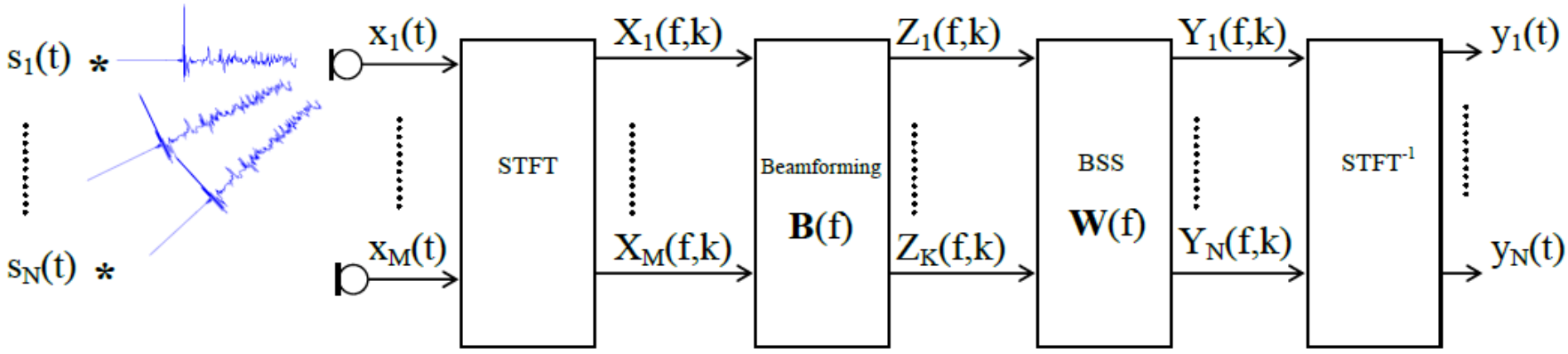
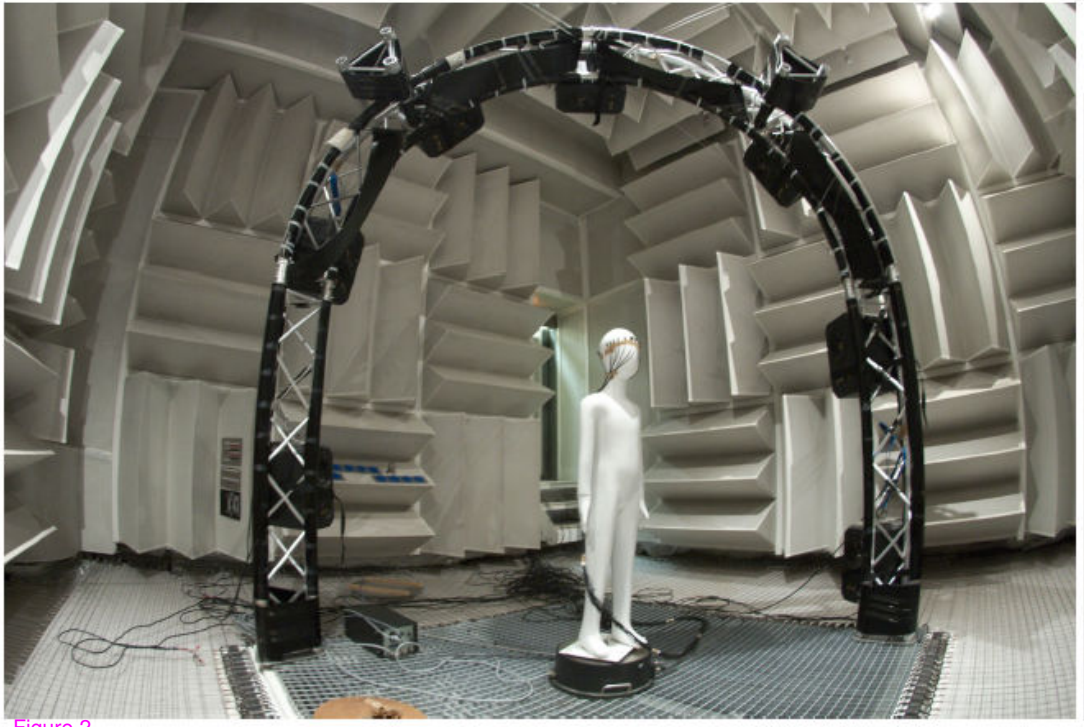

$s_1(t)$ *
$s_N(t)$ *
$x_1(t)$
$x_M(t)$
STFT
$X_1(f,k)$
$X_M(f,k)$
Beamforming
$\mathbf{B}(f)$
$Z_1(f,k)$
$Z_K(f,k)$
BSS
$\mathbf{W}(f)$
$Y_1(f,k)$
$Y_N(f,k)$
$STFT^{-1}$
$y_1(t)$
$y_N(t)$

Figure 1



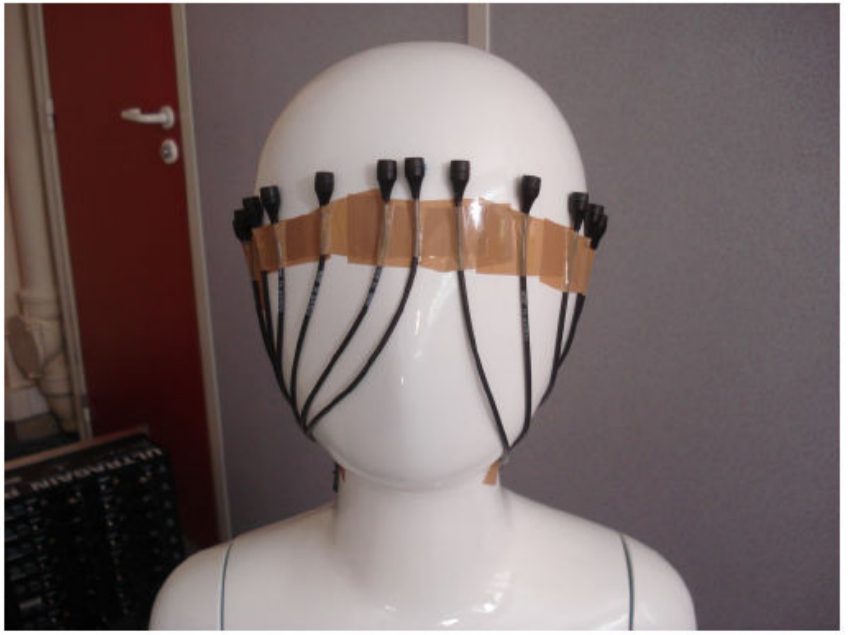


Figure 2

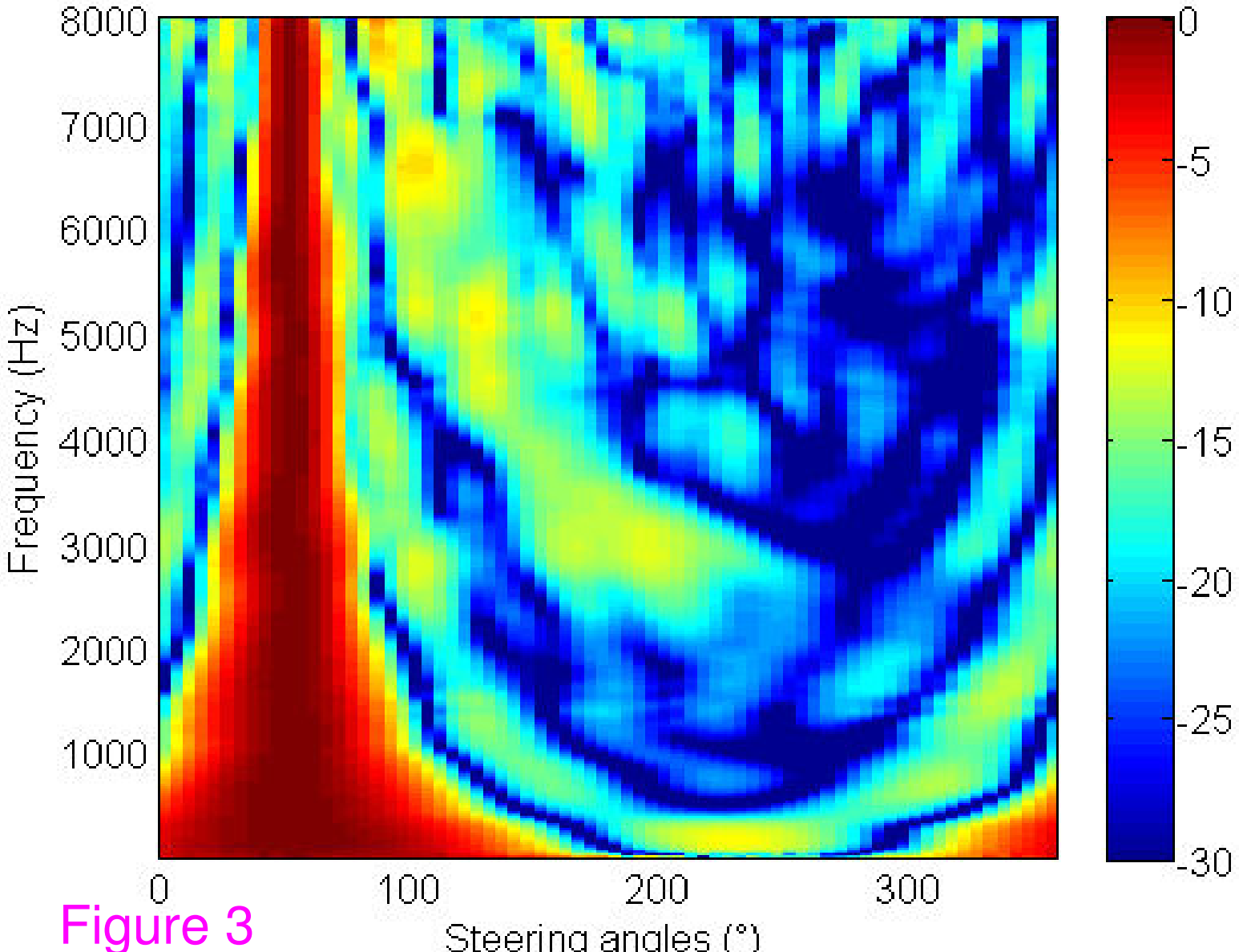

8000
7000
6000
5000
4000
3000
2000
1000
Frequency (Hz)
0
100
200
300
Steering angles (°)
0
-5
-10
-15
-20
-25
-30

Figure 3

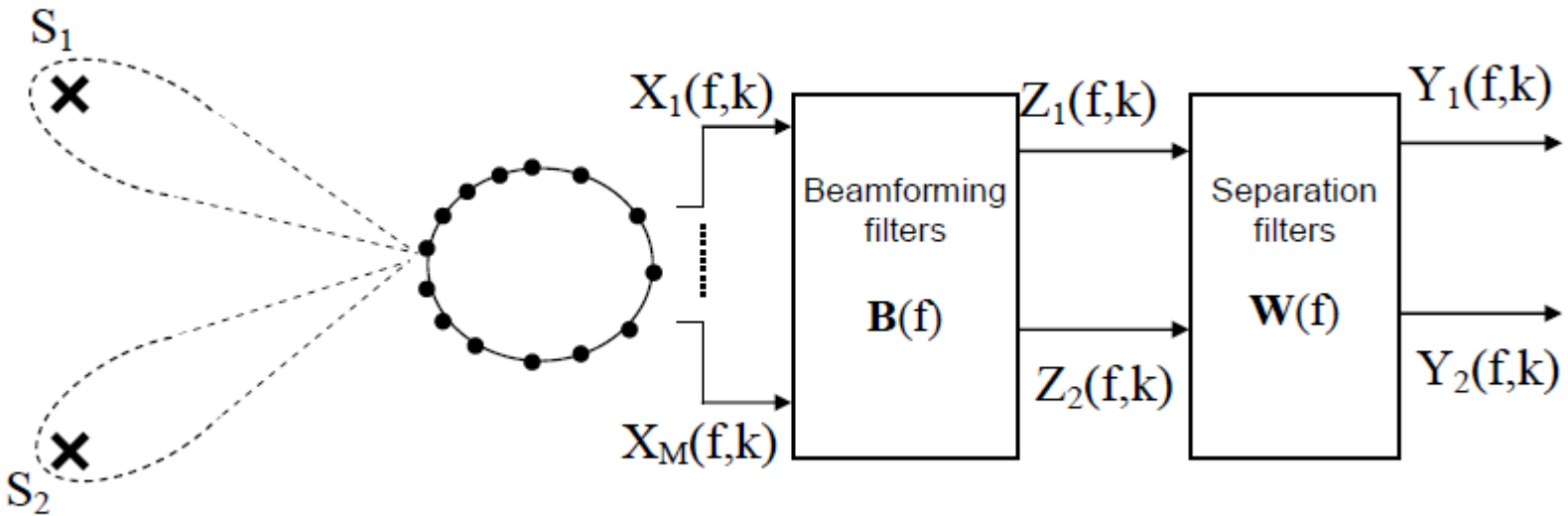

S1
S2
X1(f,k)
XM(f,k)
Beamforming filters
B(f)
Z1(f,k)
Z2(f,k)
Separation filters
W(f)
Y1(f,k)
Y2(f,k)

Figure 4

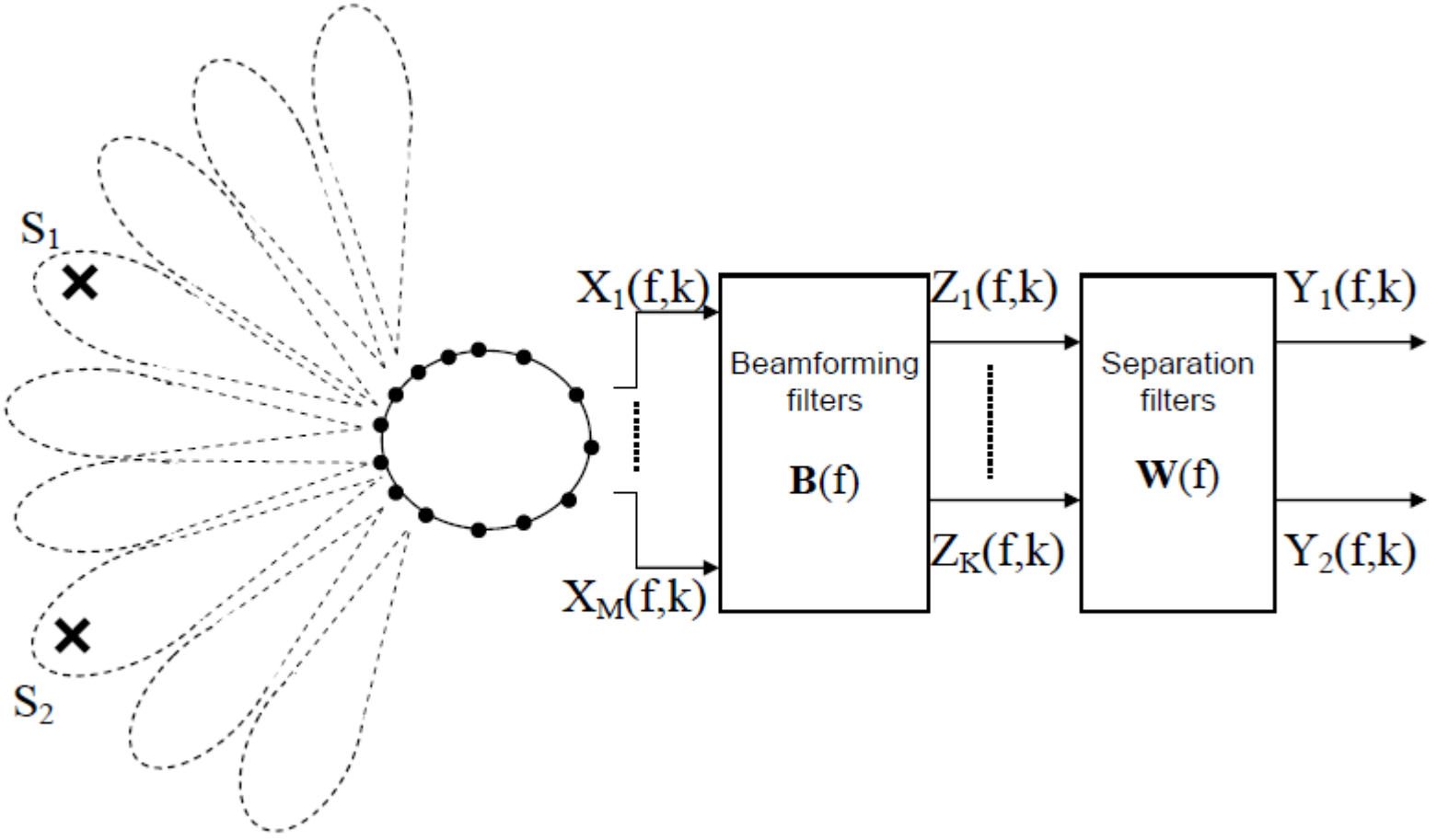

S1
S2
X1(f,k)
XM(f,k)
Beamforming filters
B(f)
Z1(f,k)
ZK(f,k)
Separation filters
W(f)
Y1(f,k)
Y2(f,k)

Figure 5

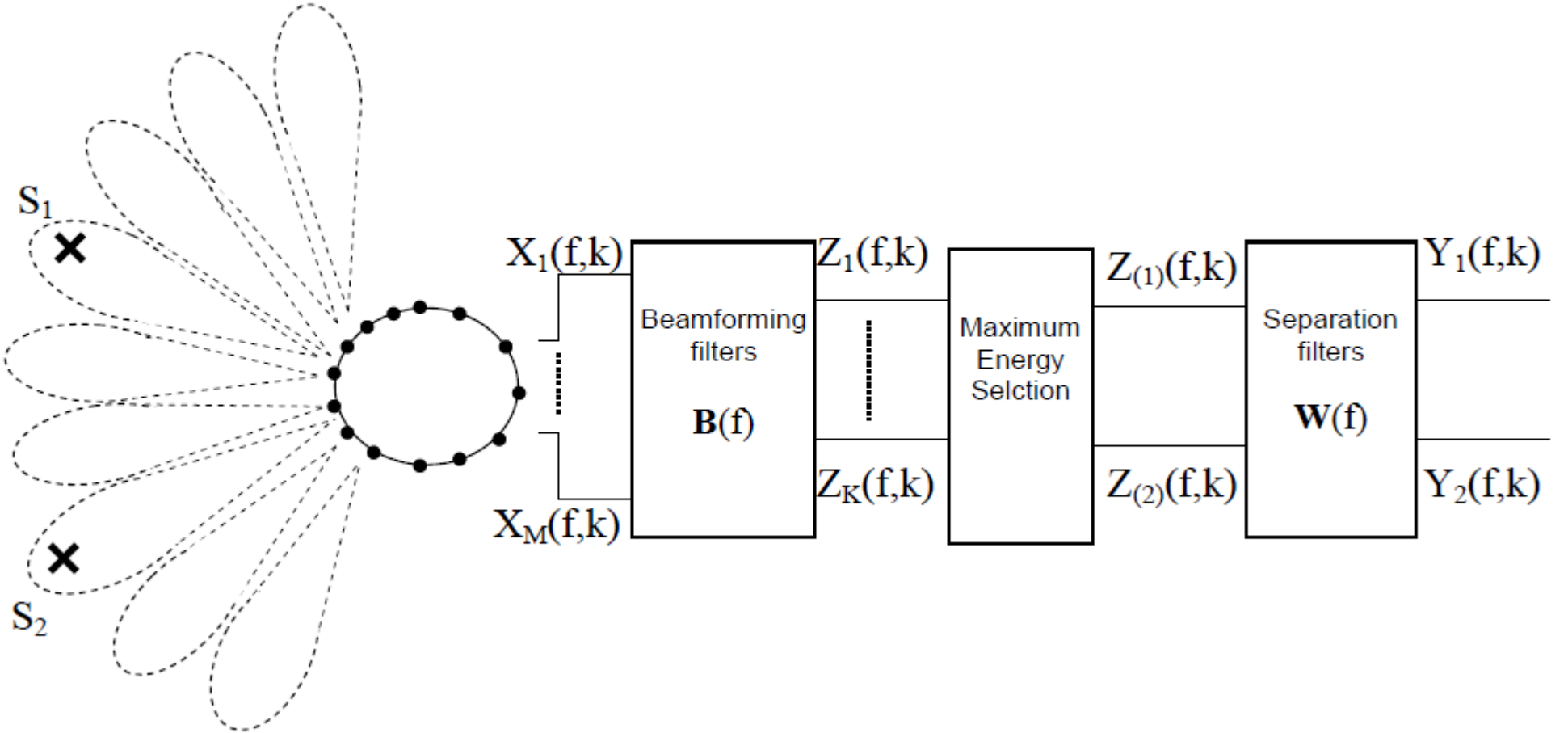

S1
S2
X1(f,k)
XM(f,k)
Beamforming filters
B(f)
Z1(f,k)
ZK(f,k)
Maximum Energy Selction
Z(1)(f,k)
Z(2)(f,k)
Separation filters
W(f)
Y1(f,k)
Y2(f,k)

Figure 6

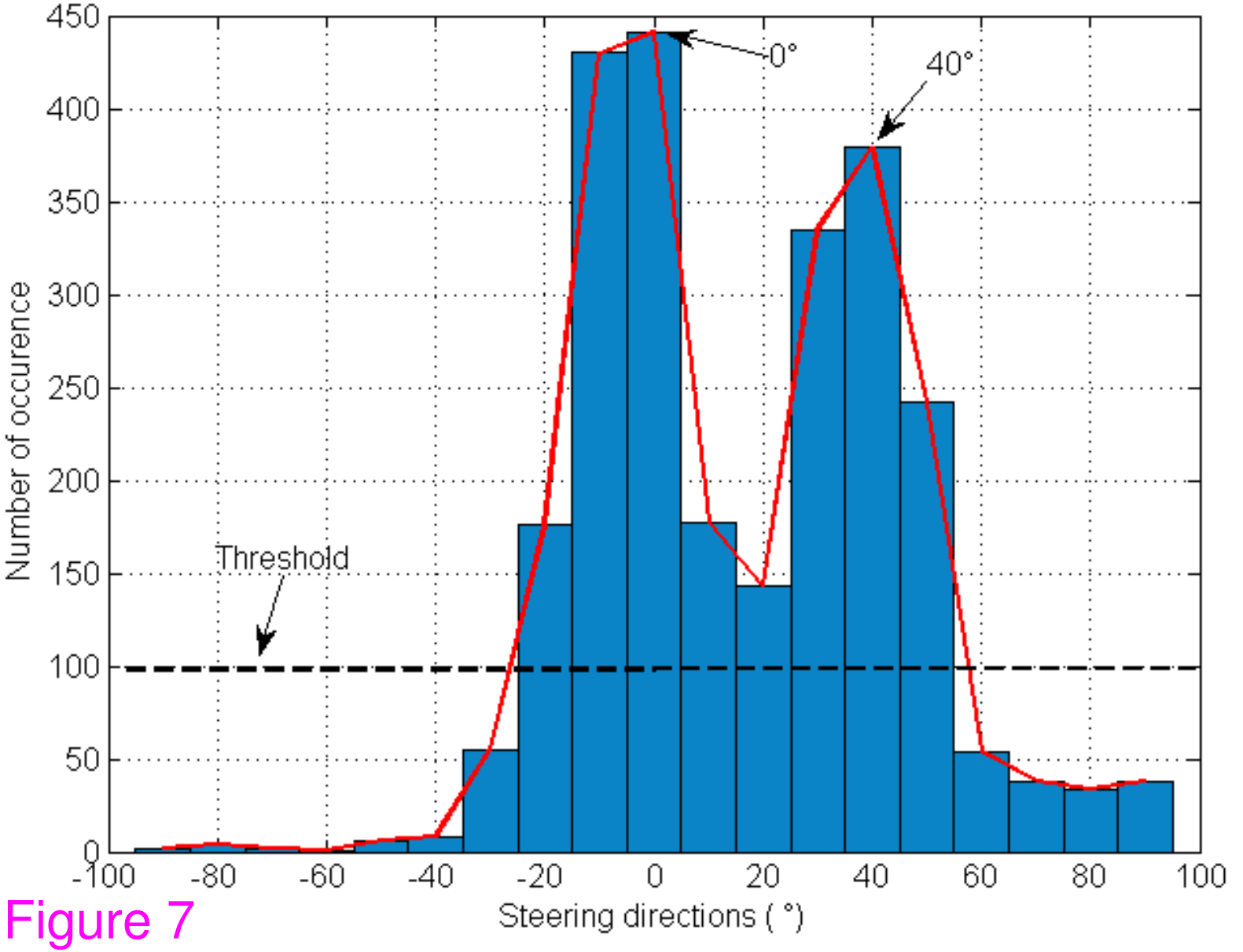
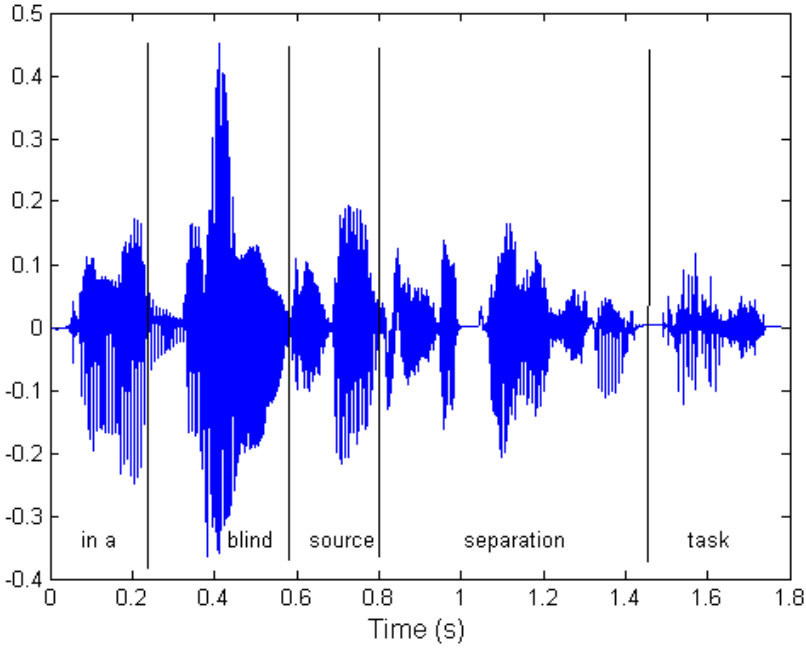
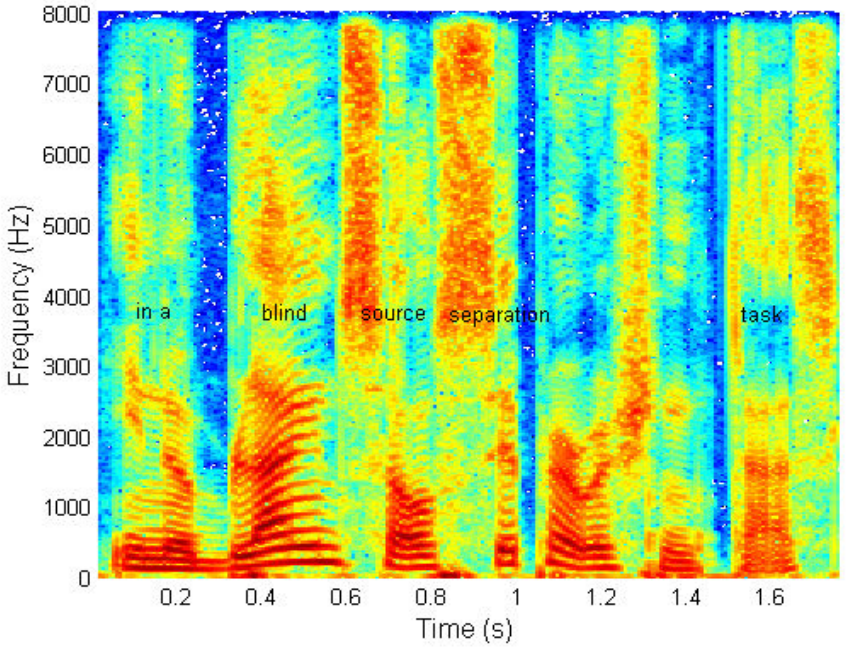

450
400
350
300
250
200
150
100
50
0
Number of occurence
-100
-80
-60
-40
-20
0
20
40
60
80
100
Steering directions (°)
0°
40°
Threshold

Figure 7


0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Time (s)
in a
blind
source
separation
task

(a)


8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency (Hz)
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Time (s)
in a
blind
source
separation
task

(b)

Figure 8

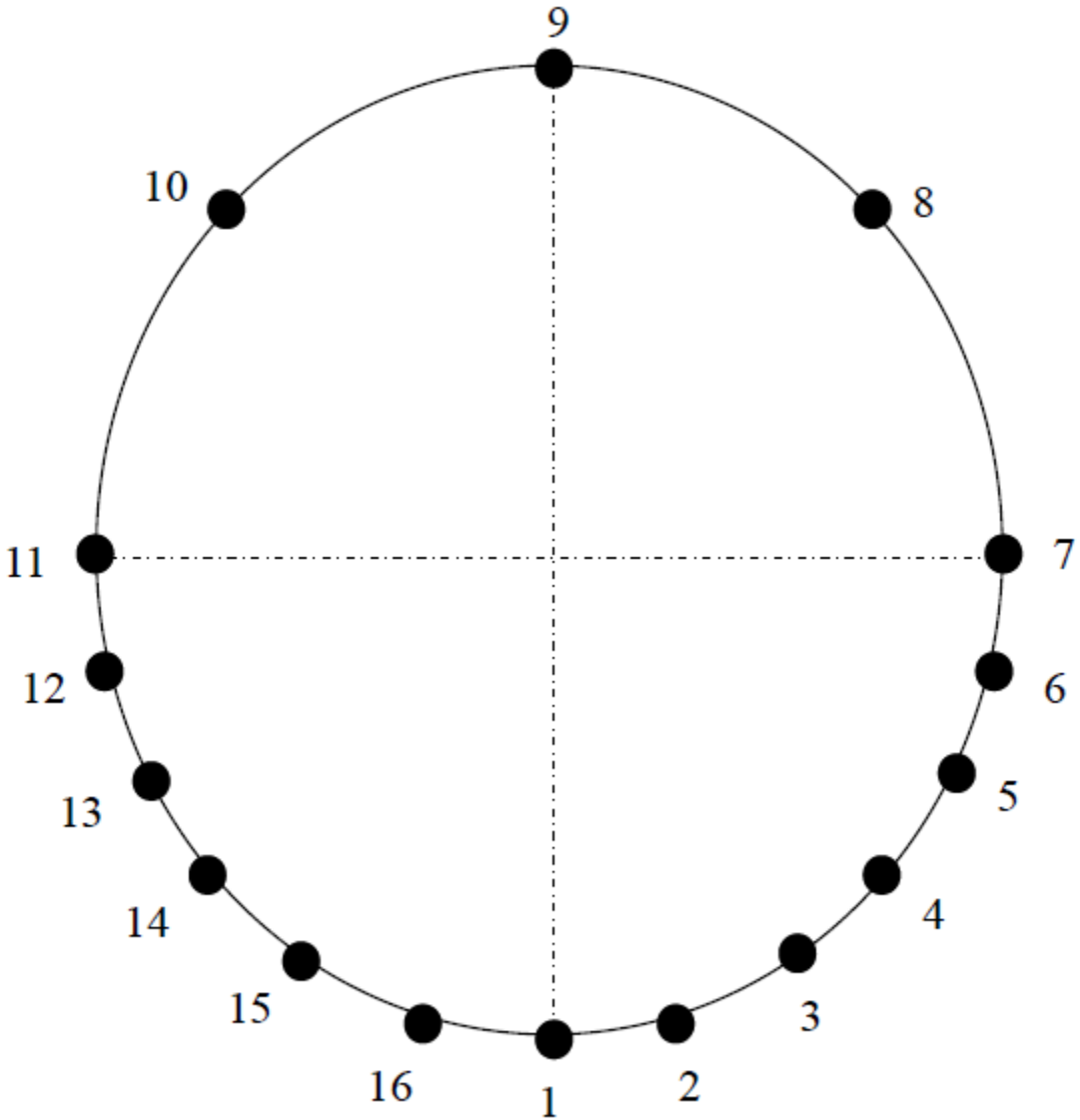


Figure 9

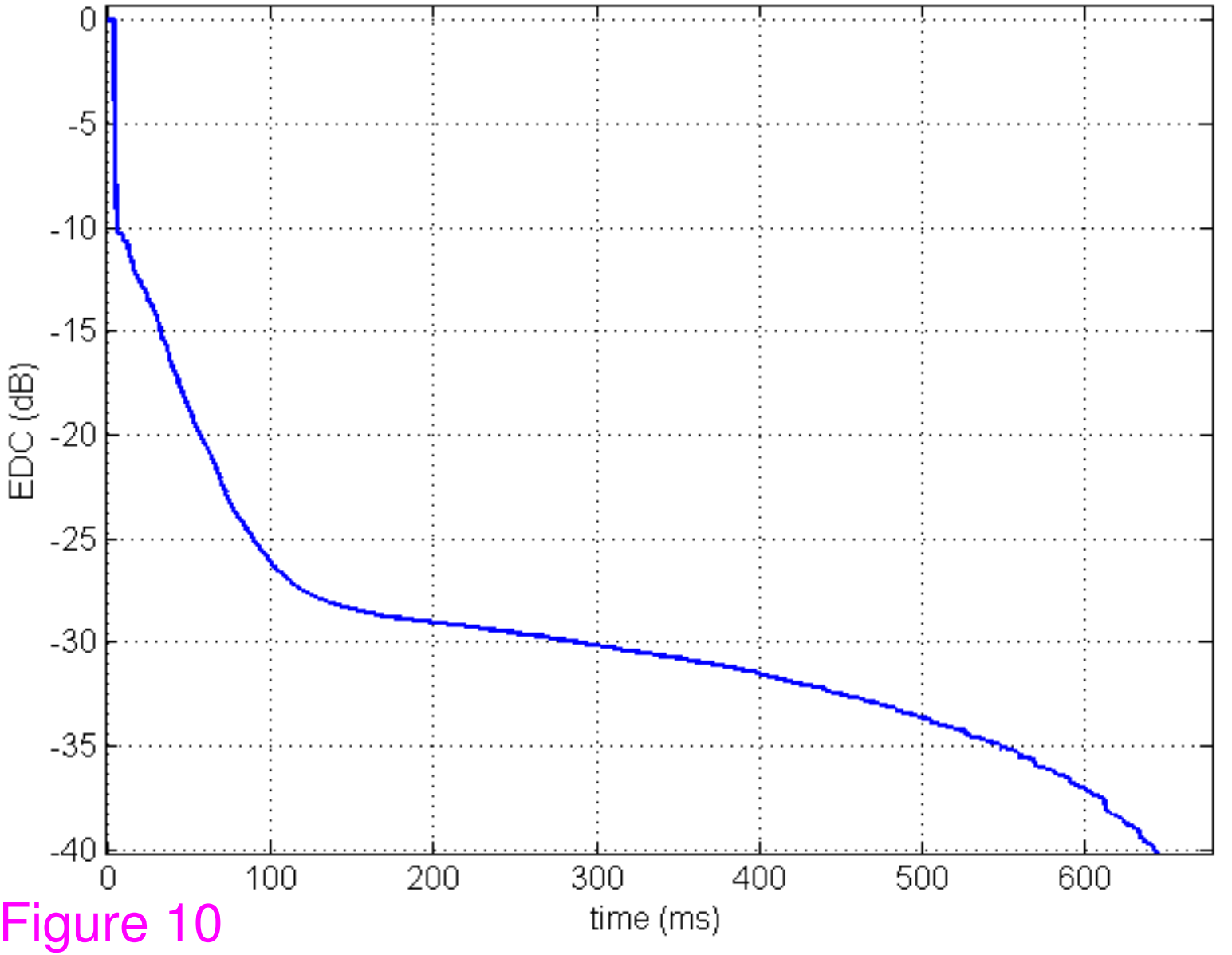


Figure 10

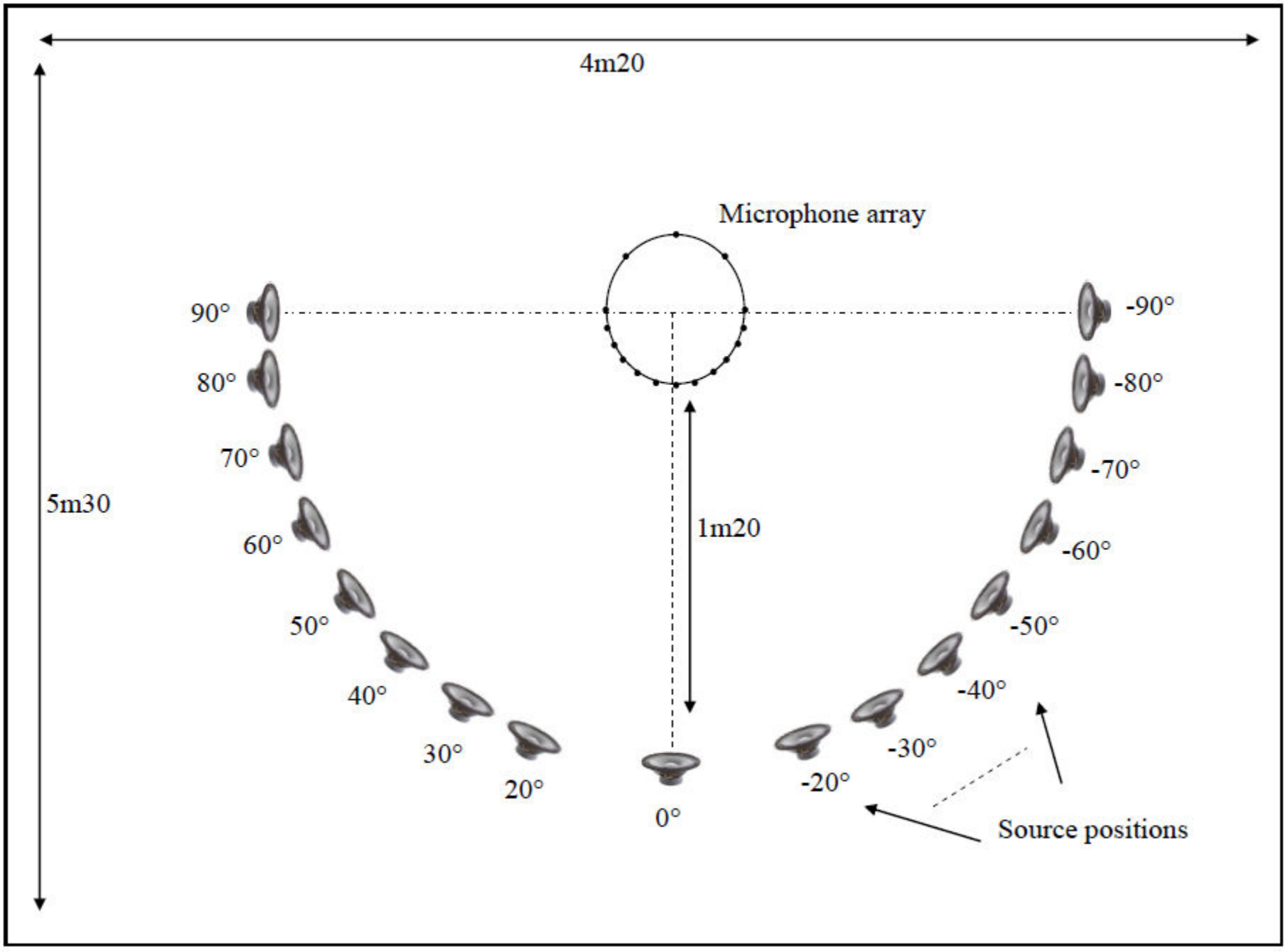

4m20
5m30
Microphone array
1m20
90°
80°
70°
60°
50°
40°
30°
20°
0°
-20°
-30°
-40°
-50°
-60°
-70°
-80°
-90°
Source positions

Figure 11

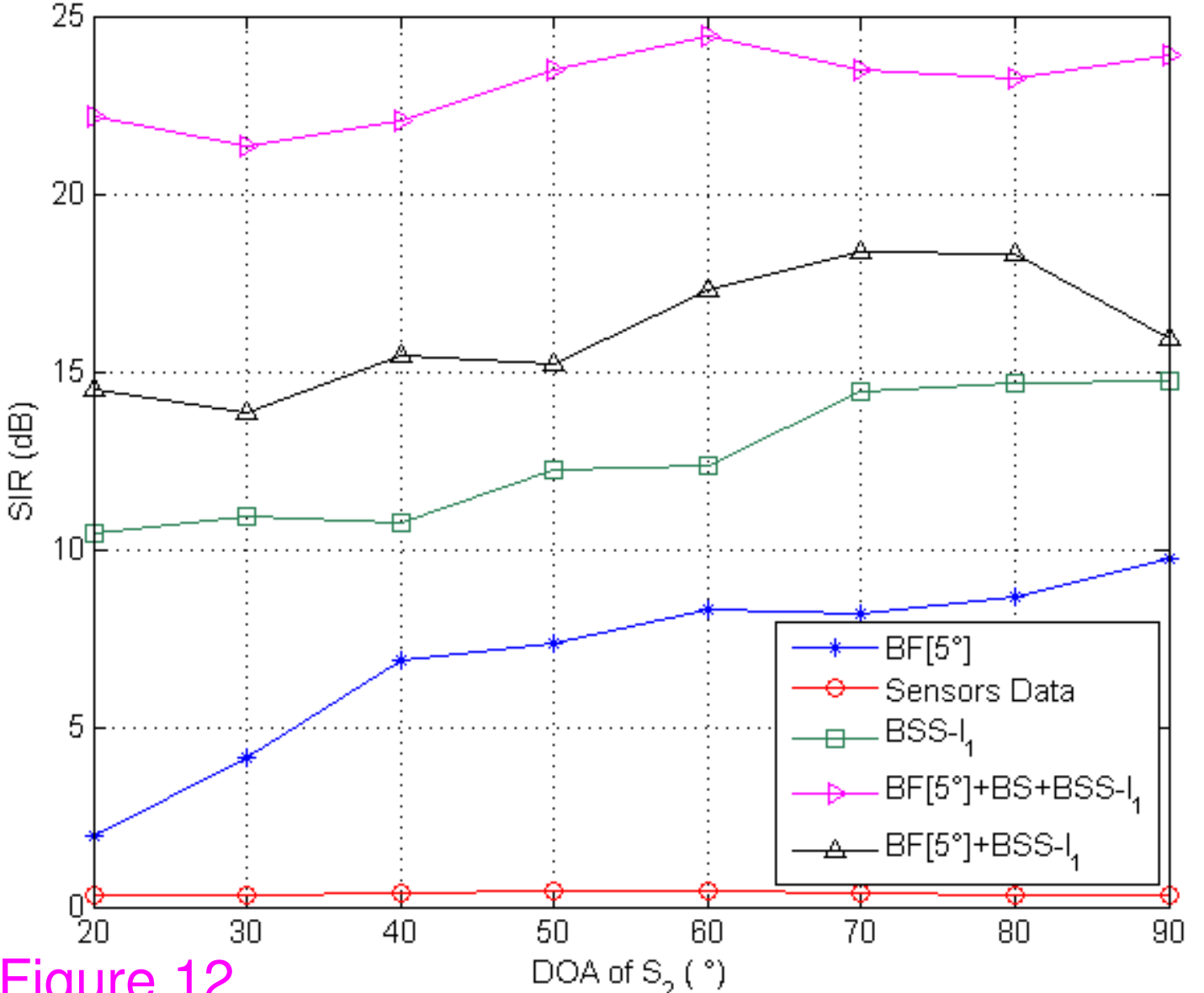

25
20
15
10
5
0
SIR (dB)
20
30
40
50
60
70
80
90
DOA of S2 (°)
BF[5°]
Sensors Data
BSS-I1
BF[5°]+BS+BSS-I1
BF[5°]+BSS-I1

Figure 12

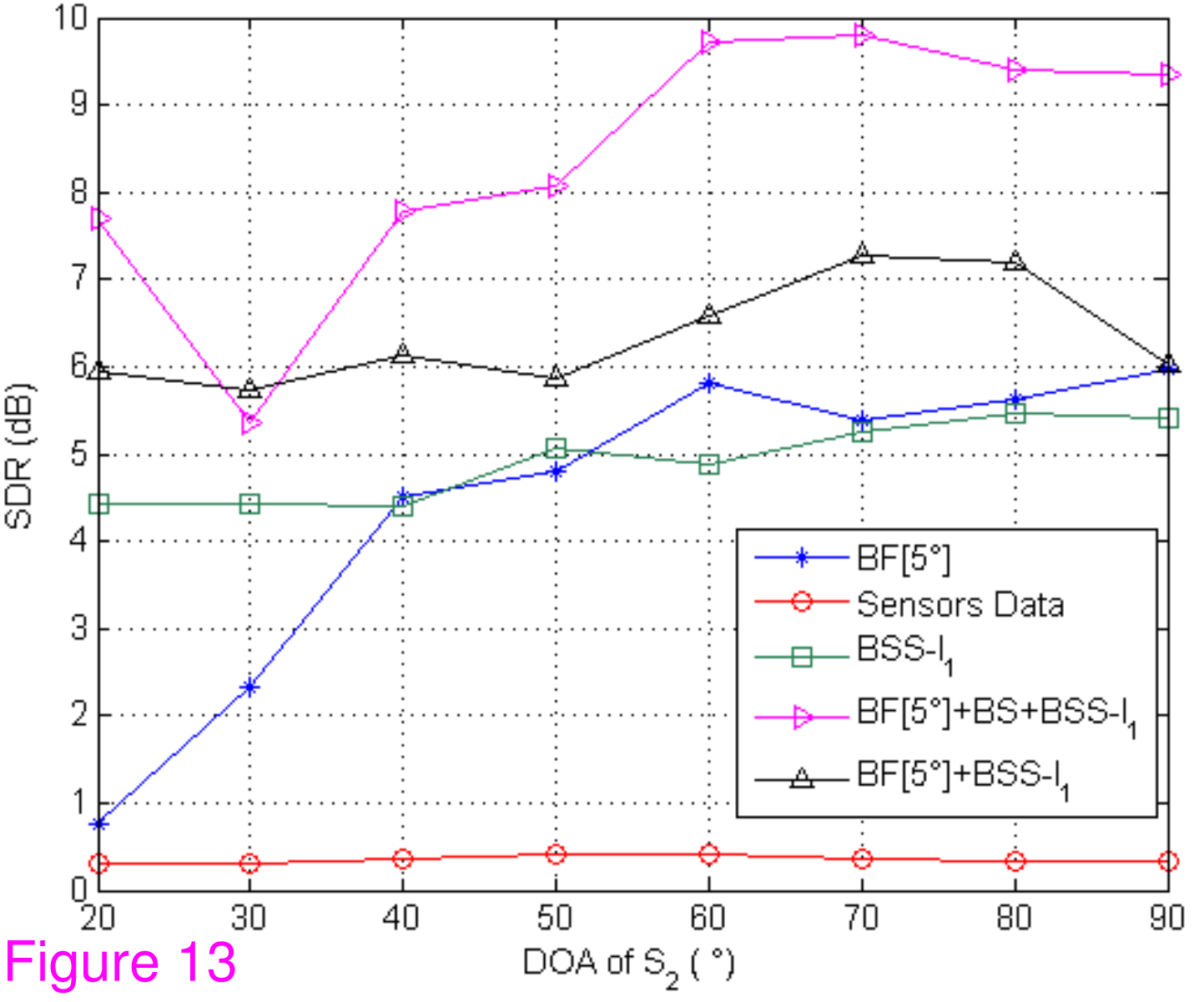


Figure 13

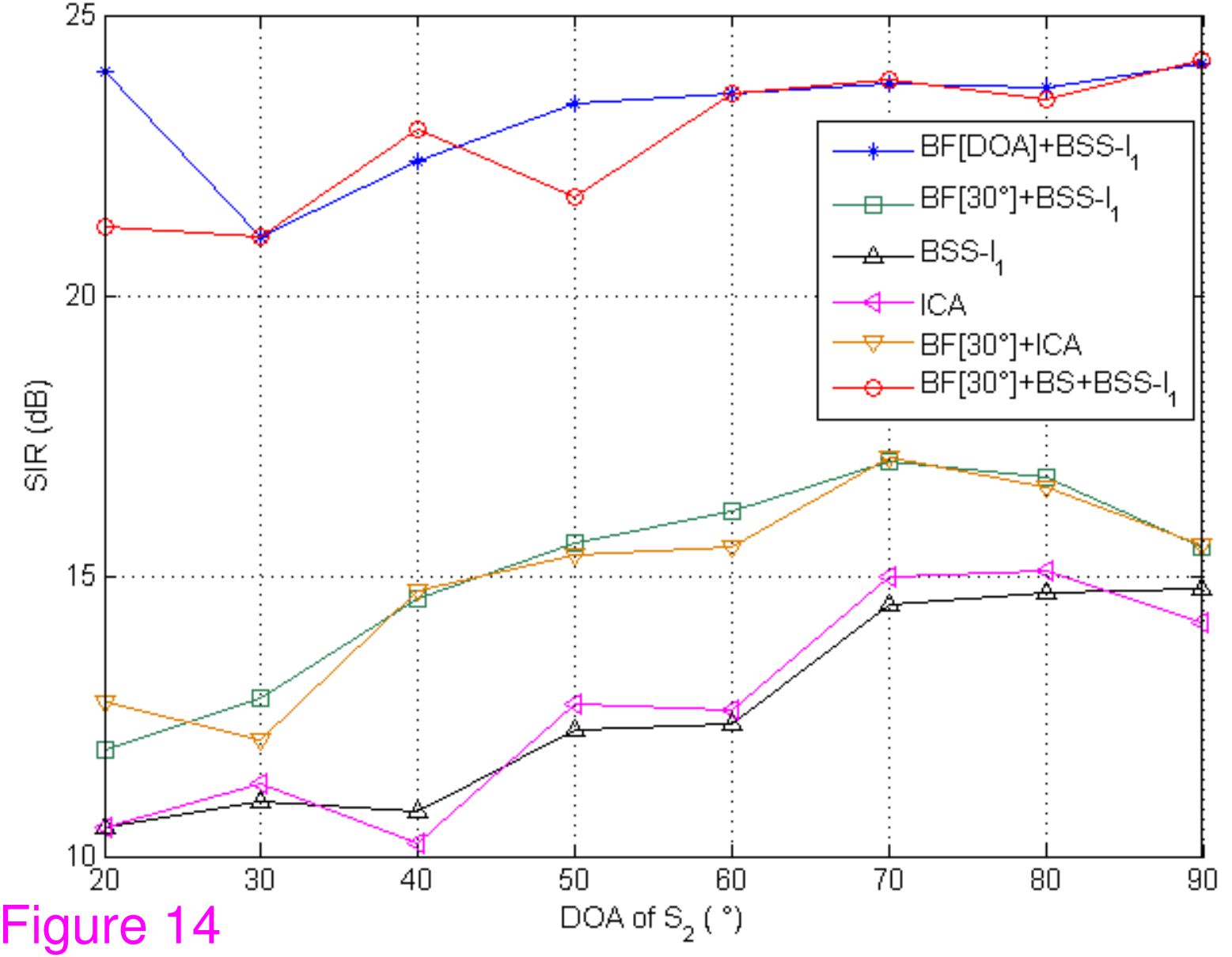

25
20
15
10
SIR (dB)
20
30
40
50
60
70
80
90
DOA of S2 (°)
BF[DOA]+BSS-I1
BF[30°]+BSS-I1
BSS-I1
ICA
BF[30°]+ICA
BF[30°]+BS+BSS-I1

Figure 14

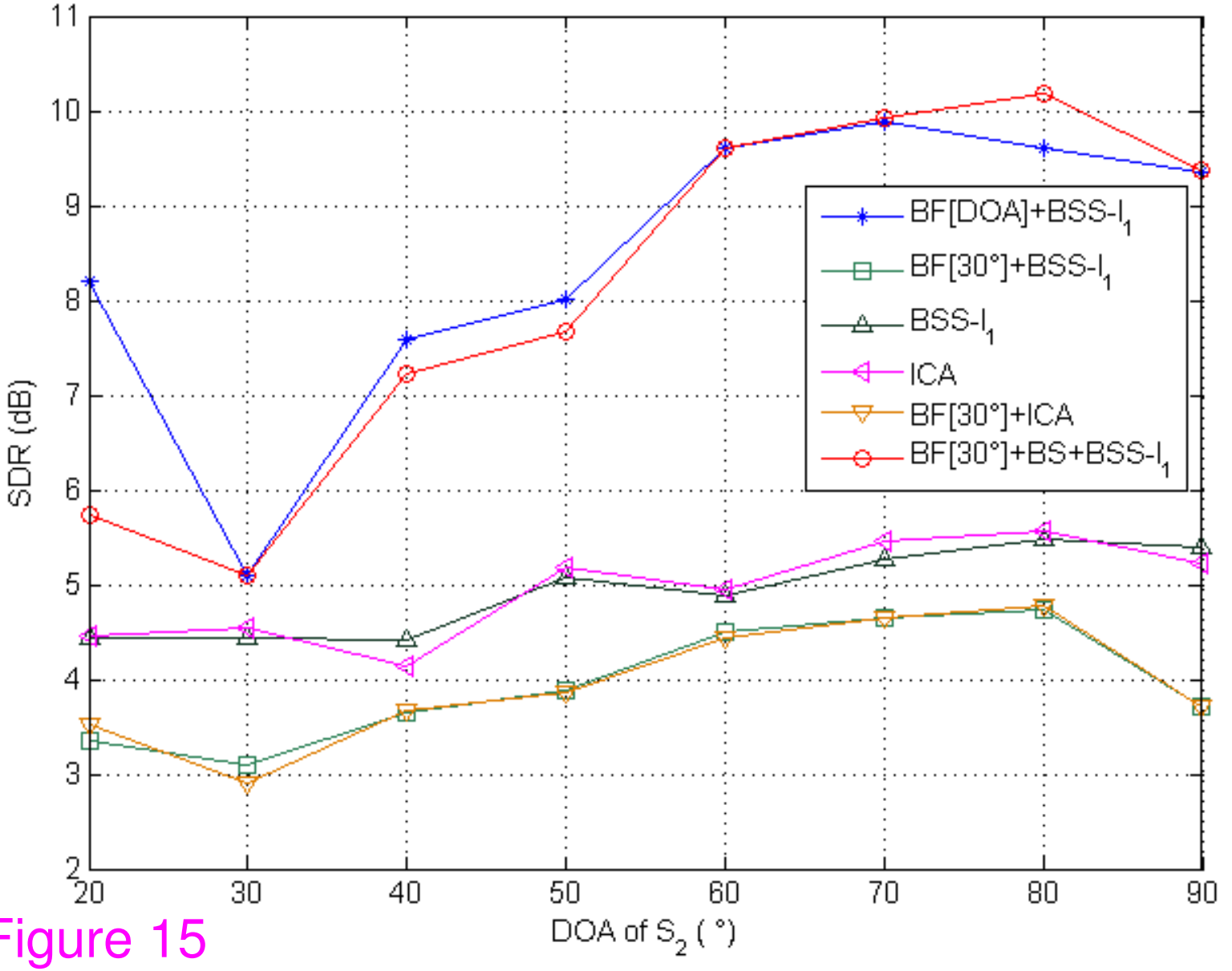

SDR (dB)
DOA of S2 (°)
BF[DOA]+BSS-I1
BF[30°]+BSS-I1
BSS-I1
ICA
BF[30°]+ICA
BF[30°]+BS+BSS-I1

Figure 15

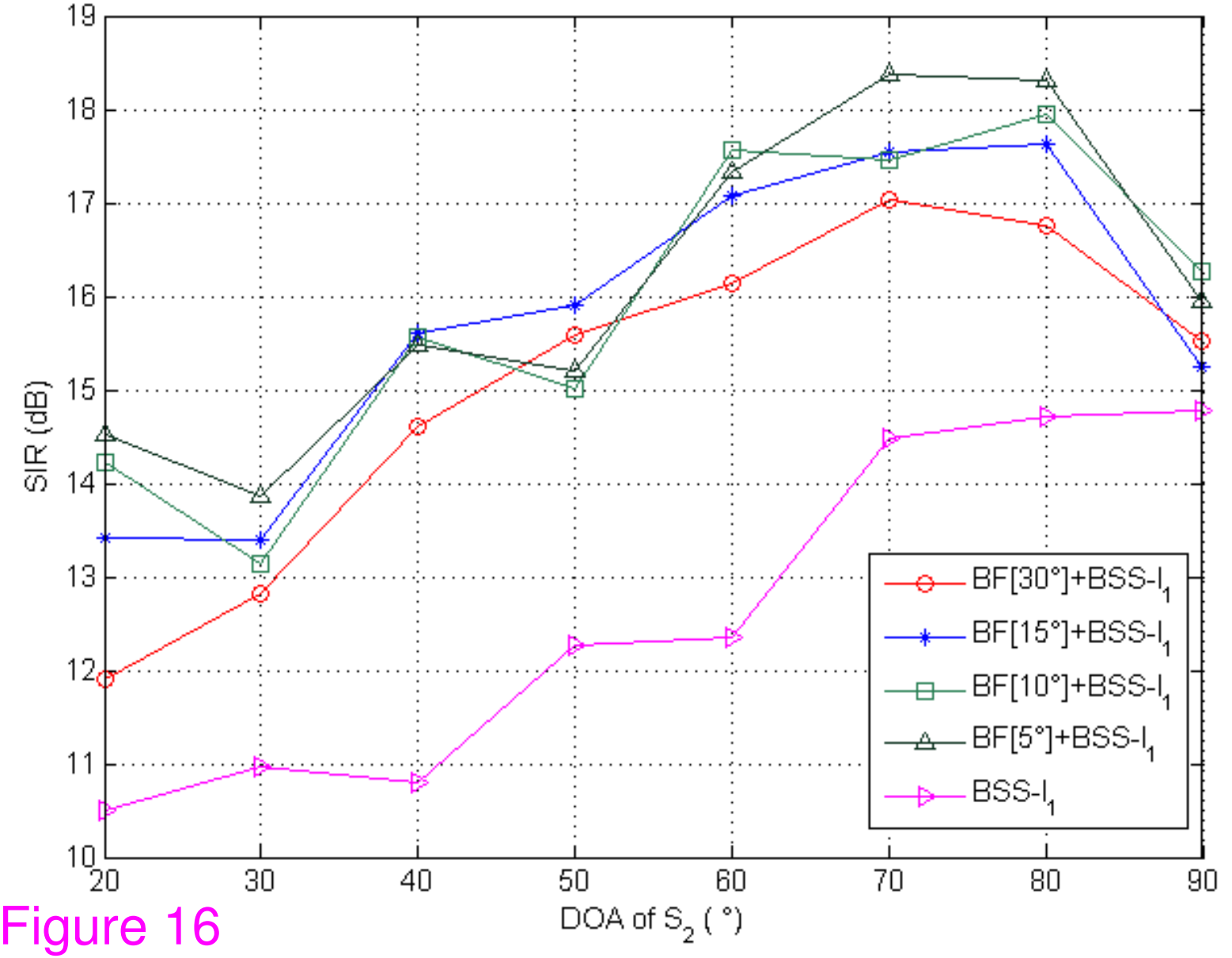

SIR (dB)
DOA of S_2 (°)
BF[30°]+BSS-I_1
BF[15°]+BSS-I_1
BF[10°]+BSS-I_1
BF[5°]+BSS-I_1
BSS-I_1

Figure 16

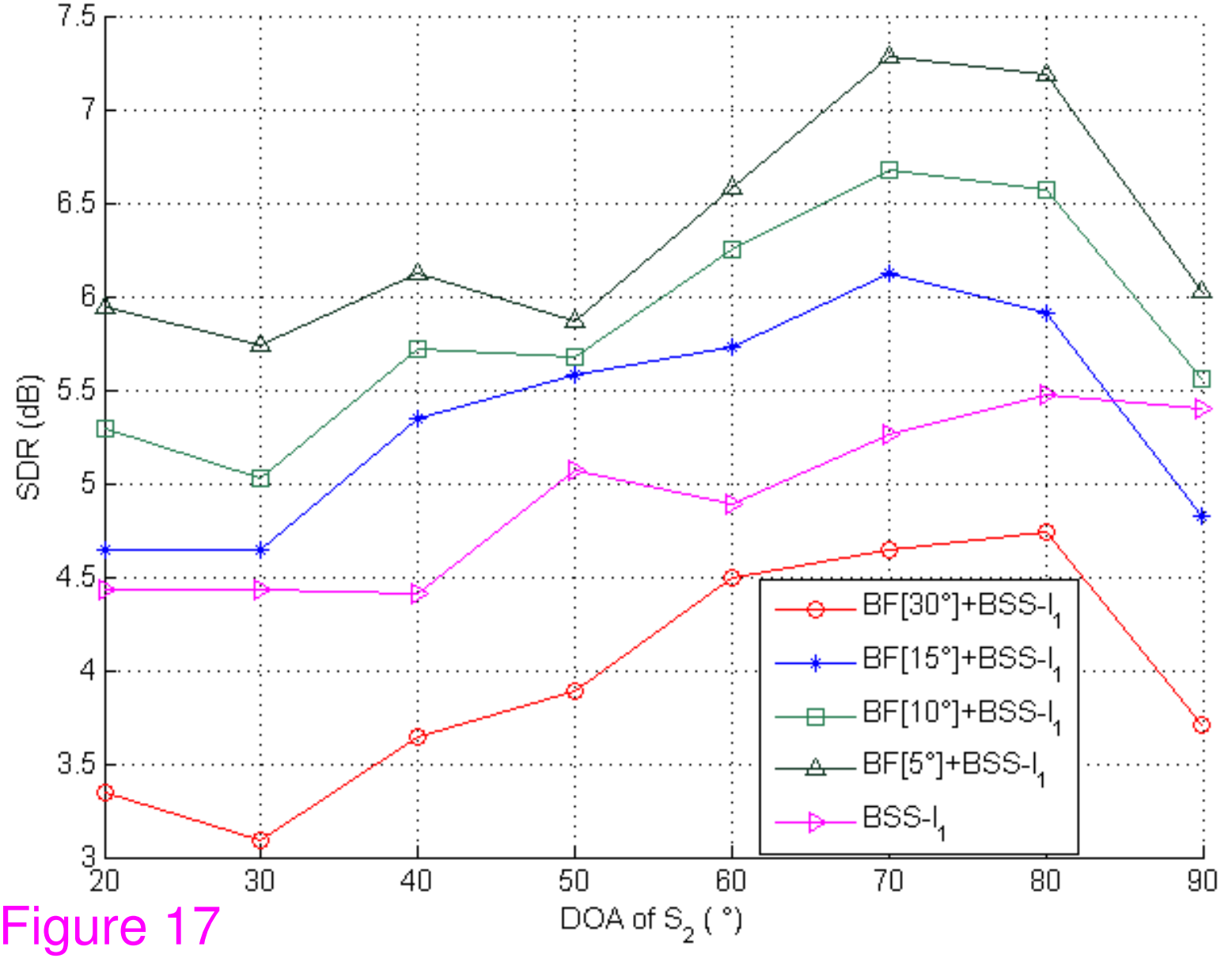
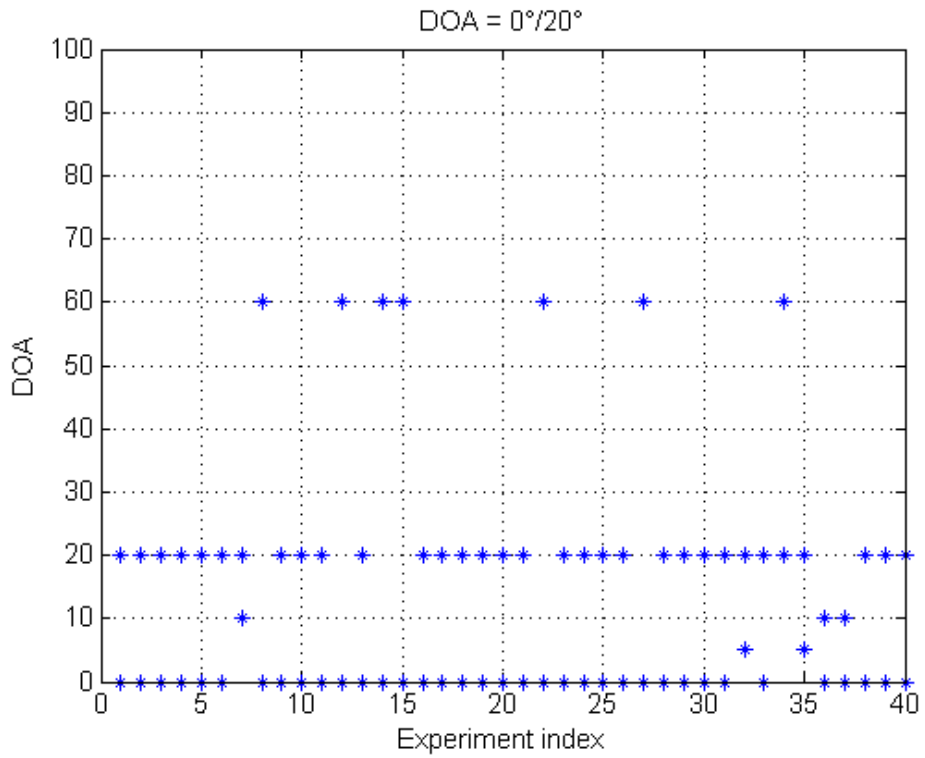
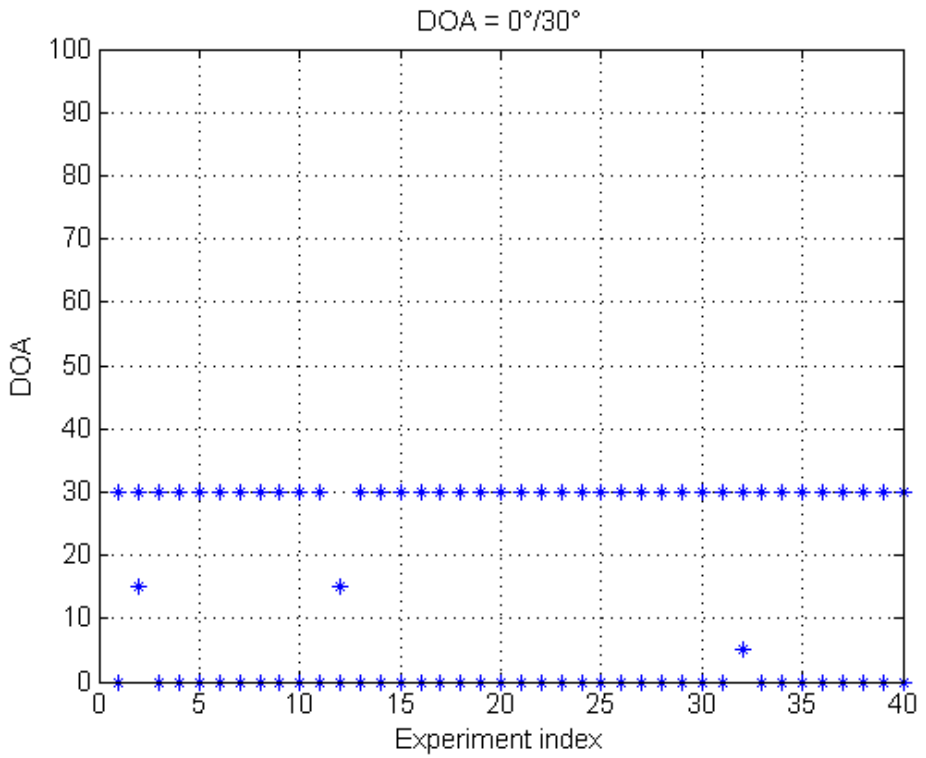
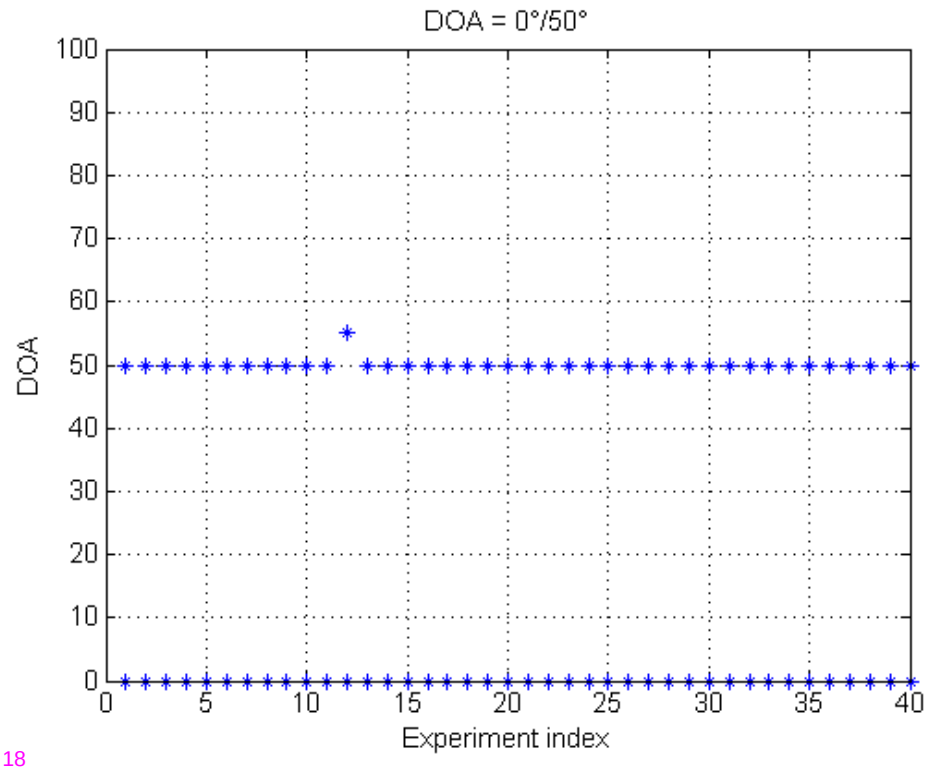

7.5
7
6.5
6
5.5
5
4.5
4
3.5
3
SDR (dB)
20
30
40
50
60
70
80
90
DOA of S2 (°)
BF[30°]+BSS-I1
BF[15°]+BSS-I1
BF[10°]+BSS-I1
BF[5°]+BSS-I1
BSS-I1

Figure 17


DOA = 0°/20°
DOA
Experiment index


DOA = 0°/30°
DOA
Experiment index


DOA = 0°/50°
DOA
Experiment index

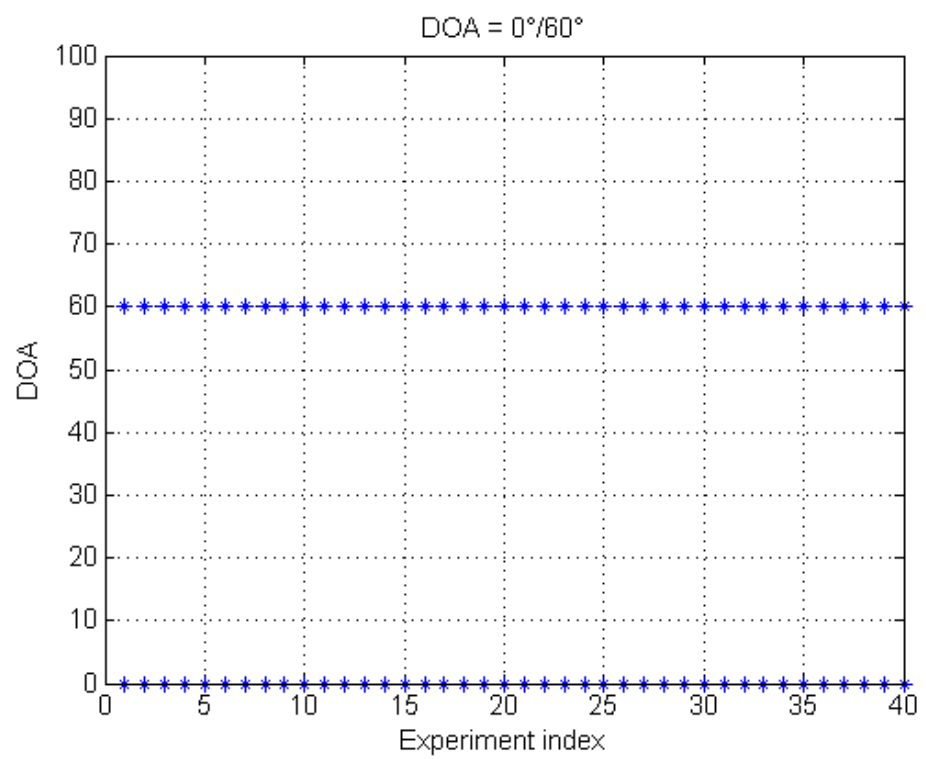
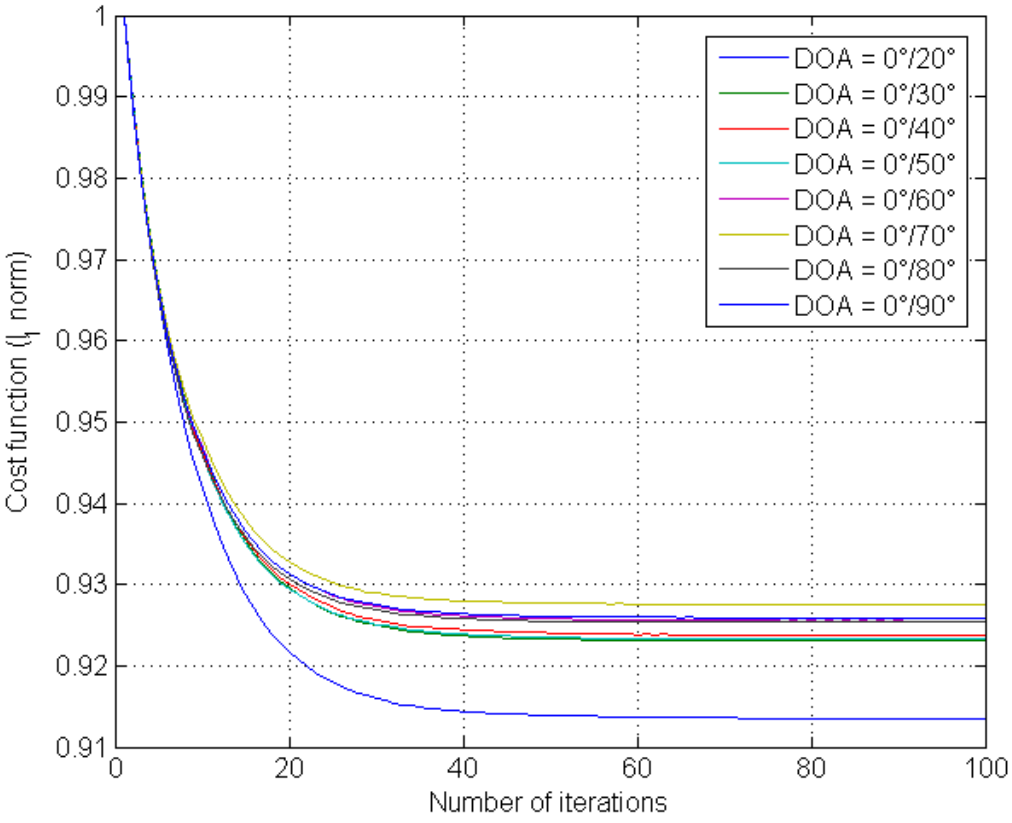
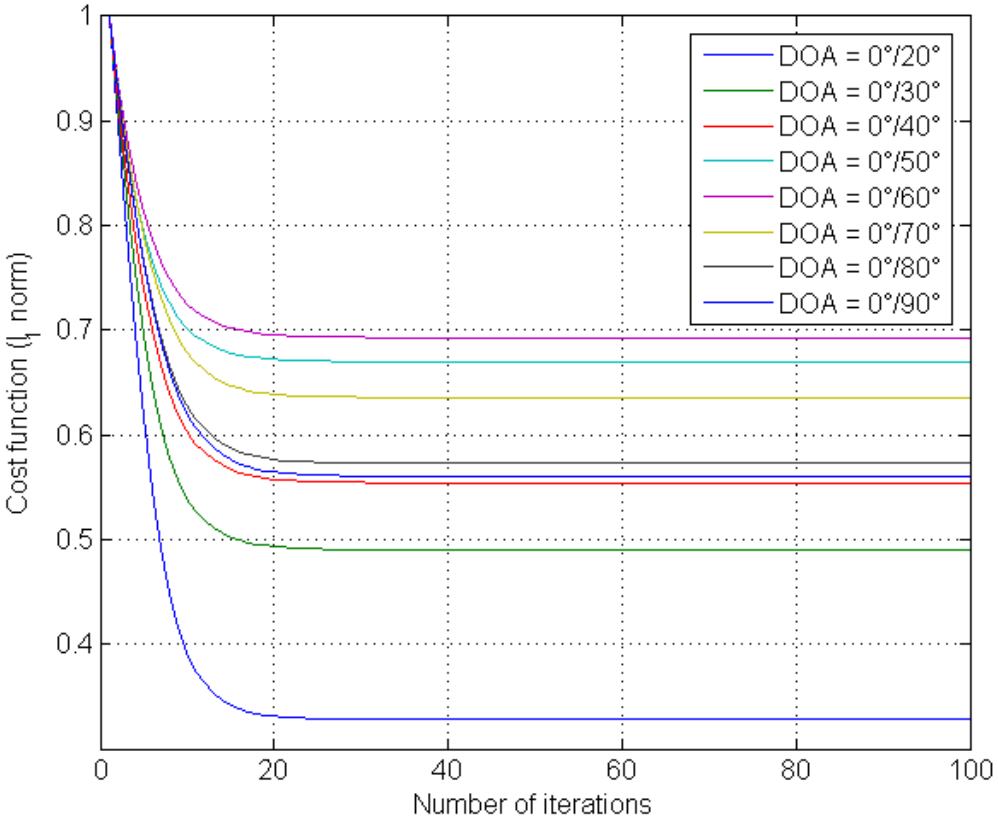

DOA = 0°/60°
DOA
Experiment index

Figure 18


DOA = 0°/20°
DOA = 0°/30°
DOA = 0°/40°
DOA = 0°/50°
DOA = 0°/60°
DOA = 0°/70°
DOA = 0°/80°
DOA = 0°/90°
Cost function (l_1 norm)
Number of iterations
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0
20
40
60
80
100

(a)


DOA = 0°/20°
DOA = 0°/30°
DOA = 0°/40°
DOA = 0°/50°
DOA = 0°/60°
DOA = 0°/70°
DOA = 0°/80°
DOA = 0°/90°
Cost function (l_1 norm)
Number of iterations
1
0.9
0.8
0.7
0.6
0.5
0.4
0
20
40
60
80
100

(b)

Figure 19