Robust microphone array signal processing against diffuse noise

Nobutaka Ito

▶ **To cite this version:**

Nobutaka Ito. Robust microphone array signal processing against diffuse noise. Signal and Image processing. University of Tokyo, 2012. English. NNT : . tel-00691931

HAL Id: tel-00691931
https://theses.hal.science/tel-00691931

Submitted on 27 Apr 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Robust Microphone Array Signal Processing against Diffuse Noise

(拡散性雑音に頑健な
マイクロフォンアレイ信号処理に関する研究)

Nobutaka Ito
伊 藤　信 貴

Contents

Abstract

We consider the general problem of microphone array signal processing in diffuse noise environments. This has various applications epitomized by speech enhancement and robust Automatic Speech Recognition (ASR) for microphone arrays. Diffuse noise arriving from almost all directions is often encountered in the real world, and has been one of the major obstacles against successful application of existing noise suppression and Direction-Of-Arrival (DOA) estimation techniques. We operate in the time-frequency domain, where signal and noise are assumed to be zero-mean Gaussian and modeled by their respective covariance matrices.

Firstly, we introduce a general linear subspace model of the noise covariance matrix that extends three state-of-the-art models, and introduce a fourth more flexible real-valued noise covariance model. We experimentally assess the fit of each model to real-world noise.

Secondly, we apply this general model to the task of diffuse noise suppression with a known target steering vector. In the state-of-the-art Wiener post-filtering approach, it is essential to accurately estimate the target power spectrogram. We propose a unified estimation framework applicable to the general noise model, which is based on projecting the observed covariance matrix onto the orthogonal complement of the noise model subspace. Ideally, this projection is noise-free, and enables accurate estimation of the target power spectrogram. The proposed framework for noise suppression is assessed through experiments with real-world noise.

Thirdly, we address the task of DOA estimation of multiple sources. The performance of the state-of-the-art MUltiple SIgnal Classification (MUSIC) algorithm is known to degrade in the presence of diffuse noise. In order to mitigate this effect, we estimate the signal covariance matrix and subsequently apply MUSIC to it. The estimation relies on the above-mentioned noise-free component of the observed covariance matrix and on the reconstruction of the remaining component belonging to the noise subspace. We design two alternative algorithms based on low-rank matrix completion and trace-norm minimization that exploit the low-rankness and the positive semidefiniteness of the signal covariance matrix. The

performance of the proposed method with each noise model was compared using a large database we created.

Finally, we present a unified framework applicable to the general noise model for diffuse noise suppression with an unknown target steering vector. This is important for effective noise suppression in the real-world, because the steering vector is usually not accurately known in practice. We jointly estimate the target steering vector and the target power spectrogram for designing the beamformer and the Wiener post-filter. The estimation is based on rank-1 completion and Principal Component Analysis (PCA). The proposed framework is shown to enable more effective noise suppression improving the SNR by about 7dB, compared to the state-of-the-art Independent Vector Analysis (IVA).

Notations

Scalars, vectors, and matrices

Scalars are denoted by regular lowercase letters (*e.g.* a), vectors by bold lowercase letters (*e.g.* $\boldsymbol{a}$), and matrices by bold uppercase letters (*e.g.* $\boldsymbol{A}$).

Operators

$(\cdot)^{\mathsf{T}}$	Transposition of a vector or a matrix
$(\cdot)^{\mathsf{H}}$	Conjugate transposition of a vector or a matrix
$(\cdot)^*$	Entry-wise conjugate of a scalar, a vector, or a matrix.
$\Re[\cdot]$	Entry-wise real part of a scalar, a vector, or a matrix
$\Im[\cdot]$	Entry-wise imaginary part of a scalar, a vector, or a matrix
$\lvert \cdot \rvert$	Absolute value of a complex number
$\lVert \cdot \rVert_{\mathrm{F}}$	Frobenius norm of a matrix
$\lVert \cdot \rVert_*$	Trace norm of a matrix
$\lVert \cdot \rVert_2$	L_2-norm of a vector
$\mathcal{E}[\cdot]$	Expectation of a random variable, vector, or matrix
$\mathcal{D}[\cdot]$	Operation of replacing the off-diagonal entries of a matrix by zeros
$\mathcal{O}[\cdot]$	Operation of replacing the diagonal entries of a matrix by zeros

Indices

l	Source index (from 1 to L)
m, n	Microphone index (from 1 to M)
t	Time
τ	Frame index
ω	Angular frequency

Constants

e	Euler's constant
j	Imaginary unit
ζ_p	p-th imaginary root of -1 (p: integer greater than 2)

Sets

$\mathbb{Z}$	Set of the integers
$\mathbb{R}$	Set of the real numbers
$\mathbb{R}^p$	Set of the p-dimensional real-valued column vectors (p: positive integer)
$\mathbb{R}^{p\times q}$	Set of the $p \times q$ real-valued matrices (p, q: positive integers)
$\mathbb{C}$	Set of the complex numbers
$\mathbb{C}^p$	Set of the p-dimensional complex-valued column vectors (p: positive integer)
$\mathbb{C}^{p\times q}$	Set of the $p \times q$ complex-valued matrices (p, q: positive integers)

Functions

$\delta(\cdot)$	Dirac's delta function
$\mathrm{sinc}(\cdot)$	Sine cardinal function $\mathrm{sinc}(x) \triangleq \frac{\sin x}{x}$
$J_0(\cdot)$	Zeroth-order Bessel function of the first kind
$\mathrm{circ}(a_1, a_2, \ldots, a_p)$	$p \times p$ circulant matrix whose first row is $\begin{bmatrix} a_1 & a_2 & \cdots & a_p \end{bmatrix}$ (p: positive integer)

Other mathematical notations

$\phi_{\alpha\alpha}(\tau, \omega)$	Power spectrogram of zero-mean scalar signal $\alpha(\tau, \omega)$, $\phi_{\alpha\alpha}(\tau, \omega) \triangleq \mathcal{E}[\lvert\alpha(\tau, \omega)\rvert^2]$
$\phi_{\alpha\beta}(\tau, \omega)$	Cross-spectrogram of zero-mean scalar signals $\alpha(\tau, \omega)$ and $\beta(\tau, \omega)$, $\phi_{\alpha\beta}(\tau, \omega) \triangleq \mathcal{E}[\alpha(\tau, \omega)\beta^*(\tau, \omega)]$
$\boldsymbol{\Phi}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\tau, \omega)$	Covariance matrix of zero-mean vector signal $\boldsymbol{\alpha}(\tau, \omega)$, $\boldsymbol{\Phi}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\tau, \omega) \triangleq \mathcal{E}[\boldsymbol{\alpha}(\tau, \omega)\boldsymbol{\alpha}^{\mathsf{H}}(\tau, \omega)]$

Abbreviations

ASR	Automatic Speech Recognition
DFT	Discrete Fourier Transform
DOA	Direction Of Arrival
DS	Delay And Sum
EVD	EigenValue Decomposition
IVA	Independent Vector Analysis
LMMSE	Linear Minimum Mean Square Error
LS	Least Squares
MUSIC	MUltiple SIgnal Classification
MVDR	Minimum Variance Distortionless Response
NRF	Noise-Reduction Factor
PCA	Principal Component Analysis
RMSE	Root Mean Square Error
SDI	Speech-Distortion Index
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
TDOA	Time Delay Of Arrival

Chapter 1

Introduction

1.1 Motivation for noise suppression and direction-of-arrival estimation

Humans have the excellent ability to recognize speech despite the simultaneous presence of many sounds in our surroundings. Let us imagine an office, for example. Employees are making sounds, while talking, walking, writing, turning pages, typing on keyboards, speaking on the phone, and so on. Also making sounds are PC fans, air conditioners, ringing telephones, doors, etc. Birds may be tweeting outside. Amazingly enough, you can recognize what a colleague says even in such an adverse sound environment.

How would it help us, if we could build systems that share this human ability? You could turn immediately the air conditioner on by using a voice command like "Air conditioner, on", even during a bustling home party. Discussions at meetings could be automatically transcribed into written form, even though more than two persons may sometimes speak simultaneously and some background noise may be present. Hearing-impaired persons could know what a friend is saying even in a noisy street by using a smartphone application that transcribes speech into text.

What is the difficulty in building such useful systems? ASR systems provide very accurate speech recognition when the microphone is placed near the speaker's mouth (close-talking scenario) [2]. In contrast, the recognition performance degrades dramatically when the microphone is placed far away (distant-talking scenario) [2]. Unfortunately, it is necessary to stick to the distant-talking scenario in many applications. For instance, it would be infeasible to always put the microphone close to the speaker's mouth in the above example

involving the hearing impaired.

The degradation of the ASR performance in this context can be attributed to the following two adverse effects of the increasing distance:

- an increased amount of reverberation relative to the direct sound,
- an increased amount of sounds from other sources (referred to as *noise*) relative to the *target* speech.

Dereverberation methods for mitigating the effect of reverberation have been widely studied in the literature (*e.g.* [3, 4]). In this thesis, we focus on the latter problem, namely noise.

A promising approach to making ASR systems robust against noise is to perform *noise suppression* prior to ASR. Noise suppression refers to the range of techniques aiming to recover the target speech from its noisy observation. The reduced amount of noise due to noise suppression is expected to result in a better recognition performance [5]. Specifically, in this paper, we focus on noise suppression with *a microphone array*, or collocated multiple microphones.

Aside from ASR, noise suppression has many other applications such as hands-free telecommunications, videoconferencing, and hearing aids. Hands-free telephony would enable you to make a phone call safely even during driving, but noise would degrade the speech quality significantly. Noise suppression would enable clear communication even in such a case. The hearing impaired have much more difficulty in understanding speech in noisy environments than the hearing unimpaired [6]. Hearing aids combined with noise suppression can enhance speech intelligibility for the hearing impaired.

A noise suppression method is typically based on knowledge of the direction from which the target signal arrives (*i.e.* the *target DOA*), and its performance largely depends on the accuracy of this knowledge. In practice, however, this information is rarely available *a priori*. Therefore, estimation of the target DOAs from the observed data is another important task.

DOA estimation has also other applications such as automatic camera pointing, source separation, *etc.* This would enable automatic camera steering activated by voice commands, whereby enabling user-friendly and realistic videoconference [7].

1.2 Difficulty and our goal

Techniques such as adaptive signal processing are known to enable efficient noise suppression and DOA estimation, if noise is *directional*. Examples of such noise include speech utterances from an interfering speaker and speech and music from a television. This kind of noise can be dealt with by adaptively controlling the *directivity* (*i.e.* direction-dependent gain) of a microphone array to form nulls into noise directions.

However, noise in the real world is not always directional, but often rather *diffuse*. Diffuse noise refers to noise that comes from many directions and has little dependency on the direction. For instance, we encounter such noise when many people are speaking at the same time in the street or at a party. Another example is noise in car or on train that is caused by the vibration of the body and the windows, which constitute *surface noise sources* instead of *point noise sources*. Such diffuse noise cannot be suppressed sufficiently by directivity control only, and causes errors in DOA estimation as well. For this reason, diffuse noise has been one of the major obstacles in applying microphone array signal processing to the real world.

This thesis aims to establish microphone array signal processing techniques that are *robust* against diffuse noise, or in other words, techniques that works well even in the presence of diffuse noise. Specifically, we focus on two tasks, namely noise suppression and DOA estimation.

1.3 Our approach

Firstly, we introduce a general linear subspace model of the noise covariance matrix that extends three state-of-the-art models, and introduce a fourth more flexible real-valued noise covariance model.

Secondly, we apply this general model to the task of diffuse noise suppression with a known target steering vector. In the state-of-the-art Wiener post-filtering approach, it is essential to accurately estimate the target power spectrogram. We propose a unified estimation framework applicable to the general noise model, which is based on orthogonal projection of the observed covariance matrix onto the orthogonal complement of the noise model subspace. Ideally, this orthogonal projection is noise-free, and enables accurate estimation of the target power spectrogram.

Thirdly, we address the task of DOA estimation of multiple sources. The performance of the state-of-the-art MUltiple SIgnal Classification (MUSIC) algorithm is known to degrade in the presence of diffuse noise. In order to mitigate this effect, we estimate the signal covariance matrix and subsequently apply MUSIC to it. The estimation relies on the above-mentioned noise-free component of the observed covariance matrix and on the reconstruction of the remaining component belonging to the noise subspace. We propose two algorithms based on different matrix completion algorithms: a first approach based on low-rank matrix completion, which uses the knowledge on the number of sources, and a second one based on trace norm minimization, which does not require that knowledge.

Finally, we present a unified framework applicable to the general noise model for diffuse noise suppression with an unknown target steering vector. This is important for effective noise suppression in the real-world, because the steering vector is usually not accurately known in practice. We jointly estimate the target steering vector and the target power spectrogram for designing the beamformer and the Wiener post-filter. The estimation is based on rank-1 completion and Principal Component Analysis (PCA).

This work has led to one journal paper [8], 3 international conference papers [9, 10, 11], and 8 domestic conference papers [12, 13, 14, 15, 16, 17, 18, 19].

1.4 Structure of the thesis

The rest of this thesis is organized as follows.

In Chapter 2, we formally define the tasks we consider in this thesis, namely noise suppression and DOA estimation in the presence of diffuse noise. We describe the standard time-frequency domain processing of the observed signals, and introduce the notion of covariance matrices. We then review the state of the art of noise suppression and DOA estimation, and summarize the limitation of existing approaches.

In Chapter 3, we present the proposed unified framework for noise modeling, and show that this includes previous noise models as special cases. Subsequently, we introduce a more flexible real-valued noise covariance model. We experimentally assess the fit of each model to real-world noise.

In Chapter 4, we describe the proposed unified framework for diffuse noise suppression with a known target steering vector. We present a unified estimator of the target power spectrogram based on the general noise model. We also derive the specific estimator for each

noise model. We assess the proposed framework for noise suppression through experiments with real-world noise.

In Chapter 5, we describe the proposed unified framework for DOA estimation. We design two alternative algorithms based on low-rank matrix completion and trace-norm minimization. Finally, we evaluate the proposed methods using a large database with real-world noise.

In Chapter 6, we present the proposed unified framework for diffuse noise suppression with an unknown target steering vector. We describe an icosahedral microphone array we fabricated and the real-world data we recorded using it. Finally, we evaluate the proposed method with the recorded data.

Finally, we conclude this thesis in Chapter 7.

Chapter 2

Tasks Considered and State of the Art

This chapter has partly been published in [8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 20].

2.1 Definition of noise suppression and direction-of-arrival estimation

Let us suppose that an array of M microphones receives L target signals emitted from point sources in the presence of diffuse noise and/or reverberation. We denote the target signals received by the first microphone by $s_l(t)$, $l = 1, 2, \ldots, L$, and stack them into the vector

$$\boldsymbol{s}(t) \triangleq \begin{bmatrix} s_1(t) & s_2(t) & \cdots & s_L(t) \end{bmatrix}^{\mathsf{T}}. \tag{2.1}$$

We denote the vector of signals received by the microphones and that of diffuse noise by $\boldsymbol{x}(t) \in \mathbb{R}^M$ and $\boldsymbol{v}(t) \in \mathbb{R}^M$, respectively. Assuming that the target sources are static, we can model the signal transmission from sources to microphones by linear time-invariant *mixing filters*. These filters are single shifted impulses for planewave or spherical wave propagation, and consist of numerous impulses in reverberant rooms. Therefore, $\boldsymbol{x}(t)$ can be modeled as follows:

$$\boldsymbol{x}(t) = \int_0^\infty \boldsymbol{H}(t')\boldsymbol{s}(t - t')dt' + \boldsymbol{v}(t), \tag{2.2}$$

where $\boldsymbol{H}(t) \in \mathbb{R}^{M \times L}$ is the matrix of the impulse responses of the mixing filters. This covers a wide range of scenarios: from a single point source ($L = 1$) with known mixing filters to multiple point sources ($L \geq 2$) with unknown mixing filters.

In practice, the observed signal is sampled in time. The corresponding discrete-time version of (2.2) is given by

$$\boldsymbol{x}[k] = \sum_{k'=0}^{\infty} \boldsymbol{H}[k']\boldsymbol{s}[k-k'] + \boldsymbol{v}[k], \tag{2.3}$$

where $k \in \mathbb{Z}$ is the time index.

Diffuse noise suppression is the task of estimating $\boldsymbol{s}[k]$ given $\boldsymbol{x}[k]$. In the following, we shall restrict ourselves to the single-source case ($L = 1$). In this case, the observation model (2.3) reduces to

$$\boldsymbol{x}[k] = \sum_{k'=0}^{\infty} s[k-k']\boldsymbol{h}[k'] + \boldsymbol{v}[k], \tag{2.4}$$

where $\boldsymbol{h}[k] \in \mathbb{R}^M$ is the vector of the impulse responses of the mixing filters, and $s[k]$ the target signal. $\boldsymbol{h}[k]$ may or may not be given, and the problem is said to be *blind* in the latter case. We focus on noise suppression for one source with known $\boldsymbol{h}[k]$ first, because this is a basic task that has been widely studied in the literature. We move to noise suppression for one source with unknown $\boldsymbol{h}[k]$ later on. We show experimentally that the proposed method for noise suppression with a known steering vector works even in the presence of several sources, *i.e.* directional interferers.

On the other hand, DOA estimation is the task of estimating the DOAs of the target signals given $\boldsymbol{x}[k]$. When the target sources are in the far field of the array, their location is specified by two parameters, namely the azimuth and the zenith angle. In this paper, we assume that the target signal is at the same height as the microphone array, and focus on the estimation of the azimuth for simplicity. However, the proposed technique can be extended to the estimation of both easily. We consider the general case of an unknown $\boldsymbol{H}[k]$ and multiple target sources ($L \geq 2$). Taking into account only the direct path in $\boldsymbol{H}[k]$ and considering the reflections on the walls to be part of noise $\boldsymbol{v}[k]$, the $\boldsymbol{H}[k]$ accounts for planewaves and is parametrized by the azimuth.

2.2 Time-frequency domain processing and covariance matrices

The Short-Time Fourier Transform (STFT) is commonly used for the analysis of time-varying signals like speech. In array signal processing in this thesis, we first analyze the observed signals by STFT, and perform processing in the time-frequency domain.

The STFT of a signal $\alpha[k]$ is defined using an *analysis window* $w[k]$ that has a compact support $[-K/2, K/2-1]$ and tapers to zero at both ends such as the Hanning window [21]:

$$w[k] = \frac{1}{2}\left(1 + \cos\frac{2\pi k}{K}\right). \tag{2.5}$$

The STFT of $\alpha[k]$ is defined as

$$\alpha(\tau,\omega) = \sum_{k=\tau S-K/2}^{\tau S+K/2-1} \alpha[k]w[k-\tau S]\mathrm{e}^{-\mathrm{j}2\pi k\omega/\omega_\mathrm{s}}, \tag{2.6}$$

with τ denoting the frame index, S the frame hop, $\omega \in \{0, \frac{\omega_\mathrm{s}}{K}, \ldots, \frac{(K-1)\omega_\mathrm{s}}{K}\}$ the angular frequency, and ω_s the sampling angular frequency.

When K is large enough compared to the number of taps of all mixing filters $h_{ml}[k]$, (2.3) is approximated by [21]:

$$\boldsymbol{x}(\tau,\omega) = \boldsymbol{H}(\omega)\boldsymbol{s}(\tau,\omega) + \boldsymbol{v}(\tau,\omega), \tag{2.7}$$

where $\boldsymbol{H}(\omega)$ is defined by $h_{ml}(\omega) = \mathcal{F}[h_{ml}[k]]$ ($\mathcal{F}[\cdot]$: discrete-time Fourier transform), and $\boldsymbol{x}(\tau,\omega)$, $\boldsymbol{s}(\tau,\omega)$, and $\boldsymbol{v}(\tau,\omega)$ denote the STFT of $\boldsymbol{x}[k]$, $\boldsymbol{s}[k]$, and $\boldsymbol{v}[k]$, respectively. In DOA estimation, $\boldsymbol{H}(\omega)$ is parametrized by the target azimuths $\xi_1, \ldots, \xi_L$ to be estimated. Indeed, the l-th column of $\boldsymbol{H}(\omega)$ is given by

$$\boldsymbol{h}_l(\omega) = \boldsymbol{h}(\omega;\xi_l) \triangleq \begin{bmatrix}\mathrm{e}^{-\mathrm{j}\omega\delta_1(\xi_l)} & \ldots & \mathrm{e}^{-\mathrm{j}\omega\delta_M(\xi_l)}\end{bmatrix}^\mathsf{T}, \tag{2.8}$$

where the delay $\delta_m(\xi_l)$ is given by

$$\delta_m(\xi_l) = -\frac{\boldsymbol{p}^\mathsf{T}(\xi_l)(\boldsymbol{r}_m - \boldsymbol{r}_1)}{c}, \tag{2.9}$$

where $\boldsymbol{r}_m$ denotes the coordinates of the m-th microphone, and $\boldsymbol{p}(\xi)$ is the unit DOA vector of the planewave from the azimuth ξ:

$$\boldsymbol{p}(\xi) \triangleq \begin{bmatrix}\cos\xi & \sin\xi & 0\end{bmatrix}^\mathsf{T}. \tag{2.10}$$

Therefore, (2.7) is written in the following form as well:

$$\boldsymbol{x}(\tau,\omega) = \sum_{l=1}^{L} s_l(\tau,\omega)\boldsymbol{h}(\omega;\xi_l) + \boldsymbol{v}(\tau,\omega). \tag{2.11}$$

Classically, $\boldsymbol{s}(\tau,\omega)$ and $\boldsymbol{v}(\tau,\omega)$ are assumed to be zero-mean Gaussian random variable, and modeled by their respective covariance matrices [22, 23]:

$$\boldsymbol{x}(\tau,\omega) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)), \tag{2.12}$$

$$\boldsymbol{s}(\tau,\omega) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi}_{\boldsymbol{ss}}(\tau,\omega)), \tag{2.13}$$

$$\boldsymbol{v}(\tau,\omega) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega)). \tag{2.14}$$

Here, the covariance matrices are defined as

$$\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega) \triangleq \mathcal{E}[\boldsymbol{x}(\tau,\omega)\boldsymbol{x}^{\mathsf{H}}(\tau,\omega)] \in \mathbb{C}^{M\times M}, \tag{2.15}$$

$$\boldsymbol{\Phi}_{\boldsymbol{ss}}(\tau,\omega) \triangleq \mathcal{E}[\boldsymbol{s}(\tau,\omega)\boldsymbol{s}^{\mathsf{H}}(\tau,\omega)] \in \mathbb{C}^{L\times L}, \tag{2.16}$$

$$\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega) \triangleq \mathcal{E}[\boldsymbol{v}(\tau,\omega)\boldsymbol{v}^{\mathsf{H}}(\tau,\omega)] \in \mathbb{C}^{M\times M}. \tag{2.17}$$

Note that covariance matrices are Hermitian positive semidefinite by definition. Assuming that the signal and noise are mutually uncorrelated, we have the following relationship among (2.15) to (2.17):

$$\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega) = \boldsymbol{H}(\omega)\boldsymbol{\Phi}_{\boldsymbol{ss}}(\tau,\omega)\boldsymbol{H}^{\mathsf{H}}(\omega) + \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega). \tag{2.18}$$

Indeed,

$$\begin{aligned}
\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega) &= \mathcal{E}[\boldsymbol{x}(\tau,\omega)\boldsymbol{x}^{\mathsf{H}}(\tau,\omega)] && (2.19)\\
&= \mathcal{E}[\{\boldsymbol{H}(\omega)\boldsymbol{s}(\tau,\omega) + \boldsymbol{v}(\tau,\omega)\}\{\boldsymbol{H}(\omega)\boldsymbol{s}(\tau,\omega) + \boldsymbol{v}(\tau,\omega)\}^{\mathsf{H}}] && (2.20)\\
&= \boldsymbol{H}(\omega)\mathcal{E}[\boldsymbol{s}(\tau,\omega)\boldsymbol{s}^{\mathsf{H}}(\tau,\omega)]\boldsymbol{H}^{\mathsf{H}}(\omega) + \boldsymbol{H}(\omega)\mathcal{E}[\boldsymbol{s}(\tau,\omega)\boldsymbol{v}^{\mathsf{H}}(\tau,\omega)] && (2.21)\\
&\quad + \mathcal{E}[\boldsymbol{v}(\tau,\omega)\boldsymbol{s}^{\mathsf{H}}(\tau,\omega)]\boldsymbol{H}^{\mathsf{H}}(\omega) + \mathcal{E}[\boldsymbol{v}(\tau,\omega)\boldsymbol{v}^{\mathsf{H}}(\tau,\omega)] && \\
&= \boldsymbol{H}(\omega)\boldsymbol{\Phi}_{\boldsymbol{ss}}(\tau,\omega)\boldsymbol{H}^{\mathsf{H}}(\omega) + \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega). && (2.22)
\end{aligned}$$

Here, $\mathcal{E}[\boldsymbol{s}(\tau,\omega)\boldsymbol{v}^{\mathsf{H}}(\tau,\omega)]$ and $\mathcal{E}[\boldsymbol{v}(\tau,\omega)\boldsymbol{s}^{\mathsf{H}}(\tau,\omega)]$ are zeros because of the uncorrelatedness of $\boldsymbol{s}(\tau,\omega)$ and $\boldsymbol{v}(\tau,\omega)$. Especially, for the case $L = 1$, this reduces to

$$\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega) = \phi_{ss}(\tau,\omega)\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega) + \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega), \tag{2.23}$$

where

$$\phi_{ss}(\tau,\omega) \triangleq \mathcal{E}[|s(\tau,\omega)|^2] \tag{2.24}$$

denotes the power spectrogram of $s(\tau,\omega)$. In practice, the short-time covariance matrix of $\boldsymbol{x}(\tau,\omega)$ is computed by empirical averaging over a few consecutive frames around the frame of interest:

$$\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega) = \frac{1}{2Q+1}\sum_{\tau'=\tau-Q}^{\tau'=\tau+Q} \boldsymbol{x}(\tau',\omega)\boldsymbol{x}^{\mathsf{H}}(\tau',\omega), \tag{2.25}$$

where $2Q+1$ is the number of frames used for averaging.

2.3 State of the art of noise suppression

2.3.1 Directivity control

Delay-and-sum beamformer [24, 25]

The most fundamental noise suppression technique is the Delay-and-Sum (DS) beamformer. Assuming that the target signal propagates as a planewave, the DS beamformer first shifts the observed signals so that the target signal components in these signals are temporally aligned. The output $y(\tau, \omega)$ of the DS beamformer is given by the average of these time-aligned signals as follows:

$$y(\tau, \omega) = \frac{1}{M} \sum_{m=1}^{M} \mathrm{e}^{\mathrm{j}\omega\delta_m} x_m(\tau, \omega), \tag{2.26}$$

where δ_m is the time it takes for the target signal to propagate from the first microphone to the m-th microphone ($\delta_1 = 0$). The delays δ_m are assumed to be known or to have been estimated. The target signal is summed up constructively, while signals coming from directions other than the target direction are summed up destructively, so that the target signal is enhanced relatively to noise. The beamformer's power response to a planewave is a function of its direction as shown in Fig. 2.1, which is called a *directivity pattern*. There is a region with high gains around the target direction (90° in the figure), which is called a beam.

The output of the DS beamformer (2.26) can be viewed as a Least-Squares (LS) estimate of the target signal. Let us define the cost function by

$$\|\boldsymbol{x}(\tau, \omega) - s(\tau, \omega)\boldsymbol{h}(\omega)\|_2^2, \tag{2.27}$$

and assume that the steering vector has the form corresponding to a planewave as follows:

$$\boldsymbol{h}(\omega) = \begin{bmatrix} 1 & \mathrm{e}^{-\mathrm{j}\omega\delta_2} & \dots & \mathrm{e}^{-\mathrm{j}\omega\delta_M} \end{bmatrix}^{\mathsf{T}}. \tag{2.28}$$

From the orthogonality principle, the solution minimizing (2.27) is given by

$$s(\tau, \omega) = \frac{\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{x}(\tau, \omega)}{\|\boldsymbol{h}(\omega)\|_2^2} = \frac{1}{M} \sum_{m=1}^{M} \mathrm{e}^{\mathrm{j}\omega\delta_m} x_m(\tau, \omega). \tag{2.29}$$

The downside of this technique is that it requires a large array aperture and a large number of microphones in order to obtain a single and sharp beam at the target DOA in the directivity pattern. Fig. 2.1 shows directivity patterns of the DS beamformer with a uniform

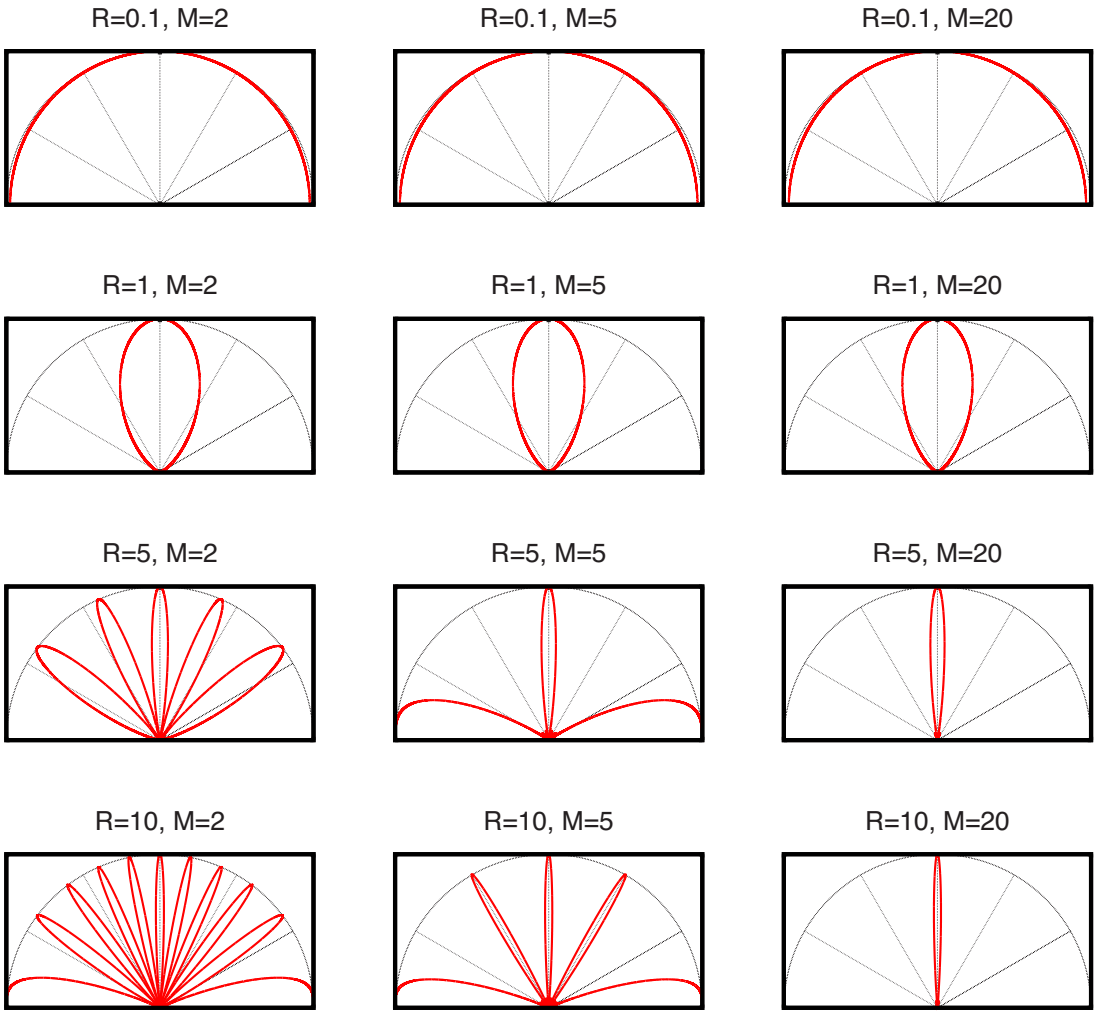


Figure 2.1 Directivity patterns of the DS beamformer for the uniform linear array for several values of R and M, where $R \triangleq \frac{MD}{\lambda}$ is the array size MD divided by the wavelength λ, and M the number of microphones (D: separation between adjacent microphones).

linear array for varying array sizes and varying numbers of microphones. The sharpness of the beam in the target direction is determined by the ratio $R \triangleq \frac{MD}{\lambda}$ of the array size MD and the wavelength λ, where D denotes the distance between adjacent microphones. R must be large for a sharp beam. Also, as seen from the figure, when $\frac{D}{\lambda} = \frac{R}{M}$ is too large, extra beams appear in directions other than the target direction. This is a phenomenon called *spatial aliasing*. To avoid this adverse effect, we must keep $\frac{D}{\lambda}$ sufficiently small. In order to make R large and make $\frac{D}{\lambda}$ small at the same time, we must make M large.

Null beamformer [26]

In contrast to the DS beamformer, null beamformers are applicable to small arrays with a few microphones. Specifically, a null beamformer can eliminate up to $M-1$ directional interferences, regardless of the array size in theory, if their steering vectors are given. It directs nulls in the directivity pattern into the directions of the interferences by applying a weight vector $\boldsymbol{w}(\omega)$ that is orthogonal to their steering vectors.

Specifically, let us consider estimating $s(\tau,\omega)$ with an estimator of the following form:

$$\hat{s}(\tau,\omega) \triangleq \boldsymbol{w}^{\mathsf{H}}(\omega)\boldsymbol{x}(\tau,\omega), \tag{2.30}$$

and let $\boldsymbol{h}_1(\omega)$ and $\boldsymbol{h}_l(\omega)$, $l=2,\ldots,I$ denote the target steering vectors and those of the interferences to which we would like to steer nulls ($I \leq M$). We assume that

$$\boldsymbol{h}_1(\omega) \notin \mathrm{span}\{\boldsymbol{h}_i(\omega) | 2 \leq i \leq I\}, \tag{2.31}$$

which is almost always the case in practice. Then, there exists $\boldsymbol{w}(\omega)$ such that $\boldsymbol{w}^{\mathsf{H}}(\omega)\boldsymbol{h}_1(\omega) = 1$ and $\boldsymbol{w}^{\mathsf{H}}(\omega)\boldsymbol{h}_l(\omega) = 0$, $l=2,\ldots,I$. Such $\boldsymbol{w}(\omega)$ eliminates the interferences without distorting the target signal.

In practice, such $\boldsymbol{w}(\omega)$ can be obtained as follows. First, we derive a vector $\tilde{\boldsymbol{w}}(\omega)$ satisfying $\tilde{\boldsymbol{w}}^{\mathsf{H}}(\omega)\boldsymbol{h}_l(\omega) = 0$, $l=2,\ldots,I$, or equivalently, $\tilde{\boldsymbol{H}}^{\mathsf{H}}(\omega)\tilde{\boldsymbol{w}}(\omega) = \boldsymbol{0}$, where $\tilde{\boldsymbol{H}}(\omega) \triangleq \begin{bmatrix}\boldsymbol{h}_2(\omega) & \ldots & \boldsymbol{h}_I(\omega)\end{bmatrix}$. This is obtained *e.g.* as a right singular vector corresponding to a zero singular value of $\tilde{\boldsymbol{H}}^{\mathsf{H}}(\omega)$. Then, a desired $\boldsymbol{w}(\omega)$ is obtained by scaling $\tilde{\boldsymbol{w}}(\omega)$ by $\boldsymbol{w}(\omega) = \frac{\tilde{\boldsymbol{w}}(\omega)}{[\tilde{\boldsymbol{w}}^{\mathsf{H}}(\omega)\boldsymbol{h}_1(\omega)]^*}$ so that it satisfies $\boldsymbol{w}^{\mathsf{H}}(\omega)\boldsymbol{h}_1(\omega) = 1$.

Although $\boldsymbol{h}_l(\omega)$, $l=2,\ldots,I$ can be known, when the DOAs of the noises are known and we assume planewave propagation, this is not always the case.

Minimum Variance Distortionless Response (MVDR) beamformer [26]

Compared to a null beamformer with known noise steering vectors, this beamformer enables null steering without knowing the noise steering vectors. It is derived as the beamformer weight that results in the minimum output power (variance) under the constraint that the target signal is not distorted.

Specifically, we would like to solve the following optimization problem:

$$\min_{\boldsymbol{w}(\tau,\omega)} \boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)\boldsymbol{w}(\tau,\omega) \quad \text{s.t.} \quad \boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{h}(\omega) = 1. \tag{2.32}$$

This optimization is attained by minimizing the following cost function:

$$\boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)\boldsymbol{w}(\tau,\omega) - \lambda_1(\tau,\omega)\{\Re[\boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{h}(\omega)] - 1\} - \lambda_2(\tau,\omega)\Im[\boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{h}(\omega)] \tag{2.33}$$

$$= \boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)\boldsymbol{w}(\tau,\omega) - \Re\{\lambda(\tau,\omega)[\boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{h}(\omega) - 1]\}, \tag{2.34}$$

where $\lambda_1(\tau,\omega), \lambda_2(\tau,\omega) \in \mathbb{R}$ are real-valued Lagrangian multipliers, and $\lambda(\tau,\omega) \triangleq \lambda_1(\tau,\omega) + \mathrm{j}\lambda_2(\tau,\omega)$. Differentiating this with respect to $\boldsymbol{w}^*(\tau,\omega)$ and equating the result to zero, we have

$$\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)\boldsymbol{w}(\tau,\omega) - \frac{1}{2}\lambda(\tau,\omega)\boldsymbol{h}(\omega) = \boldsymbol{0}. \tag{2.35}$$

Solving this equation with respect to $\boldsymbol{w}(\tau,\omega)$, we obtain

$$\boldsymbol{w}(\tau,\omega) = \frac{1}{2}\lambda(\tau,\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\tau,\omega)\boldsymbol{h}(\omega). \tag{2.36}$$

Substituting this into the constraint as

$$\frac{1}{2}\lambda^*(\tau,\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\tau,\omega)\boldsymbol{h}(\omega) = 1, \tag{2.37}$$

we obtain $\lambda(\tau,\omega)$ as follows:

$$\lambda(\tau,\omega) = \frac{2}{\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\tau,\omega)\boldsymbol{h}(\omega)}. \tag{2.38}$$

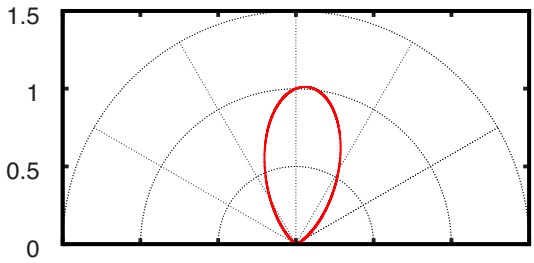
Thus, the optimal weight vector is

$$\boldsymbol{w}_{\mathrm{MVDR}}(\tau,\omega) \triangleq \frac{\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\tau,\omega)\boldsymbol{h}(\omega)}{\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\tau,\omega)\boldsymbol{h}(\omega)}. \tag{2.39}$$

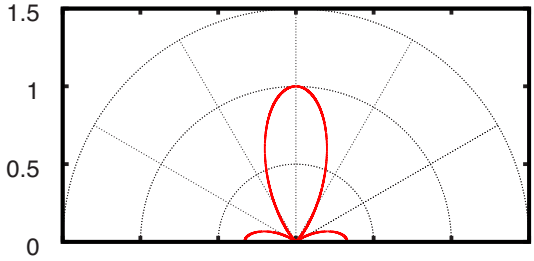
Similarly to null beamformers, the MVDR beamformer can perfectly suppress up to $M-1$ directional interferences, but the suppression of diffuse noise from a large number of directions is insufficient. Fig. 2.2 shows examples of the directivity pattern of the MVDR beamformer (a) for directional interferences and (b) for diffuse noise. The target DOA was 90° in both cases. The beamformer formed nulls in the noise directions (0° and 150°) for directional interferences, but did not manage to suppress all noise directions for diffuse noise.

2.3.2 Post-filtering

Recently, an approach of post-filtering, *i.e.* time-frequency masking at the output of a beamformer, has been studied as a promising technique for diffuse noise suppression [27, 28,



(a)



(b)

Figure 2.2 Examples of the directivity pattern of the MVDR beamformer for a target located at 90° in the presence of (a) directional noise (DOAs: 0°, 150°) or (b) diffuse noise (spherically isotropic noise).

29, 30, 31, 5, 32, 33, 34, 35]. The directivity control approach described in Section 2.3.1 has only a limited ability of suppressing diffuse noise because of the limited number of nulls that can be formed. On the other hand, time-frequency masking is based on the diversity between the target signal and noise in the time-frequency domain rather than directional diversity. Therefore, it can be effective even for diffuse noise. Especially, Simmer *et al.* [30] and Van Trees [36] showed that the Linear Minimum Mean Square Error (LMMSE) estimator of the target signal is obtained by the MVDR beamformer followed by a time-frequency mask called the Wiener post-filter [30, 5, 35]:

$$\hat{s}(\tau,\omega) = \frac{\phi_{ss}(\tau,\omega)}{\phi_{yy}(\tau,\omega)} \cdot \underbrace{\frac{\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\tau,\omega)\boldsymbol{x}(\tau,\omega)}{\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\tau,\omega)\boldsymbol{h}(\omega)}}_{=\,y(\tau,\omega)}. \tag{2.40}$$

Here,

$$p(\tau,\omega) \triangleq \frac{\phi_{ss}(\tau,\omega)}{\phi_{yy}(\tau,\omega)} \tag{2.41}$$

is the Wiener post-filter.

Equation (2.40) constitutes the LMMSE estimator of the target signal, or the multichannel

Wiener filter. Let us consider a linear estimator of the form

$$\hat{s}(\tau,\omega) \triangleq \boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{x}(\tau,\omega), \tag{2.42}$$

and consider minimizing the mean square error

$$\mathcal{E}[|\boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{x}(\tau,\omega) - s(\tau,\omega)|^2]. \tag{2.43}$$

Partial differentiation of this with respect to $\boldsymbol{w}^*(\tau,\omega)$ leads to

$$\hat{s}(\tau,\omega) \triangleq \phi_{ss}(\tau,\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\tau,\omega)\boldsymbol{x}(\tau,\omega). \tag{2.44}$$

Noting that

$$\phi_{yy}(\tau,\omega) = \frac{1}{\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\tau,\omega)\boldsymbol{h}(\omega)}, \tag{2.45}$$

we obtain (2.40).

In the design of the Wiener post-filter (2.41), it is essential to accurately estimate the target power spectrogram or equivalently the short-time target autocorrelation function from the noisy signals observed at the microphones. We can compute the covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)$ of the observed signal, *e.g.* by (2.25), but cannot compute $\phi_{ss}(\tau,\omega)$ in this way, because we cannot observe $s(\tau,\omega)$. $\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)$ is linked to $\phi_{ss}(\tau,\omega)$ as in (2.23). Therefore, for known $\boldsymbol{h}(\omega)$, we obtain $\phi_{ss}(\tau,\omega)$, if we can eliminate some entries or components of $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega)$ by some transformation.

Zelinski's design of Wiener post-filter [27]

Zelinski's estimator of $\phi_{ss}(\tau,\omega)$ is based on the assumption of spatially uncorrelated noise. This assumption implies that $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega)$ is diagonal, and $\phi_{ss}(\tau,\omega)$ can be obtained from the noise-free off-diagonal entries. Although Zelinski's method was originally presented in the time domain, we describe here its equivalence in the time-frequency domain for the ease of comparison.

Specifically, the interchannel cross-spectra of observed signals are noise-free as follows:

$$\phi_{x_m x_n}(\tau,\omega) = \phi_{ss}(\tau,\omega)h_m(\omega)h_n^*(\omega) \ (m \neq n). \tag{2.46}$$

Substituting the planewave model (2.28) to $h_m(\omega)$, (2.46) becomes

$$\phi_{x_m x_n}(\tau,\omega) = \phi_{ss}(\tau,\omega)\mathrm{e}^{-\mathrm{j}\omega(\delta_m-\delta_n)} \ (m \neq n). \tag{2.47}$$

Solving (2.46) for $\phi_{ss}(\tau,\omega)$ and averaging the result over the microphone pairs, we obtain the estimator

$$\hat{\phi}_{ss}^{\mathrm{Zel}}(\tau,\omega) = \frac{1}{M(M-1)} \sum_{m \neq n} \phi_{x_m x_n}(\tau,\omega) \mathrm{e}^{\mathrm{j}\omega(\delta_m - \delta_n)} \tag{2.48}$$

$$= \frac{2}{M(M-1)} \sum_{m<n} \Re[\phi_{x_m x_n}(\tau,\omega) \mathrm{e}^{\mathrm{j}\omega(\delta_m - \delta_n)}]. \tag{2.49}$$

The cross-spectrogram $\phi_{x_m x_n}(\tau,\omega)$ in this equation can be estimated by, for example, averaging $x_m(\tau,\omega)x_n^*(\tau,\omega)$ temporally over several adjacent frames.

A detailed analysis on Zelinski's post-filter can be found in [37].

McCowan's design of Wiener post-filter [5]

Zelinski's model is inappropriate for modeling diffuse noise observed by a small array, because the noise highly correlates between microphones in that case. Instead of neglecting inter-channel noise correlation as in Zelinski's method, McCowan's method is based on the assumption that the inter-channel noise coherences are given. Here we present a slightly modified version of McCowan's method, which is more theoretically sound as explained later.

The assumption is based on the fact that the noise coherences are known for some ideal noise fields. For example, consider *spherically isotropic noise*, which is composed of noise planewaves with an equal power spectrum propagating in any directions in the three-dimensional space. In this case, the noise coherence between the m-th and the n-th microphones is given by [38]

$$\gamma_{v_m v_n}(\tau,\omega) \triangleq \frac{\phi_{v_m v_n}(\tau,\omega)}{\sqrt{\phi_{v_m v_m}(\tau,\omega)}\sqrt{\phi_{v_n v_n}(\tau,\omega)}} \tag{2.50}$$

$$= \mathrm{sinc}\left(\frac{r_{mn}\omega}{c}\right), \tag{2.51}$$

where r_{mn} is the distance between the microphones, and c the velocity of sound. Another example is *cylindrically isotropic noise*, which is defined in the same way as spherically isotropic noise, except that noise propagates two-dimensionally in the horizontal directions. In this case, the noise coherence is given by [39]

$$\gamma_{v_m v_n}(\tau,\omega) = J_0\left(\frac{l_{mn}\omega}{c}\right). \tag{2.52}$$

Here, $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind, and l_{mn} is the distance between the orthogonal projections onto the horizontal plane of the m-th and the n-th microphones.

McCowan's method also assumes that the power spectrogram of noise is identical at all microphones, which is true for spherically/cylindrically isotropic noise:

$$\phi_{v_1 v_1}(\tau,\omega) = \cdots = \phi_{v_M v_M}(\tau,\omega) =: \phi_{vv}(\tau,\omega). \tag{2.53}$$

In this case, from (2.50), we have

$$\phi_{v_m v_n}(\tau,\omega) = \phi_{vv}(\tau,\omega)\gamma_{v_m v_n}(\tau,\omega), \tag{2.54}$$

and therefore

$$\phi_{x_m x_n}(\tau,\omega) = \phi_{ss}(\tau,\omega)h_m(\omega)h_n^*(\omega) + \phi_{vv}(\tau,\omega)\gamma_{v_m v_n}(\tau,\omega). \tag{2.55}$$

McCowan's estimator of $\phi_{ss}(\tau,\omega)$ is obtained based on solving the system of equations (2.55) for $\phi_{ss}(\tau,\omega)$ and averaging the result over all microphone pairs as follows:

$$\hat{\phi}_{ss}^{\mathrm{Mc}}(\tau,\omega) = \frac{2}{M(M-1)}\sum_{m<n}\frac{\Re\left[\dfrac{\phi_{x_m x_n}(\tau,\omega)}{h_m(\omega)h_n^*(\omega)}\right] - \dfrac{\phi_{x_m x_m}(\tau,\omega)+\phi_{x_n x_n}(\tau,\omega)}{2}\Re\left[\dfrac{\gamma_{v_m v_n}(\tau,\omega)}{h_m(\omega)h_n^*(\omega)}\right]}{1 - \dfrac{|h_m(\omega)|^2+|h_n(\omega)|^2}{2}\Re\left[\dfrac{\gamma_{v_m v_n}(\tau,\omega)}{h_m(\omega)h_n^*(\omega)}\right]}. \tag{2.56}$$

Let us comment here about the difference between McCowan's original estimator and the modified estimator (2.56). The difference lies in the term $\frac{|h_m(\omega)|^2+|h_n(\omega)|^2}{2}$. In the original estimator, the noise signals time-shifted so that the target signal is in phase, namely $v_m(\tau,\omega)/h_m(\omega)$, is assumed to be spherically/cylindrically isotropic [40, 38]. This resulted in

$$\frac{\phi_{v_m v_n}(\tau,\omega)}{h_m(\omega)h_n^*(\omega)} = \begin{cases} \phi_{vv}(\tau,\omega)\mathrm{sinc}\left(\dfrac{r_{mn}\omega}{c}\right) \text{ (spherical)}, \\ \phi_{vv}(\tau,\omega)J_0\left(\dfrac{l_{mn}\omega}{c}\right) \text{ (cylindrical)}. \end{cases} \tag{2.57}$$

However, the factor $\frac{1}{h_m(\omega)h_n^*(\omega)}$ caused by the alignment changes the phase for the planewave case and both the phase and the magnitude in general. The modified version (2.56) models the original noise signals $v_m(\tau,\omega)$ as spherically/cylindrically isotropic, which results in (2.54). Therefore, we have

$$\phi_{v_m v_n}(\tau,\omega) = \begin{cases} \phi_{vv}(\tau,\omega)\mathrm{sinc}\left(\dfrac{r_{mn}\omega}{c}\right) \text{ (spherical)}, \\ \phi_{vv}(\tau,\omega)J_0\left(\dfrac{l_{mn}\omega}{c}\right) \text{ (cylindrical)}. \end{cases} \tag{2.58}$$

This is theoretically sounder and resulted in much better results in our preliminary experiments.

In the experiment in Chapter 4.3, the cylindrically isotropic model was used as the model of inter-channel noise coherences for McCowan's method, because it always resulted in a better result than the spherically isotropic model in our preliminary experiment.

2.3.3 Blind noise decorrelation

In this section, we review another relevant technique of Blind Noise Decorrelation (BND) proposed by Shimizu *et al.* [41, 42, 43]. This approach models diffuse noise as isotropic, and diagonalizes the noise covariance matrix with a constant unitary matrix exploiting symmetrical arrays. Shimizu *et al.* applied this technique to estimation of the target power spectrogram in diffuse noise environments, but not to post-filtering. We will use it for post-filtering and DOA estimation in Chapters 4, 5, and 6.

Zelinski [27] assumed that diffuse noise is spatially uncorrelated, but real-world diffuse noise is highly correlated, especially for small arrays. McCowan *et al.* [5] assumed that diffuse noise has known fixed coherences depending on microphone distances to take noise correlation into account. However, the noise coherence can deviate from a nominal value due to the geometry of noise sources and the room or the diffraction by a rigid mount. Consequently, the assumption of explicit values of noise coherences in McCowan's method may be inaccurate. In comparison, Shimizu *et al.* showed that certain classes of symmetrical arrays enables the diagonalization of the noise covariance matrix by a constant unitary matrix, under the isotropic noise model defined as follows:

1) The noise power spectrogram at all microphones is identical as in Eq. (2.53).

2) The short-time noise cross-spectrogram is identical for all microphone pairs with an equal distance:

$$r_{mn} = r_{m'n'} \;\Rightarrow\; \phi_{v_m v_n}(\tau, \omega) = \phi_{v_{m'} v_{n'}}(\tau, \omega). \tag{2.59}$$

Note that the explicit value of the noise coherence is not assumed.

As an example of BND, consider a 4-element array with its microphones at the vertices of a square (Fig. 2.3). From assumption 1), we have

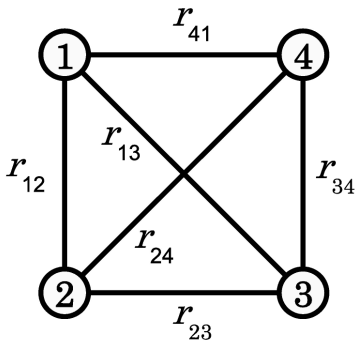


Figure 2.3 Square array. Since $r_{12} = r_{23} = r_{34} = r_{41}$ and $r_{13} = r_{24}$, the covariance matrix of isotropic noise becomes circulant.

$$\phi_{v_1v_1} = \phi_{v_2v_2} = \phi_{v_3v_3} = \phi_{v_4v_4} =: \alpha. \tag{2.60}$$

Furthermore, we have, from assumption 2),

$$\phi_{v_1v_2} = \phi_{v_2v_1} = \phi_{v_2v_3} = \phi_{v_3v_2} = \cdots = \phi_{v_1v_4} =: \beta, \tag{2.61}$$

$$\phi_{v_1v_3} = \phi_{v_3v_1} = \phi_{v_2v_4} = \phi_{v_4v_2} =: \gamma, \tag{2.62}$$

because of $r_{12} = r_{23} = r_{34} = r_{41}$ and $r_{13} = r_{24}$. Consequently, $\boldsymbol{\Phi}_{\boldsymbol{vv}}$ has the following structure: $\boldsymbol{\Phi}_{\boldsymbol{vv}} = \mathrm{circ}(\alpha, \beta, \gamma, \beta)$. Here, $\mathrm{circ}(a_1, a_2, \ldots, a_p)$ denotes the $p \times p$ circulant matrix whose first row is $\begin{bmatrix} a_1 & a_2 & \ldots & a_p \end{bmatrix}$, where p is a positive integer. Being a circulant matrix, $\boldsymbol{\Phi}_{\boldsymbol{vv}}$ is diagonalized by the 4×4 Discrete Fourier Transform (DFT) matrix $\boldsymbol{F}_4$ for any values of α, β, and γ [44].

$\boldsymbol{\Phi}_{\boldsymbol{vv}}$ can also be diagonalized by the following real-valued orthogonal matrix:

$$\boldsymbol{P} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix}. \tag{2.63}$$

Indeed,

$$\boldsymbol{P}^{\mathsf{H}} \boldsymbol{\Phi}_{\boldsymbol{vv}} \boldsymbol{P} = \begin{bmatrix} \alpha + 2\beta + \gamma & 0 & 0 & 0 \\ 0 & \alpha - \gamma & 0 & 0 \\ 0 & 0 & \alpha - \gamma & 0 \\ 0 & 0 & 0 & \alpha - 2\beta + \gamma \end{bmatrix}. \tag{2.64}$$

This means that isotropic noise is decorrelated by the basis transformation $\boldsymbol{P}^{\mathsf{H}} \boldsymbol{v}$. In the limit of small array aperture (practically, sufficiently small aperture compare to the wavelength),

this decorrelation basis can be regarded as a transformation into the sound pressure itself and its spatial gradients w.r.t. x, y, and both.

The array geometries for which the covariance matrix of isotropic noise is diagonalized by a constant unitary matrix are called *crystal arrays* and include the polygonal, rectangular, polyhedral, polygonal prism, and rectangular solid arrays. The proof of diagonalization and the specific form of the diagonalization matrix for each class is given in [45, 8].

This approach is similar to the so-called phase-mode processing with spherical microphone arrays [46, 47, 48, 49, 50], in the sense that both methods basis-transform the observed signals into an orthonormal basis. However, their objectives are different: the BND aims to decorrelate isotropic noise, while the phase-mode processing basically aims to make the beamformer design independent of the array configuration. Also, there is a difference that the basis vectors of blind noise decorrelation are discrete, whereas those of phase-mode processing are continuous.

Note that spherically isotropic noise can be decorrelated via spherical harmonic decomposition [51]. However noise is not always spherically isotropic, because of the distribution of the noise sources, the room shape, the diffraction by a rigid mount, etc. In this case, the noise correlation matrix, or the normalized noise covariance matrix, deviates from the ideal sine cardinal form for spherically isotropic noise. In comparison, the BND can decorrelate isotropic noise with an arbitrary coherence matrix.

BND is also related to the spatio-temporal gradient method proposed by Ando *et al.* [52, 53, 54]. In this approach, the sound pressure and its spatio-temporal gradients at a single point were utilized for acquiring the geometrical information and estimating the location of sound sources. This approach enables source localization with small array aperture and a short observation interval. They used a square array and the same basis vectors as in (2.63) to approximate spatial gradients. In comparison, BND exploits this basis for decorrelation of diffuse noise.

In the next section, we will review the state of the art of direction-of-arrival estimation.

2.4 State of the art of direction-of-arrival estimation

2.4.1 Approach based on time difference of arrival

One of the most common approaches is that based on the TDOA between microphones of the incident wavefront, which is defined as the time it takes for the wavefront to propagate between them. Let us assume for simplicity that the source and the microphones are on the same plane. Given the TDOA for a microphone pair, the set of source positions explaining this TDOA is given by a hyperbolic curve whose focal points are these microphones. Therefore, given the TDOAs for more than two microphone pairs, the source position is determined as the intersection of two hyperbolic curves. Specifically, when the source is in the far field of the microphone array, the hyperbolic curve can be approximated by its asymptotic line. In practice, these hyperbolic curves or lines do not intersect at a single point, if there are more than two microphone pairs, due to the errors in TDOA estimation. Therefore, the source location is determined by minimizing the squared error between the observed TDOAs and those corresponding to the assumed source location [55, 56, 57].

The most fundamental technique for estimating TDOA between a microphone pair is based on the cross-correlation function between microphones. Indeed, when there is only one source and no noise, the TDOA between microphones m and n can be obtained by picking the largest peak in the cross-correlation function of $x_m(t)$ and $x_n(t)$:

$$R_{x_m x_n}(\tau) \triangleq \mathcal{E}[x_m(t) x_n(t-\tau)]. \tag{2.65}$$

A generalized cross-correlation function [58], or the inverse Fourier transform of the cross-spectrum $\phi_{x_m x_n}(\omega)$ weighted by some function $G(\omega)$, is also used aiming at robustness against reverberation and noise:

$$\tilde{R}_{x_m x_n}(\tau) \triangleq \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) \phi_{x_m x_n}(\omega) \mathrm{e}^{\mathrm{j}\omega\tau} d\omega, \tag{2.66}$$

where $\phi_{x_m x_n}(\omega)$ is the cross-spectrum of x_m and x_n. The main limitation of this TDOA-based approach is that only one source is assumed, which results in degraded estimation accuracy in the presence of multiple sources.

2.4.2 Beamforming approach

The beamforming approach is based on steering a beamformer to various directions to obtain a *steered response power*, that is, the power of the beamformer output $y(\tau, \omega)$ as a

function of the beam direction (*look direction*). DOA estimates are obtained as largest peaks in the steered response power. If we denote the beamformer weight corresponding to the look direction ξ by $\boldsymbol{w}(\omega;\xi)$, the steered response power is given by

$$\phi_{yy}(\omega;\xi) = \mathcal{E}[|\boldsymbol{w}^{\mathsf{H}}(\omega;\xi)\boldsymbol{x}|^2] \tag{2.67}$$

$$= \boldsymbol{w}^{\mathsf{H}}(\omega;\xi)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\omega)\boldsymbol{w}^{\mathsf{H}}(\omega;\xi). \tag{2.68}$$

Different designs of $\boldsymbol{w}(\omega;\xi)$ lead to different methods.

Delay-and-sum beamformer

The DS beamformer is the most fundamental example, and corresponds to

$$\boldsymbol{w}(\omega;\xi) = \frac{\boldsymbol{h}(\omega;\xi)}{\|\boldsymbol{h}(\omega;\xi)\|_2^2} \tag{2.69}$$

from the definition of $\boldsymbol{w}(\omega;\xi)$ and (2.29). Here, $\boldsymbol{h}(\omega;\xi)$ depends on ξ as follows:

$$\boldsymbol{h}(\omega;\xi) = \left[1 \quad \mathrm{e}^{-\mathrm{j}\omega\delta_2(\xi)} \quad \dots \quad \mathrm{e}^{-\mathrm{j}\omega\delta_M(\xi)}\right]^{\mathsf{T}}, \tag{2.70}$$

where $\delta_m(\xi)$ denotes the time it takes for a planewave from the azimuth ξ to propagate from the first microphone to the m-th microphone. Therefore,

$$\phi_{yy}(\omega;\xi) = \frac{\boldsymbol{h}^{\mathsf{H}}(\omega;\xi)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega;\xi)}{\|\boldsymbol{h}(\omega;\xi)\|_2^4} \tag{2.71}$$

$$= \frac{1}{M^2}\boldsymbol{h}^{\mathsf{H}}(\omega;\xi)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega;\xi). \tag{2.72}$$

This beamformer necessitates an array with large aperture and many microphones to form a sharp beam in the target direction without spatial aliasing as pointed out in Section 2.3.1. Otherwise, the DOA estimate by this method becomes unreliable due to a broad beam or the spatial aliasing.

MVDR beamformer

The MVDR beamformer (or *Capon's method*), on the other hand, enables DOA estimation with a small array with a few microphones. This is by directing nulls into the direction of incident waves from directions other than the look direction. The weight vector is given by

$$\boldsymbol{w}(\omega;\xi) = \frac{\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\omega)\boldsymbol{h}(\omega;\xi)}{\boldsymbol{h}^{\mathsf{H}}(\omega;\xi)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\omega)\boldsymbol{h}(\omega;\xi)}. \tag{2.73}$$

This leads to

$$\phi_{yy}(\omega;\xi) = \frac{1}{\boldsymbol{h}^{\mathsf{H}}(\omega;\xi)\boldsymbol{\Phi}_{\boldsymbol{xx}}^{-1}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega;\xi)}. \tag{2.74}$$

Although the MVDR beamformer can deal with multiple sources by steering nulls, diffuse noise causes estimation errors, which cannot be eliminated by null steering.

2.4.3 MUltiple SIgnal Classification (MUSIC) [1]

The MVDR beamformer steers nulls into DOAs other than the look direction whereas forming a beam into the look direction. By contrast, MUSIC is based on null steering into *all* incident directions. This approach is applicable to multiple sources as well, and features higher angular resolution compared to the beamforming approach.

The observation model (2.11) implies that, if there are less target sources than the microphones ($L < M$), the target component $\sum_{l=1}^{L} s_l(\tau,\omega)\boldsymbol{h}(\omega;\xi_l)$ resides in the low-dimensional space

$$\mathcal{S}(\omega) \triangleq \mathrm{span}\{\boldsymbol{h}(\omega;\xi_l)\}_l. \tag{2.75}$$

Therefore, each of the basis vectors $\boldsymbol{e}_1(\omega),\ldots,\boldsymbol{e}_{M-L}(\omega)$ of the orthogonal complement of $\mathcal{S}(\omega)$ forms a directivity pattern with nulls in the target directions:

$$|\boldsymbol{e}_i^{\mathsf{H}}(\omega)\boldsymbol{h}(\omega;\xi)|^2\Big|_{\xi=\xi_1,\ldots,\xi_L} = 0 \tag{2.76}$$

The harmonic average of the reciprocals of these directivity patterns

$$f_{\mathrm{N}}(\omega;\xi) \triangleq \frac{1}{\sum_{i=1}^{M-L}|\boldsymbol{e}_i^{\mathsf{H}}(\omega)\boldsymbol{h}(\omega;\xi)|^2} \tag{2.77}$$

$$= \frac{1}{\boldsymbol{h}^{\mathsf{H}}(\omega;\xi)\boldsymbol{E}(\omega)\boldsymbol{E}^{\mathsf{H}}(\omega)\boldsymbol{h}(\omega;\xi)} \tag{2.78}$$

attains peaks at $\xi = \xi_1,\ldots,\xi_L$. Here, $\boldsymbol{E}(\omega) \triangleq \begin{bmatrix}\boldsymbol{e}_1(\omega) & \ldots & \boldsymbol{e}_{M-L}(\omega)\end{bmatrix}$. In this paper, we call $f_{\mathrm{N}}(\omega;\xi)$ the *narrowband MUSIC spectrum*.

It is important in deriving the narrowband MUSIC spectrum to identify the basis $\{\boldsymbol{e}_i\}_{i=1}^{M-L}$. When there is no noise, we can observe the target spatial covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{cc}}(\tau,\omega) \triangleq \boldsymbol{H}(\omega)\boldsymbol{\Phi}_{\boldsymbol{ss}}(\tau,\omega)\boldsymbol{H}^{\mathsf{H}}(\omega)$ in (2.18), and the basis is obtained as its null space. In the case of spatially white noise, the basis is also obtained as the eigenspace of $\boldsymbol{\Phi}_{\boldsymbol{cc}}(\tau,\omega)$ corresponding to the smallest eigenvalue. Even if noise is not spatially white, if the noise covariance

matrix is known up to a scalar, the basis is obtained as the generalized eigenvectors of the matrix pencil $(\boldsymbol{\Phi}_{\boldsymbol{xx}}(\omega), \boldsymbol{\Gamma}_{\boldsymbol{vv}}(\omega))$ corresponding to the smallest generalized eigenvalues [1]. Here, $\boldsymbol{\Gamma}_{\boldsymbol{vv}}(\omega)$ denotes the scaled noise spatial covariance matrix. For some ideal noise fields, $\boldsymbol{\Gamma}_{\boldsymbol{vv}}(\omega)$ is known *a priori*. An example is the spherically isotropic noise field [38, 5], for which

$$\boldsymbol{\Gamma}_{\boldsymbol{vv}}(\omega) = \begin{bmatrix} 1 & \mathrm{sinc}(\frac{\omega r_{12}}{c}) & \cdots & \mathrm{sinc}(\frac{\omega r_{1M}}{c}) \\ \mathrm{sinc}(\frac{\omega r_{21}}{c}) & 1 & \cdots & \mathrm{sinc}(\frac{\omega r_{2M}}{c}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{sinc}(\frac{\omega r_{M1}}{c}) & \mathrm{sinc}(\frac{\omega r_{M2}}{c}) & \cdots & 1 \end{bmatrix}. \tag{2.79}$$

In practice, however, the real-world noise deviates from such ideal model as pointed out, and the identification of $\{\boldsymbol{e}_i\}_{i=1}^{M-L}$ by this method can be unreliable. In such case, estimation of $\boldsymbol{\Phi}_{\boldsymbol{cc}}(\tau, \omega)$ from the observed noisy spatial covariance matrix is important, so that we can identify $\{\boldsymbol{e}_i\}_{i=1}^{M-L}$ as its null space.

In order to integrate the information at different frequencies, we average the narrowband MUSIC spectra over frequencies. Arithmetic, geometric and harmonic averaging lead respectively to [59, 60]:

$$f_{\mathrm{W,A}}(\xi) \triangleq \frac{1}{K} \sum_{\omega_{\min}}^{\omega_{\max}} f_{\mathrm{N}}(\omega; \xi) \tag{2.80}$$

$$f_{\mathrm{W,G}}(\xi) \triangleq \left[\prod_{\omega_{\min}}^{\omega_{\max}} f_{\mathrm{N}}(\omega; \xi) \right]^{1/K} \tag{2.81}$$

$$f_{\mathrm{W,H}}(\xi) \triangleq \frac{K}{\sum_{\omega_{\min}}^{\omega_{\max}} 1/f_{\mathrm{N}}(\omega; \xi)}. \tag{2.82}$$

Here $[\omega_{\min}, \omega_{\max}]$ denotes the frequency range over which averaging is performed, and K denotes the number of frequency bins in this range. We call $f_{\mathrm{W}}(\xi)$ the *wideband MUSIC spectrum*.

The DOA estimates $\{\hat{\xi}_j\}_{j=1}^{J}$ are obtained by picking peaks in $f_{\mathrm{W}}(\xi)$, where J is the assumed number of sources. In practice, $f_{\mathrm{W}}(\xi)$ is discretized by evaluating $f_{\mathrm{W}}(\xi)$ on a finite grid $\{2\pi \frac{i}{I} | i = 0, \cdots, I-1\}$ as follows:

$$f_{\mathrm{W}}[i] = f_{\mathrm{W}}\left(2\pi \frac{i}{I}\right), \ i = 0, \ldots, I-1, \tag{2.83}$$

where I denotes the number of points in the grid. $\{\hat{\xi}_j\}_{j=1}^{J}$ are calculated up to a minimum angular distance Δ by the following algorithm:

Algorithm 1. *Define* $\Omega^{(0)} \triangleq \{0, \cdots, I-1\}$. *Given the number* J *of peaks to be selected, iterate the following for* $j = 1, 2, \cdots, J$*:*

- $\hat{\xi}_j = 2\pi\frac{i_{opt}}{I}$, *where* i_{opt} *denotes the index* i *maximizing* $f_W[i]$ *subject to* $i \in \Omega^{(j)}$ *and* $f_W[i-1] < f_W[i] > f_W[i+1]$.
- $\Omega^{(j+1)} = \Omega^{(j)} - \{i | d(2\pi\frac{i}{I}, \hat{\xi}_j) \leq \Delta\}$, *where* $d(x,y) \triangleq \min\{|x-y|, 2\pi - |x-y|\}$ *is the angular distance.*

There is a trade-off regarding the length of the data from which the subspace $\boldsymbol{E}(\omega)$ in (2.78) is estimated. Longer data is favorable for obtaining a reliable estimate of $\boldsymbol{\Phi}_{\boldsymbol{xx}}$ from which $\boldsymbol{E}(\omega)$ is derived. On the other hand, the overdetermined assumption $L < M$ is more likely to be violated when longer data is used. We shall experimentally investigate the impact of this data duration in Section 5.3.

2.5 Summary

Both diffuse noise suppression and DOA estimation in diffuse noise boil down to the same problem of denoising the observed covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{xx}}$. Here, denoising at different difficulty levels is required for these tasks.

The state-of-the-art Wiener post-filtering for diffuse noise suppression necessitates an accurate estimation of the signal power spectrogram. The signal power spectrogram $\phi_{ss}(\tau,\omega)$ cannot be observed, but is linked to the observed covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)$ as follows:

$$\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega) = \phi_{ss}(\tau,\omega)\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega) + \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega). \tag{2.84}$$

When the steering vector $\boldsymbol{h}(\omega)$ is known, we only need to estimate the scalar coefficient $\phi_{ss}(\tau,\omega)$ of the known matrix $\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)$. Therefore, it suffices to eliminate *some* entries or components of $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega)$.

Regarding DOA estimation, state-of-the-art MUSIC requires the accurate identification of the null space of the signal covariance matrix. If noise is spatially white, meaning that the noise covariance matrix is a scalar matrix of the form $\sigma^2\boldsymbol{I}$, the eigenspace corresponding to the smallest eigenvalue of the observed covariance matrix coincides with the null space. However, spatially correlated diffuse noise makes it difficult to identify the null space from the observed data. We tackle this identification problem by estimating the signal covariance matrix, so that we can obtain the desired null space directly through the eigenanalysis of the estimated matrix. The signal covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{cc}}(\tau,\omega)$ is related to the observed

covariance matrix $\boldsymbol{\Phi_{xx}}(\tau, \omega)$ as follows:

$$\boldsymbol{\Phi_{xx}}(\tau, \omega) = \boldsymbol{\Phi_{cc}}(\tau, \omega) + \boldsymbol{\Phi_{vv}}(\tau, \omega). \tag{2.85}$$

Since we need the recovery of the *whole* signal covariance matrix $\boldsymbol{\Phi_{cc}}(\tau, \omega)$ in this case, we need to eliminate the whole matrix $\boldsymbol{\Phi_{vv}}(\tau, \omega)$, not only some entries or components.

Chapter 3

Unified Modeling of Noise Covariance in Matrix Linear Space

This chapter has partly been published in [12, 13].

As pointed out in Chapter 2, both diffuse noise suppression and DOA estimation in diffuse noise boil down to denoising of the observed covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{xx}}$. This chapter introduces a unified framework for modeling the noise covariance matrix for covariance matrix denoising.

The rest of this chapter is organized as follows. In Section 3.1, we introduce the unified model of the noise covariance matrix. In Section 3.2, we show that some conventional models of diffuse noise are specific cases of the proposed unified model. Section 3.3 introduces the real-valued noise covariance model. In Section 3.4, we assess the validity of different noise models with real-world noise data.

3.1 Unified framework for modeling noise in matrix linear space

Before formally introducing the unified noise model, let us illustrate the motivation behind this, taking the spatially uncorrelated noise model, which assumes that the noise at different microphones are uncorrelated to each other. In the 3-microphone case, the noise covariance matrix is modeled as

$$\boldsymbol{\Phi}_{\boldsymbol{vv}} = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{bmatrix}, \tag{3.1}$$

where $\alpha \geq 0$, $\beta \geq 0$, and $\gamma \geq 0$. This can be decomposed as

$$\boldsymbol{\Phi}_{\boldsymbol{vv}} = \alpha \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \beta \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.2}$$

Therefore, $\boldsymbol{\Phi}_{\boldsymbol{vv}}$ belongs to a matrix linear subspace $\mathcal{V}$ spanned by the following three matrices:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.3}$$

embedded in the space $\mathcal{H}$ spanned by the 3×3 Hermitian matrices:

$$\mathcal{H} \triangleq \{\boldsymbol{A} \in \mathbb{C}^{3\times 3} | \boldsymbol{A}^{\mathsf{H}} = \boldsymbol{A}\}. \tag{3.4}$$

Note here that, as $\boldsymbol{\Phi}_{\boldsymbol{vv}}$ is Hermitian by definition, we restrict $\mathcal{H}$ to the Hermitian matrices. Indeed, $\boldsymbol{\Phi}_{\boldsymbol{vv}}$ does not span the whole subspace $\mathcal{V}$ but only the positive semidefinite matrices. Nevertheless, this formalism leads to efficient algorithms as will be shown in Chapters 4 to 6.

Formally, our general noise model is that the noise covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau, \omega)$ belongs to a subspace $\mathcal{V}(\omega)$ of the linear space

$$\mathcal{H} \triangleq \{\boldsymbol{A} \in \mathbb{C}^{M\times M} | \boldsymbol{A}^{\mathsf{H}} = \boldsymbol{A}\} \tag{3.5}$$

over $\mathbb{R}$. Note that $\mathcal{H}$ does not form a linear space on $\mathbb{C}$, because it is not closed under the multiplication by a complex number. Note that we allow the noise subspace $\mathcal{V}$ to depend on ω in general. $\mathcal{H}$ is endowed with the inner product

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle \triangleq \sum_{m=1}^{M} \sum_{n=1}^{M} a_{mn} b_{mn}^{*} \tag{3.6}$$

$$= \mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^{\mathsf{H}}) \tag{3.7}$$

$$= \mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) \tag{3.8}$$

and the Frobenius norm

$$\|\boldsymbol{A}\|_{\mathrm{F}} \triangleq \sqrt{\langle \boldsymbol{A}, \boldsymbol{A} \rangle}. \tag{3.9}$$

The subspace $\mathcal{V}(\omega)$ can be characterized either by a set of basis vectors or by the orthogonal projection operator onto $\mathcal{V}(\omega)$. We denote the orthogonal projection operators onto $\mathcal{V}(\omega)$

and its orthogonal complement $\mathcal{V}^{\perp}(\omega)$ by $\mathcal{P}_{\omega}$ and $\mathcal{P}_{\omega}^{\perp}$ respectively, and their general forms are given by

$$\mathcal{P}_{\omega}[\boldsymbol{A}] \triangleq \sum_{i=1}^{P} \langle \boldsymbol{A}, \boldsymbol{Q}_i(\omega) \rangle \boldsymbol{Q}_i(\omega), \tag{3.10}$$

$$\mathcal{P}_{\omega}^{\perp}[\boldsymbol{A}] \triangleq \sum_{i=P+1}^{M^2} \langle \boldsymbol{A}, \boldsymbol{Q}_i(\omega) \rangle \boldsymbol{Q}_i(\omega), \tag{3.11}$$

where $\{\boldsymbol{Q}_i(\omega)\}_{i=1}^{P}$ and $\{\boldsymbol{Q}_i(\omega)\}_{i=P+1}^{M^2}$ denote an orthonormal basis of $\mathcal{V}(\omega)$ and $\mathcal{V}^{\perp}(\omega)$ respectively, where $P \triangleq \dim\mathcal{V}(\omega)$. Noting that $\mathcal{H}$ is a linear space on $\mathbb{R}$, not $\mathbb{C}$, we see that its dimension is $\dim\mathcal{H} = M^2$. Explicit forms of these projectors for specific noise models are given in Sections 3.2 and 3.3.

This general modeling has several benefits. First, it highlights the theoretical connections between the previous noise models to be presented in Section 3.2 and the proposed model to be presented in Section 3.3. Second, as shown in Chapters 4 to 6, it enables the design of new general algorithms applicable to all specific noise models, instead of multiple specific algorithms each applicable to a single model. Third, it facilitates the design of new noise models by restricting the search space for these models; instead of searching for arbitrary *e.g.* nonlinear models, we restrict ourselves to linear subspace models.

3.2 New interpretation of conventional noise models as subspaces

3.2.1 Spatially uncorrelated noise model

Zelinski [27] assumed that noise is uncorrelated between microphones and proposed a post-filter design based on this assumption as described in Section 2.3.2. The noise covariance matrix is diagonal in this case, and belongs to the M-dimensional subspace $\mathcal{V}$ of $\mathcal{H}$ spanned by

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{3.12}$$

The orthogonal projection operators $\mathcal{P}$ and $\mathcal{P}^{\perp}$ are given by

$$\mathcal{P}[\boldsymbol{A}] = \mathcal{D}(\boldsymbol{A}), \tag{3.13}$$

$$\mathcal{P}^{\perp}[\boldsymbol{A}] = \mathcal{O}(\boldsymbol{A}), \tag{3.14}$$

where $\mathcal{D}(\cdot)$ is the operation of replacing the off-diagonal entries by zeros, and $\mathcal{O}(\cdot)$ that of replacing the diagonal entries by zeros. Indeed, any $\boldsymbol{A} \in \mathcal{H}$ can be uniquely expressed as the sum of a component belonging to $\mathcal{V}$ and one belonging to $\mathcal{V}^{\perp}$ as follows:

$$\boldsymbol{A} = \underbrace{\mathcal{D}(\boldsymbol{A})}_{\in \mathcal{V}} + \underbrace{\mathcal{O}(\boldsymbol{A})}_{\in \mathcal{V}^{\perp}}. \tag{3.15}$$

3.2.2 Fixed noise coherence model

McCowan *et al.* [5] assumed that the noise coherence matrix is fixed, and presented a post-filter design based on this model as described in Section 2.3.2. This model corresponds to the subspace

$$\mathcal{V}(\omega) \triangleq \{k\boldsymbol{\Gamma}(\omega) | k \in \mathbb{R}\}, \tag{3.16}$$

where $\boldsymbol{\Gamma}(\omega)$ denotes the noise coherence matrix.

The orthogonal projection operators are given by

$$\mathcal{P}_{\omega}[\boldsymbol{A}] = \frac{\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Gamma}(\omega))}{\mathrm{tr}(\boldsymbol{\Gamma}^2(\omega))}\boldsymbol{\Gamma}(\omega), \tag{3.17}$$

$$\mathcal{P}_{\omega}^{\perp}[\boldsymbol{A}] = \boldsymbol{A} - \frac{\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Gamma}(\omega))}{\mathrm{tr}(\boldsymbol{\Gamma}^2(\omega))}\boldsymbol{\Gamma}(\omega). \tag{3.18}$$

3.2.3 Blind noise decorrelation model

Shimizu *et al.* [41, 42, 43] assumed that $\boldsymbol{\Phi}_{\boldsymbol{vv}}$ is diagonalized by a known constant unitary matrix, and proposed a method for estimating the target power spectrogram based on this model as described in Section 2.3.3. The assumption implies that $\boldsymbol{\Phi}_{\boldsymbol{vv}}$ is expressed as

$$\boldsymbol{\Phi}_{\boldsymbol{vv}} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{\mathsf{H}} \tag{3.19}$$

for some unknown diagonal matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{M \times M}$ and some known unitary diagonalization matrix $\boldsymbol{P} \in \mathbb{C}^{M \times M}$. This equation can be rewritten as

$$\boldsymbol{\Phi}_{\boldsymbol{vv}} = \sum_{m=1}^{M} \lambda_m \boldsymbol{p}_m \boldsymbol{p}_m^{\mathsf{H}} \tag{3.20}$$

with λ_m denoting m-th diagonal entry of $\boldsymbol{\Lambda}$ and $\boldsymbol{p}_m$ the m-th column of $\boldsymbol{P}$. This implies that $\boldsymbol{\Phi}_{\boldsymbol{vv}}$ belongs to

$$\mathcal{V} = \text{span}\{\boldsymbol{p}_m \boldsymbol{p}_m^{\mathsf{H}}\}_{m=1}^{M}. \tag{3.21}$$

The projectors $\mathcal{P}$ and $\mathcal{P}^{\perp}$ are given by

$$\mathcal{P}[\boldsymbol{A}] = \boldsymbol{P}\mathcal{D}(\boldsymbol{P}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{P})\boldsymbol{P}^{\mathsf{H}}, \tag{3.22}$$

$$\mathcal{P}^{\perp}[\boldsymbol{A}] = \boldsymbol{P}\mathcal{O}(\boldsymbol{P}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{P})\boldsymbol{P}^{\mathsf{H}}, \tag{3.23}$$

because any $\boldsymbol{A} \in \mathcal{H}$ can be uniquely decomposed as

$$\boldsymbol{A} = \underbrace{\boldsymbol{P}\mathcal{D}(\boldsymbol{P}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{P})\boldsymbol{P}^{\mathsf{H}}}_{\in \mathcal{V}} + \underbrace{\boldsymbol{P}\mathcal{O}(\boldsymbol{P}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{P})\boldsymbol{P}^{\mathsf{H}}}_{\in \mathcal{V}^{\perp}}. \tag{3.24}$$

3.3 Real-valued noise covariance model

The BND model in Section 3.2.3 is based on crystal arrays. In contrast, motivated by the proposed general noise modeling framework, we propose a model applicable to arrays with arbitrary geometries [10]. This relaxation widens the application range greatly. For example, this enables us to utilize ready-made microphone arrays, which does not necessarily belong to the crystal array category. Furthermore, it is often the case for consumer products that we can only place microphones in a restricted area. The approach enables to place many microphones in the area for a better performance, which is not the case for crystal arrays.

Instead of utilizing the symmetry of the whole array, we exploit the pair-wise symmetry. The isotropy model in Section 3.2.3 implies

$$\phi_{v_m v_n} = \phi_{v_n v_m}. \tag{3.25}$$

By definition of the cross-spectrum, we have

$$\phi_{v_n v_m} = \phi^*_{v_m v_n}. \tag{3.26}$$

From (3.25) and (3.26), we have $\phi_{v_m v_n} \in \mathbb{R}$. Therefore, $\boldsymbol{\Phi}_{\boldsymbol{vv}}$ belongs to the following $M(M+1)/2$-dimensional subspace $\mathcal{V}$:

$$\mathcal{V} \triangleq \{\boldsymbol{A} \in \mathbb{R}^{M\times M} | \boldsymbol{A}^{\mathsf{T}} = \boldsymbol{A}\} \subset \mathcal{H} = \{\boldsymbol{A} \in \mathbb{C}^{M\times M} | \boldsymbol{A}^{\mathsf{H}} = \boldsymbol{A}\}. \tag{3.27}$$

This is spanned by

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \cdots, \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{3.28}$$

The projectors are given by

$$\mathcal{P}[\boldsymbol{A}] = \Re[\boldsymbol{A}], \tag{3.29}$$

$$\mathcal{P}^{\perp}[\boldsymbol{A}] = \mathrm{j}\Im[\boldsymbol{A}], \tag{3.30}$$

because any $\boldsymbol{A} \in \mathcal{H}$ is decomposed uniquely as

$$\boldsymbol{A} = \underbrace{\Re[\boldsymbol{A}]}_{\in \mathcal{V}} + \underbrace{\mathrm{j}\Im[\boldsymbol{A}]}_{\in \mathcal{V}^{\perp}}. \tag{3.31}$$

This model is more flexible than the spatially uncorrelated noise model and the fixed noise coherence model for spherically/cylindrically isotropic noise. Indeed, these models are real-valued, and thus subspaces of the real-valued noise covariance model.

3.4 Assessment of noise models with real-world noise

Two different aspects are important to predict the performance of a certain model:

- the number of parameters of the model (*i.e.* $\dim\mathcal{V}(\omega)$) compared to the number of observations (*i.e.* $\dim\mathcal{H}$),
- the fit between this model and real-world covariance matrices.

Ideally, for *e.g.* twice as many parameters, we expect the fit to increase a lot. If the fit is only marginally better, the increased number of parameters is likely to result in a poorer performance in a practical blind setting. These two pieces of information together hence enable to predict the outcomes of subsequent experiments to a certain degree.

Another important issue is how much the signal covariance matrix diverge from the noise model. If the signal covariance matrix also lies in $\mathcal{V}$, this model is useless in distinguishing between the signal and the noise.

We conducted an experiment to investigate the fit of real-world noise covariance matrices to the noise models presented in Sections 3.2 and 3.3 relative to the number of parameters of the model, as well as the discrepancy of the signal covariance matrix from the model.

We used the following three data/databases we created:

- Rennes database: a database of noise recorded in Rennes, France, using a uniform linear array with four microphones and the distance between adjacent microphones of 0.086 m. Three noise samples are included, which are recorded in a cafeteria, on a subway train, and in a station square. The duration of each noise sample is one minute.

- Shinjuku database: a database of noise recorded in Shinjuku, Tokyo, using a square array with four microphones and the diameter of 0.05 m. Four noise samples are included, which are recorded in a station building, on a platform, on a subway train, and in a station square. The duration of each noise sample is one minute.

- University data: noise recorded in an experiment room at the University of Tokyo, using an icosahedral array with twelve microphones and the diameter of 0.15 m. The windows of the room were open during the recording. The data duration is 10 s.

Each noise sample was converted into the time-frequency representation $\boldsymbol{v}(\tau, \omega)$ by STFT, and an empirical noise covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\omega)$ was computed by long-term temporal average of $\boldsymbol{v}(\tau, \omega)\boldsymbol{v}^{\mathsf{H}}(\tau, \omega)$ over the whole data duration. We define a measure of the discrepancy between $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\omega)$ and the noise model $\mathcal{V}(\omega)$ as follows. $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\omega)$ can be expressed as the sum of two components: $\mathcal{P}_\omega[\boldsymbol{\Phi}_{\boldsymbol{vv}}](\omega)$ belonging to $\mathcal{V}(\omega)$ and $\mathcal{P}_\omega^\perp[\boldsymbol{\Phi}_{\boldsymbol{vv}}](\omega)$ orthogonal to $\mathcal{V}(\omega)$. Therefore, the discrepancy between $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\omega)$ and $\mathcal{V}(\omega)$ can be evaluated by

$$\epsilon(\omega) \triangleq \frac{\|\mathcal{P}^\perp[\boldsymbol{\Phi}_{\boldsymbol{vv}}](\omega)\|_{\mathrm{F}}}{\|\boldsymbol{\Phi}_{\boldsymbol{vv}}(\omega)\|_{\mathrm{F}}}, \tag{3.32}$$

which is the distance between $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\omega)$ and $\mathcal{V}(\omega)$ normalized by $\|\boldsymbol{\Phi}_{\boldsymbol{vv}}(\omega)\|_{\mathrm{F}}$. Here, the normalization is aimed at removing the dependency of $\epsilon(\omega)$ on the scale of $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\omega)$. We can define the *discrepancy index* for each database and noise model as the arithmetic average of $\epsilon(\omega)$ over the frequency and the noise samples in the database.

Fig. 3.1-3.3 two-dimensionally plots the discrepancy index of each model versus its dimensionality. These correspond to the Rennes database, the Shinjuku database, and the University data, respectively. A model closer to the origin is a good model that is able to fit

Table 3.1 The dimensions of the space $\mathcal{H}$ and each model subspace $\mathcal{V}(\omega)$ as a function of the number of microphones.

number of microphones	$\mathcal{H}$	uncor	coh	BND	real
M	M^2	M	1	M	$M(M+1)/2$
4	16	4	1	4	10
12	144	12	1	12	78

the real-world noise better with a smaller number of parameters. Note that the dimensions of the space $\mathcal{H}$ and the model subspaces $\mathcal{V}(\omega)$ are functions of M as shown in Table 3.1.

Having a significantly higher dimension (10 for the Rennes and Shinjuku databases; 78 for the University data) than the other models, the real-valued noise covariance model gave the smallest discrepancy index of 0.16 to 0.27. This was smallest among all models, which means the best fit to real-world noise. However, the high dimensionality compared to the number of observations can lead to the overfitting to the data. The ratio $(M+1)/(2M)$ of the dimension and the number of observation is larger than 0.5, and approaches 0.5 when $M \to \infty$.

In comparison, the BND model reduced the dimensionality significantly compared to the real-valued noise covariance model with only a small increase in the discrepancy index. For the Shinjuku database (resp. the University data), it has only 0.4 (resp. 0.13) time as high a dimension as the real-valued noise covariance model, with an increase in the discrepancy index of 0.06 (resp. 0.17). Furthermore, it gave a lower discrepancy index compared to the spatially uncorrelated noise model and the fixed noise coherence model. Note that the BND model was excluded from Fig. 3.1, because it is inapplicable to the linear array employed for recording the Rennes database.

The discrepancy index of the spatially uncorrelated noise model was the largest among all models except for the University data, for which the fixed noise coherence model failed. It was larger than that of the BND model having the same dimensionality by 0.31-0.35 except for the Rennes database for which the BND model was inapplicable. The poor fit is due to the high spatial correlation of real-world noise for small microphone spacing. Indeed, this model worked better for the Rennes database than for the other databases, for which microphone distances were larger and thus the noise correlation was lower.

The fixed noise coherence model has dimension 1 independent of the number of microphones, but nevertheless it fitted the Rennes and the Shinjuku databases reasonably well.

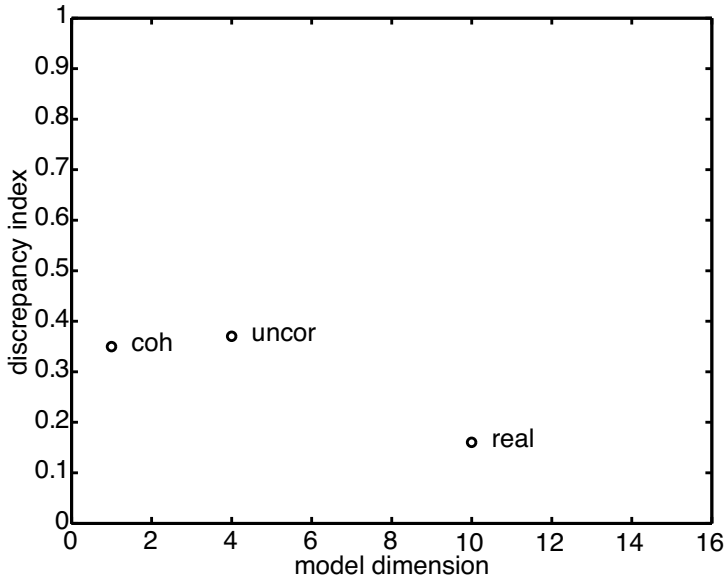


Figure 3.1 The discrepancy index vs. the model dimension for each noise model for the Rennes database.

For these databases, it outperformed the spatially uncorrelated noise model having M times as high a dimension. For the Rennes database, the reduction in the discrepancy index was only 0.02, because the spatially uncorrelated noise model worked relatively well due to the large microphone distances. In contrast, it reduced the discrepancy index compared to the spatially uncorrelated noise model by as much as 0.21 for the Shinjuku database by taking into account the noise correlation. However, for the University data recorded using an array mounted on a rigid mount, it gave the highest discrepancy index of 0.82. This can be interpreted as a result of the diffraction due to the mount.

We also evaluated the discrepancy index between the signal covariance matrix and the noise models. Since this is expected to depend on the target DOA, we need to evaluate it for various DOAs. To facilitate this, we utilized a theoretical signal covariance matrix under the planewave propagation model. The steering vector of the planewave from the zenith angle θ and the azimuth ξ is given by (2.28) with the delay δ_m given as a function of θ and ξ by

$$\delta_m(\theta, \xi) = -\frac{\boldsymbol{p}^{\mathsf{T}}(\theta, \xi)(\boldsymbol{r}_m - \boldsymbol{r}_1)}{c}, \tag{3.33}$$

where $\boldsymbol{r}_m$ denotes the coordinates of the m-th microphone, and $\boldsymbol{p}(\theta, \xi)$ is the unit DOA vector

$$\boldsymbol{p}(\theta, \xi) \triangleq \begin{bmatrix} \sin\theta\cos\xi & \sin\theta\sin\xi & \cos\theta \end{bmatrix}^{\mathsf{T}}. \tag{3.34}$$

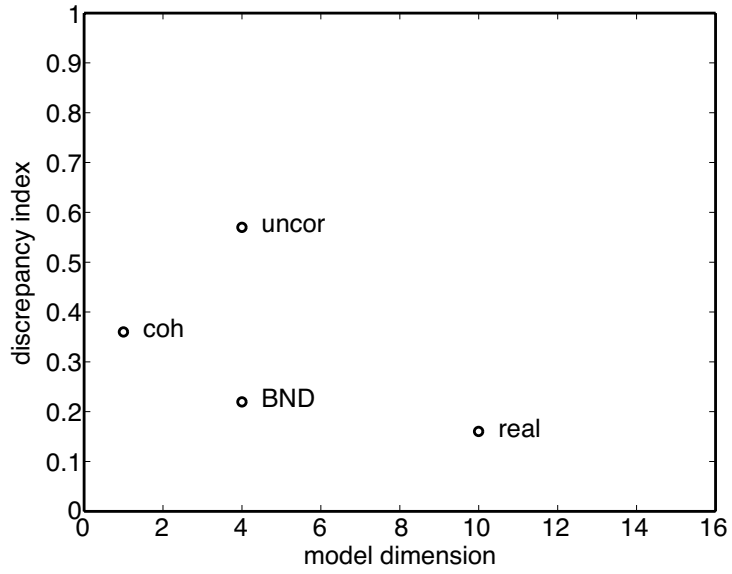


Figure 3.2 The discrepancy index vs. the model dimension for each noise model for the Shinjuku database.

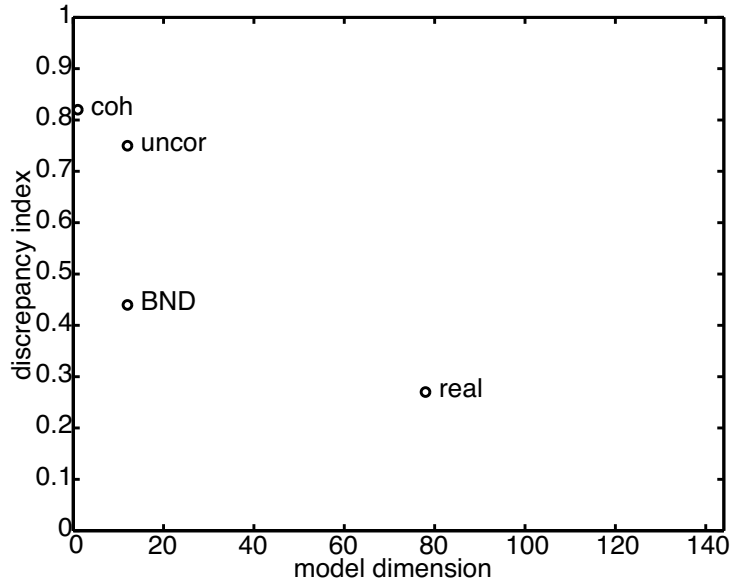


Figure 3.3 The discrepancy index vs. the model dimension for each noise model for the University data.

Using this steering vector, denoted by $\boldsymbol{h}(\omega;\theta,\xi)$, the discrepancy index is calculated as follows:

$$\frac{1}{K}\sum_{\omega}\frac{\|\mathcal{P}^{\perp}[\boldsymbol{h}(\omega;\theta,\xi)\boldsymbol{h}^{\mathsf{H}}(\omega;\theta,\xi)]\|_{\mathrm{F}}}{\|\boldsymbol{h}(\omega;\theta,\xi)\boldsymbol{h}^{\mathsf{H}}(\omega;\theta,\xi)\|_{\mathrm{F}}}, \tag{3.35}$$

where K denotes the number of frequency bins, and the scalar $\phi_{ss}(\tau, \omega)$ in the numerator and the denominator has been cancelled. Ideally, for the icosahedral array, the effects of the diffraction by the rigid mount should be taken into account, but we used the simple planewave model here for a preliminary investigation.

We calculated the discrepancy index as a function of θ and ξ defined by (3.35) for each noise model and each array configuration used for recording the databases (Fig. 3.4 to 3.14). Figures 3.4 to 3.6 are the two-dimensional plot of the absolute value of the discrepancy index as a function of θ and ξ for the uniform linear array configuration used for recording of the Rennes database for the spatially uncorrelated noise model, the fixed noise coherence model, and the real-valued noise covariance model, respectively. Figures 3.7 to 3.10 (resp. Figs. 3.11 to 3.14) are the results for the square (resp. icosahedral) array for the spatially uncorrelated noise model, the fixed noise coherence model, the BND model, and the real-valued noise covariance model, respectively.

As we see from Figs. 3.4, 3.7, and 3.11, the discrepancy index for the spatially uncorrelated noise model was independent of θ and ξ for all array configuration. Indeed, as each entry of $\boldsymbol{h}(\omega; \theta, \xi)\boldsymbol{h}^{\mathsf{H}}(\omega; \theta, \xi)$ has a unit magnitude,

$$\frac{\|\mathcal{P}^{\perp}[\boldsymbol{h}(\omega; \theta, \xi)\boldsymbol{h}^{\mathsf{H}}(\omega; \theta, \xi)]\|_{\mathrm{F}}}{\|\boldsymbol{h}(\omega; \theta, \xi)\boldsymbol{h}^{\mathsf{H}}(\omega; \theta, \xi)\|_{\mathrm{F}}} = \frac{\|\mathcal{O}[\boldsymbol{h}(\omega; \theta, \xi)\boldsymbol{h}^{\mathsf{H}}(\omega; \theta, \xi)]\|_{\mathrm{F}}}{\|\boldsymbol{h}(\omega; \theta, \xi)\boldsymbol{h}^{\mathsf{H}}(\omega; \theta, \xi)\|_{\mathrm{F}}} \tag{3.36}$$

$$= \frac{\sqrt{M^2 - M}}{\sqrt{M^2}} \tag{3.37}$$

$$= \sqrt{\frac{M-1}{M}}. \tag{3.38}$$

This becomes 0.87 for $M = 4$, and 0.96 for $M = 12$.

As we see from Figs. 3.5, 3.8, and 3.12, the discrepancy index for the fixed noise coherence model was high for all θ and ξ. The minimum and the maximum values among the evaluated points were 0.75 and 0.84 for the uniform linear array, 0.66 and 0.75 for the square array, and 0.85 and 0.93 for the icosahedral array.

As seen from Figs. 3.9 and 3.13, the discrepancy index for the BND model was much more dependent on θ and ξ. For the square array, it became zero at the zenith angles of 0° and 180° corresponding to the orthogonal DOAs to the array plane. Since the signal wavefront arrives at the microphones in phase in these cases, the signal cannot be distinguished from diffuse noise. In contrast, for the three-dimensional icosahedral array, the discrepancy index was high for all θ and ξ. The minimum and the maximum values among the evaluated points were 0 and 0.72 for the square array, and 0.71 and 0.87 for the icosahedral array.

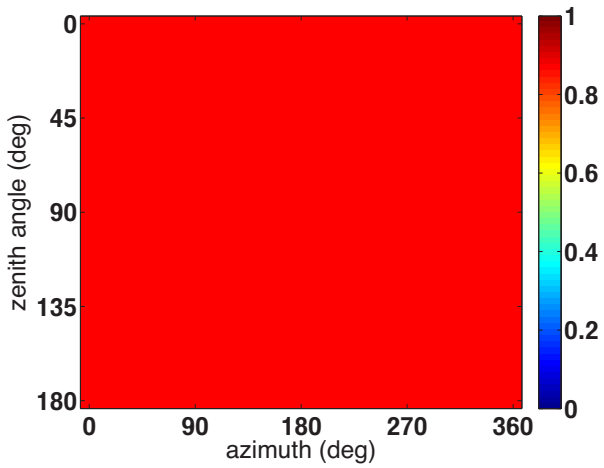


Figure 3.4 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (uniform linear array; spatially uncorrelated noise model).

Finally, as seen from Figs. 3.10 and 3.14, the discrepancy index for the real-valued noise covariance model was globally lower than that for the BND model. Furthermore, for the uniform linear array, it became zero at the zenith angles of 0° and 180° and the azimuths of 90° and 270° corresponding to the orthogonal DOAs to the array axis. Also, for the square array, it became zero at the zenith angles of 0° and 180° corresponding to the orthogonal DOAs to the array plane. For the three-dimensional icosahedral array, this effect was absent. The minimum and the maximum values among the evaluated points were 0 and 0.63 for the linear array, 0 and 0.61 for the square array, and 0.54 and 0.67 for the icosahedral array.

Figures 3.15-3.19 show the noise coherence matrices before and after the BND at the frequency of 1 kHz. The entry in the m-th row and the n-th column of this matrix is the correlation coefficient of the noise signals at the m-th and n-th microphones. Figures 3.15-3.19 correspond to noise samples recorded in a station building, on a platform, on a subway train, and in a station square (all in Shinjuku), and that recorded in an experiment room at the University, respectively. We see that the noise correlation is quite high before BND, especially for the Shinjuku database, for which we used a very small microphone array with a diameter of 0.05 m. Therefore, the spatially uncorrelated noise model is far from accurate. The proportion of the magnitude of the diagonal entries increased significantly through the BND, verifying the BND model.

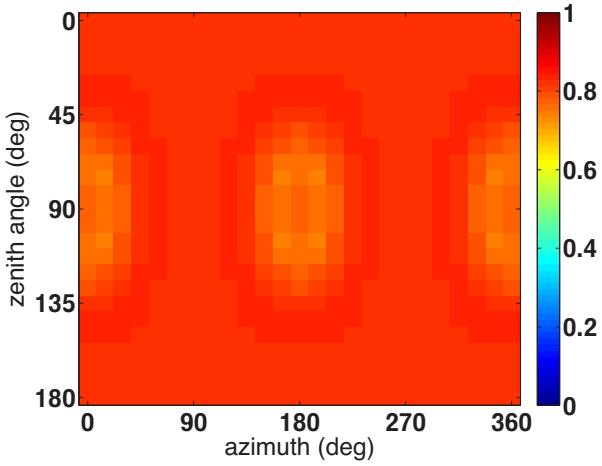


Figure 3.5 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (uniform linear array; fixed noise coherence model).

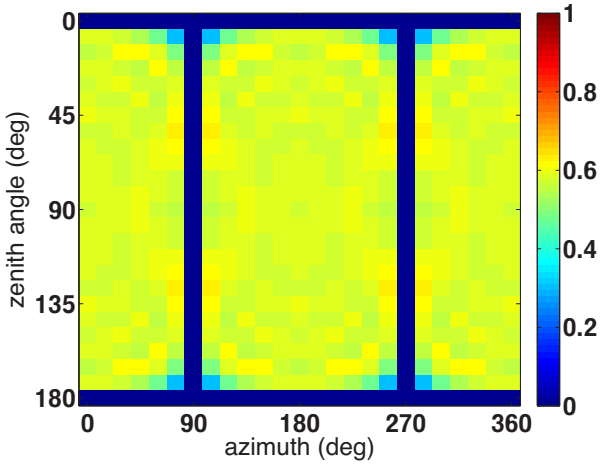


Figure 3.6 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (uniform linear array; real-valued noise covariance model).

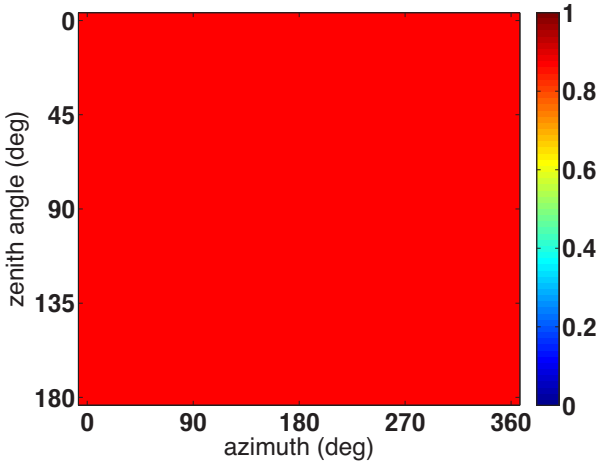


Figure 3.7 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (square array; spatially uncorrelated noise model).

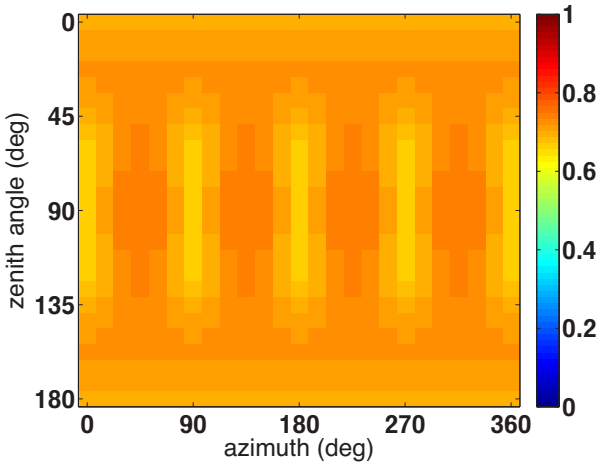


Figure 3.8 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (square array; fixed noise coherence model).

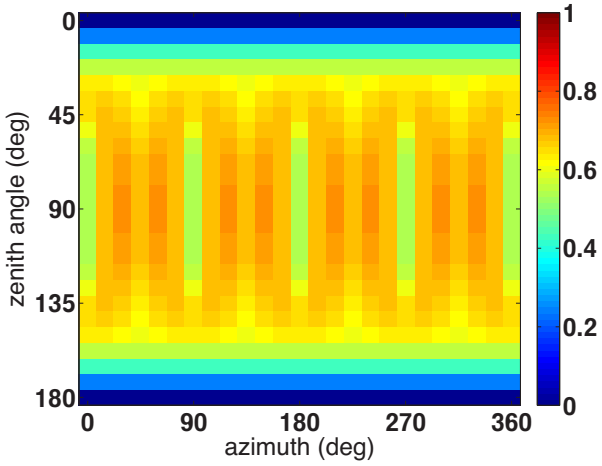


Figure 3.9 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (square array; BND model).

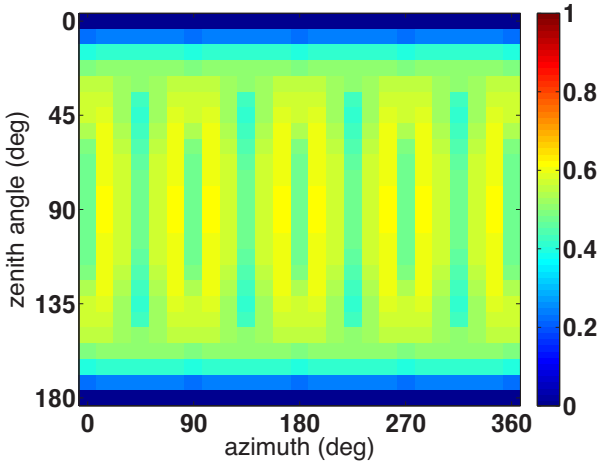


Figure 3.10 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (square array; real-valued noise covariance model).

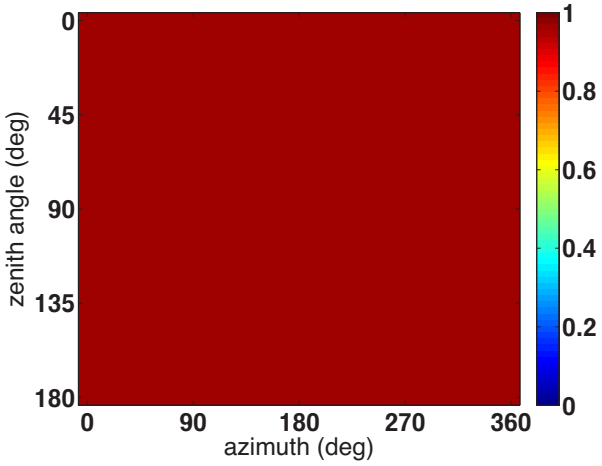


Figure 3.11 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (icosahedral array; spatially uncorrelated noise model).

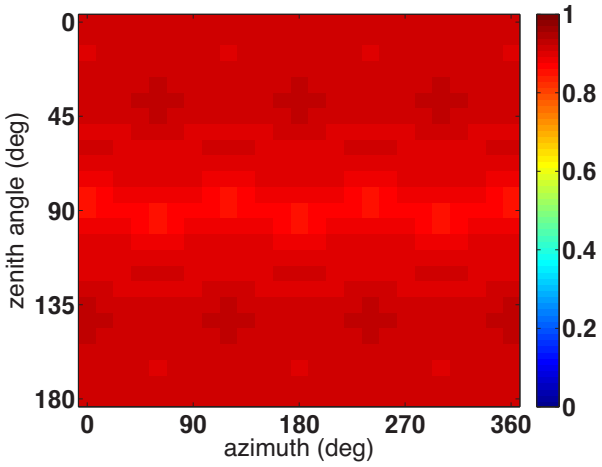


Figure 3.12 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (icosahedral array; fixed noise coherence model).

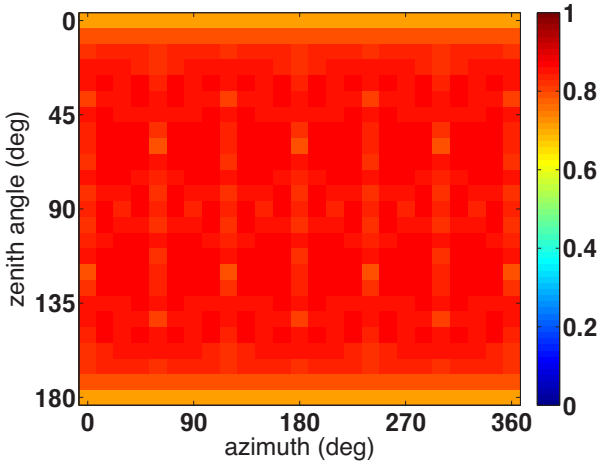


Figure 3.13 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (icosahedral array; BND model).

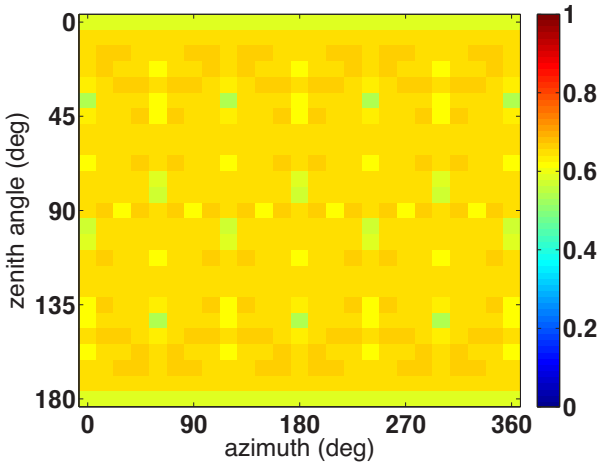
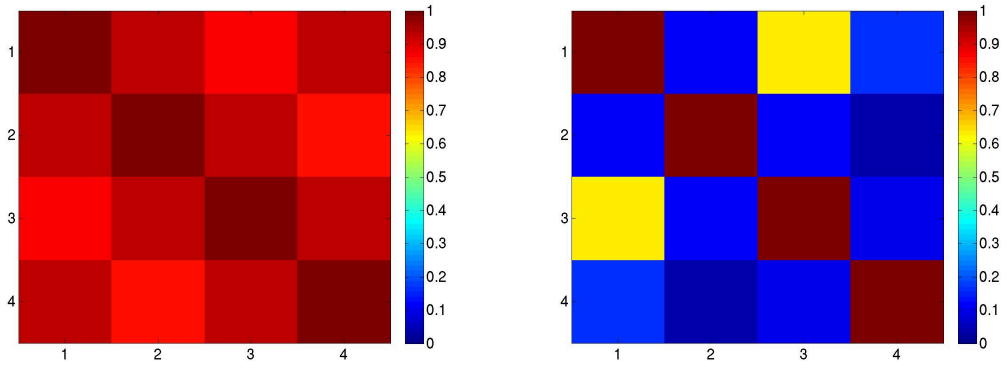
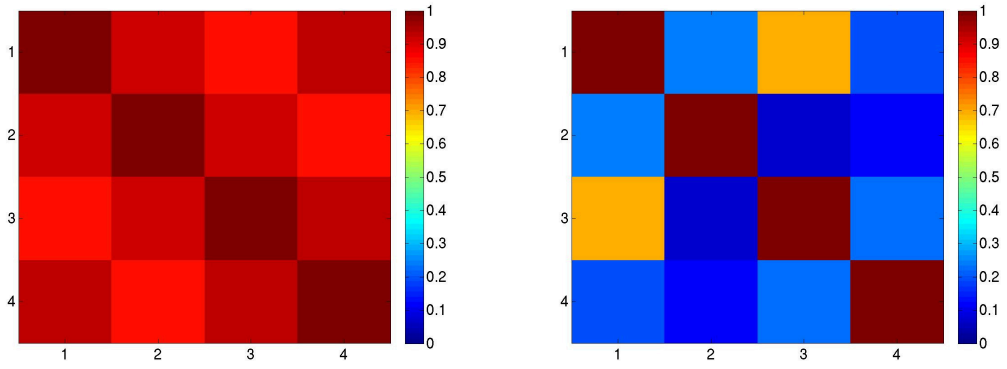


Figure 3.14 The discrepancy index between the signal covariance matrix and the noise model as a function of the azimuth and the zenith angle of the target DOA (icosahedral array; real-valued noise covariance model).



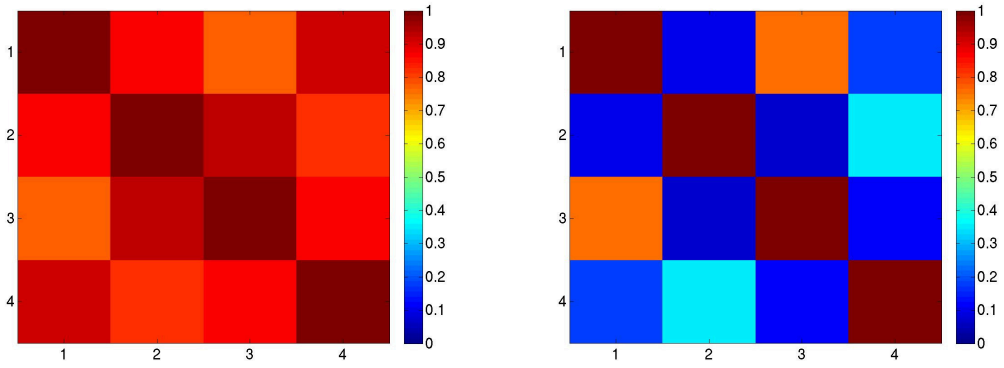
(a) (b)

Figure 3.15 The noise coherence matrices (a) before BND and (b) after BND (noise environment: station building in Shinjuku).



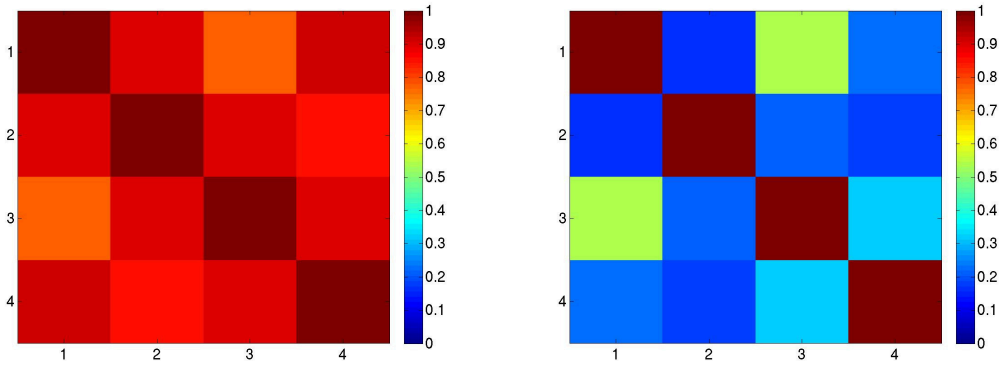
(a) (b)

Figure 3.16 The noise coherence matrices (a) before BND and (b) after BND (noise environment: platform in Shinjuku).



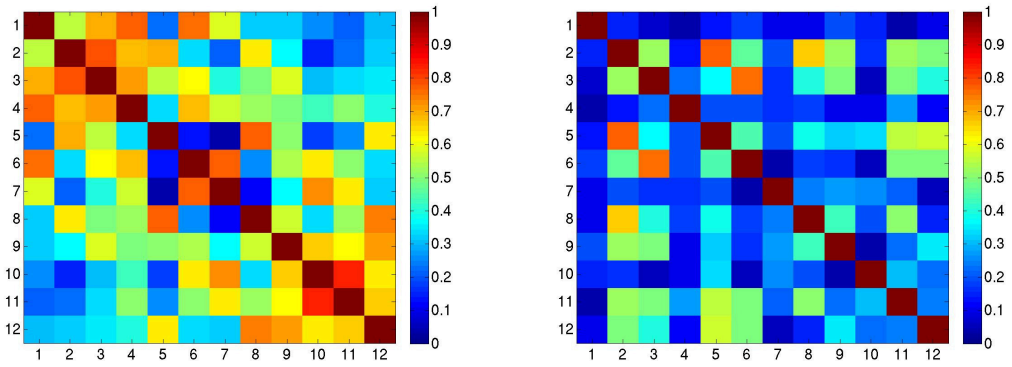
(a) (b)

Figure 3.17 The noise coherence matrices (a) before BND and (b) after BND (noise environment: subway train in Shinjuku).



(a) (b)

Figure 3.18 The noise coherence matrices (a) before BND and (b) after BND (noise environment: station square in Shinjuku).



(a) (b)

Figure 3.19 The noise coherence matrices (a) before BND and (b) after BND (noise environment: experiment room at the University of Tokyo).

Chapter 4

Diffuse Noise Suppression for Target Signal from Known Direction

This chapter has partly been published in [8, 10, 11, 14, 16, 17, 18, 19, 20].

We have seen in Chapter 2 that noise suppression can be addressed by MVDR beamforming followed by Wiener post-filtering as in (2.40). In this framework, it is essential to accurately estimate the power spectrogram $\phi_{ss}(\tau, \omega)$ and the steering vector $\boldsymbol{h}(\omega)$ from the observed signals $\boldsymbol{x}(\tau, \omega)$.

In this chapter, we focus on the estimation of $\phi_{ss}(\tau, \omega)$ given $\boldsymbol{h}(\omega)$. In practice, $\boldsymbol{h}(\omega)$ can be calculated using the planewave propagation model (2.28) given the target DOA. We describe a unified estimation framework applicable to the general noise model presented in Chapter 3. We also derive the explicit forms of the estimator for specific noise models.

The rest of this chapter is organized as follows. Section 4.1 presents the unified framework for the estimation of $\phi_{ss}(\tau, \omega)$. In Section 4.2, we derive the explicit form of the estimator for each specific noise model. In Section 4.3, we assess the noise suppression performance of the proposed approach through simulation with real-world noise.

4.1 Unified framework for diffuse noise suppression based on orthogonal projection in matrix linear space

The estimation of $\phi_{ss}(\tau, \omega)$ in this chapter is based on orthogonal projection of the observed covariance matrix in a matrix linear space. The unified noise model in Chapter 3 assumes that the noise covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau, \omega)$ belongs to a subspace $\mathcal{V}(\omega)$ in the matrix linear

space $\mathcal{H}$. This implies that the orthogonal projection operation $\mathcal{P}_\omega^\perp$ onto the orthogonal complement $\mathcal{V}^\perp(\omega)$ eliminates $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega)$:

$$\mathcal{P}_\omega^\perp[\boldsymbol{\Phi}_{\boldsymbol{vv}}](\tau,\omega) = \boldsymbol{O}. \tag{4.1}$$

Therefore, we can obtain a noise-free component by applying $\mathcal{P}_\omega^\perp$ to the observed covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)$. Specifically, applying $\mathcal{P}_\omega^\perp$ to the both sides of (2.23), we have

$$\mathcal{P}_\omega^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau,\omega) = \phi_{ss}(\tau,\omega)\mathcal{P}_\omega^\perp[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]. \tag{4.2}$$

Here, $\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)$ can be calculated from the data and $\boldsymbol{h}(\omega)$ is assumed to be known. Therefore, we can obtain $\phi_{ss}(\tau,\omega)$ using (4.2). In practice, (4.2) contains some errors because of the misestimation of $\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)$ due to the limited data and the imperfection of the noise model $\mathcal{V}(\omega)$. For simplicity, we estimate $\phi_{ss}(\tau,\omega)$ through the Least-Squares (LS) fitting. Specifically, we minimize the following squared error of (4.2) with respect to $\phi_{ss}(\tau,\omega)$:

$$J \triangleq \sum_\tau \|\mathcal{P}_\omega^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau,\omega) - \phi_{ss}(\tau,\omega)\mathcal{P}_\omega^\perp[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\|_{\mathrm{F}}^2. \tag{4.3}$$

From the orthogonality principle, the optimum solution $\hat{\phi}_{ss}(\tau,\omega)$ is given by

$$\hat{\phi}_{ss}(\tau,\omega) = \frac{\langle \mathcal{P}_\omega^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau,\omega), \mathcal{P}_\omega^\perp[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\rangle}{\|\mathcal{P}_\omega^\perp[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\|_{\mathrm{F}}^2}. \tag{4.4}$$

On the other hand, the estimate $\hat{\phi}_{yy}(\tau,\omega)$ of the beamformer output power $\phi_{yy}(\tau,\omega)$ in the denominator of the Wiener post-filter is calculated by Zelinski's estimator

$$\hat{\phi}_{yy}^{\mathrm{Zel}}(\tau,\omega) \triangleq \frac{1}{M}\sum_{m=1}^{M} \phi_{x_m x_m}(\tau,\omega). \tag{4.5}$$

We observed through a preliminary experiment that this estimator resulted in a higher noise suppression performance compared to the direct calculation by averaging $|y(\tau,\omega)|^2$ temporally over several adjacent frames. Consequently, the post-filter is designed as follows:

$$\hat{p}(\tau,\omega) \triangleq \frac{\hat{\phi}_{ss}(\tau,\omega)}{\hat{\phi}_{yy}^{\mathrm{Zel}}(\tau,\omega)}. \tag{4.6}$$

Since the Wiener post-filter $p(\tau,\omega)$ lies in the range $0 \le p(\tau,\omega) \le 1$ in theory, we perform the following simple post-processing:

$$\hat{p}(\tau,\omega) \leftarrow \begin{cases} 0, & \text{if } \hat{p}(\tau,\omega) < 0, \\ 1, & \text{if } \hat{p}(\tau,\omega) > 1. \end{cases} \tag{4.7}$$

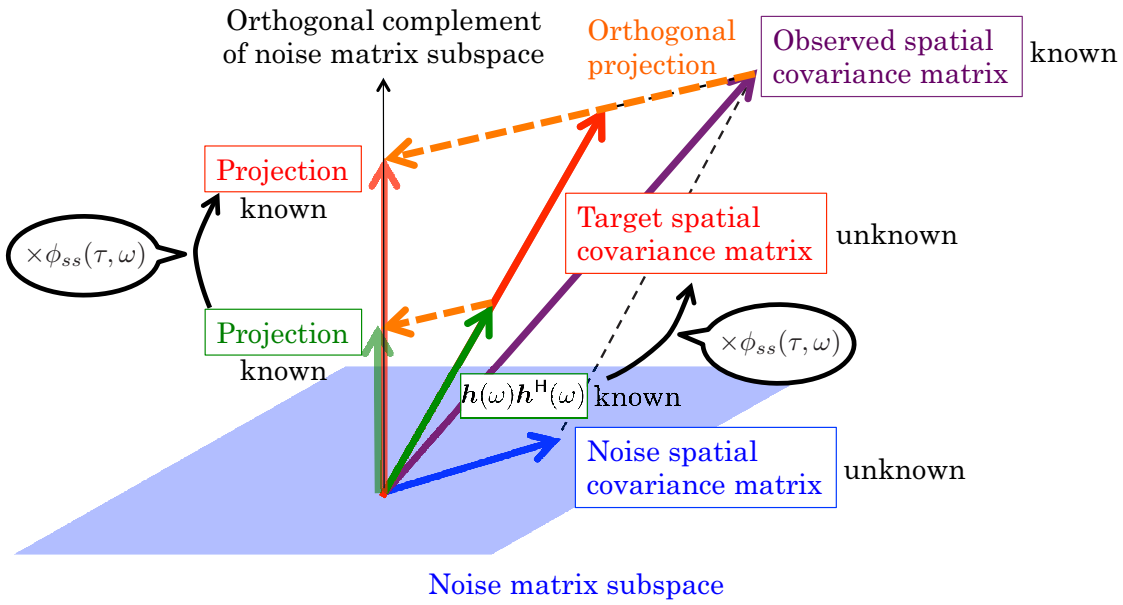


Figure 4.1 Illustration of target power spectrogram estimation based on orthogonal projection in the matrix linear space.

4.2 Application to specific noise models

In Section 4.1, we have seen that the general form of the estimator of $\phi_{ss}(\tau,\omega)$ is given by (4.4). In this section, we derive the specific estimator for each noise model presented in Chapter 3.

4.2.1 Application to the spatially uncorrelated noise model

The estimator for the uncorrelated noise model is obtained by substituting (3.14) to (4.4) as follows:

$$\hat{\phi}_{ss}^{\mathrm{uncor}}(\tau,\omega)=\frac{\langle\mathcal{O}[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau,\omega),\mathcal{O}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\rangle}{\|\mathcal{O}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\|_{\mathrm{F}}^{2}}. \tag{4.8}$$

This can be rewritten more explicitly as follows:

$$\hat{\phi}_{ss}^{\mathrm{uncor}}(\tau,\omega)=\frac{\sum\limits_{m,n,m\neq n}\phi_{x_mx_n}(\tau,\omega)h_m^*(\omega)h_n(\omega)}{\sum\limits_{m,n,m\neq n}|h_m(\omega)|^2|h_n(\omega)|^2}. \tag{4.9}$$

This can be seen as a generalization of Zelinski's estimator $\hat{\phi}_{ss}^{\mathrm{Zel}}(\tau,\omega)$ in (2.49) to the general steering vector. Indeed, substituting the steering vector (2.28) of a planewave, $\hat{\phi}_{ss}^{\mathrm{uncor}}(\tau,\omega)$ coincides with $\hat{\phi}_{ss}^{\mathrm{Zel}}(\tau,\omega)$.

4.2.2 Application to fixed noise coherence model

The estimator for the fixed noise coherence model is obtained by substituting (3.18) to (4.4) as follows:

$$\hat{\phi}_{ss}^{\mathrm{coh}}(\tau,\omega) = \frac{\left\langle \boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega) - \frac{\mathrm{tr}[\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)\boldsymbol{\Gamma}(\omega)]}{\mathrm{tr}[\boldsymbol{\Gamma}^2(\omega)]}\boldsymbol{\Gamma}(\omega), \boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega) - \frac{\mathrm{tr}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Gamma}(\omega)]}{\mathrm{tr}[\boldsymbol{\Gamma}^2(\omega)]}\boldsymbol{\Gamma}(\omega) \right\rangle}{\left\| \boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega) - \frac{\mathrm{tr}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Gamma}(\omega)]}{\mathrm{tr}[\boldsymbol{\Gamma}^2(\omega)]}\boldsymbol{\Gamma}(\omega) \right\|_{\mathrm{F}}^2}. \tag{4.10}$$

This is different from McCowan's estimator $\hat{\phi}_{ss}^{\mathrm{Mc}}(\tau,\omega)$ in (2.56) based on the same noise model, because the latter is based on a suboptimal estimator instead of LS. We compare the performance of both estimators in the experiment in Section 4.3.

4.2.3 Application to the blind noise decorrelation model

The estimator for the BND model is obtained by substituting (3.23) to (4.4) as follows:

$$\hat{\phi}_{ss}^{\mathrm{BND}}(\tau,\omega) = \frac{\langle \boldsymbol{P}\mathcal{O}[\boldsymbol{P}^{\mathsf{H}}\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)\boldsymbol{P}]\boldsymbol{P}^{\mathsf{H}}, \boldsymbol{P}\mathcal{O}[\boldsymbol{P}^{\mathsf{H}}\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{P}]\boldsymbol{P}^{\mathsf{H}}\rangle}{\|\boldsymbol{P}\mathcal{O}[\boldsymbol{P}^{\mathsf{H}}\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{P}]\boldsymbol{P}^{\mathsf{H}}\|_{\mathrm{F}}^2}. \tag{4.11}$$

Equation (4.11) can be simplified using the following properties of unitary matrices. For a unitary matrix $\boldsymbol{P} \in \mathbb{C}^{M\times M}$ and Hermitian matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{H}$, the following equations hold:

$$\langle \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathsf{H}}, \boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^{\mathsf{H}}\rangle = \langle \boldsymbol{A}, \boldsymbol{B}\rangle, \tag{4.12}$$

$$\|\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathsf{H}}\|_{\mathrm{F}} = \|\boldsymbol{A}\|_{\mathrm{F}}. \tag{4.13}$$

Therefore, (4.11) becomes

$$\hat{\phi}_{ss}^{\mathrm{BND}}(\tau,\omega) = \frac{\langle \mathcal{O}[\boldsymbol{P}^{\mathsf{H}}\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)\boldsymbol{P}], \mathcal{O}[\boldsymbol{P}^{\mathsf{H}}\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{P}]\rangle}{\|\mathcal{O}[\boldsymbol{P}^{\mathsf{H}}\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{P}]\|_{\mathrm{F}}^2}. \tag{4.14}$$

Moreover, defining $\tilde{\boldsymbol{h}}(\omega) \triangleq \boldsymbol{P}^{\mathsf{H}}\boldsymbol{h}(\omega)$ and $\tilde{\boldsymbol{x}}(\tau,\omega) \triangleq \boldsymbol{P}^{\mathsf{H}}\boldsymbol{x}(\tau,\omega)$, we have

$$\hat{\phi}_{ss}^{\mathrm{BND}}(\tau,\omega) = \frac{\langle \mathcal{O}[\boldsymbol{\Phi}_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}}](\tau,\omega), \mathcal{O}[\tilde{\boldsymbol{h}}(\omega)\tilde{\boldsymbol{h}}^{\mathsf{H}}(\omega)]\rangle}{\|\mathcal{O}[\tilde{\boldsymbol{h}}(\omega)\tilde{\boldsymbol{h}}^{\mathsf{H}}(\omega)]\|_{\mathrm{F}}^2}. \tag{4.15}$$

This can be rewritten more explicitly as follows [8]:

$$\hat{\phi}_{ss}^{\mathrm{BND}}(\tau,\omega) = \frac{\sum_{m,n,m\neq n} \phi_{\tilde{x}_m\tilde{x}_n}(\tau,\omega)\tilde{h}_m^*(\omega)\tilde{h}_n(\omega)}{\sum_{m,n,m\neq n} |\tilde{h}_m(\omega)|^2|\tilde{h}_n(\omega)|^2}. \tag{4.16}$$

Note that this is identical to (4.9) except that there are tildes on h_m and x_m. Therefore, (4.16) can be seen as estimation by (4.9) after spatial decorrelation.

4.2.4 Application to the real-valued noise covariance model

The estimator for the real-valued noise covariance model is obtained by substituting (3.30) to (4.4) as follows [10]:

$$\hat{\phi}_{ss}^{\mathrm{real}}(\tau,\omega) = \frac{\langle \mathrm{j}\Im[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau,\omega), \mathrm{j}\Im[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\rangle}{\|\mathrm{j}\Im[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\|_{\mathrm{F}}^2} \tag{4.17}$$

$$= \frac{\sum_{m\neq n} \Im[\phi_{x_mx_n}(\tau,\omega)]\Im[h_m(\omega)h_n^*(\omega)]}{\sum_{m\neq n} \Im[h_m(\omega)h_n^*(\omega)]^2}. \tag{4.18}$$

Note here that the summation excludes the diagonal entries because they are real-valued by definition.

4.3 Performance evaluation with real-world noise

To evaluate the diffuse noise suppression performance of proposed/conventional beamforming and post-filtering approaches, we conducted a simulation with real-world diffuse noise. The rest of this section is organized as follows. In Section 4.3.1, we define performance metrics. In Section 4.3.2, we describe the experimental conditions, and finally in Section 4.3.3, we present the results.

4.3.1 Evaluation metrics

We evaluate the noise suppression performance with the following three objective metrics: the output Signal-to-Noise Ratio (SNR), the Speech-Distortion Index (SDI), and the Noise-Reduction Factor (NRF). This is motivated by the fact that there is a trade-off between the

amount of noise reduction and that of target distortion in general: the more the former, the more the latter as well. Therefore, we evaluate the overall performance with the output SNR, and target distortion and noise reduction with the SDI and the NRF, separately.

We decompose the estimated target signal (the output of a noise suppression system) into the sum of two components: one proportional to the target signal and one orthogonal to it. Note that we can access to the target signal for evaluation, because we artificially mixed it with the noise component in this simulation. More specifically, let us denote by

$$\boldsymbol{s}' \triangleq \begin{bmatrix} s[0] & s[1] & \cdots & s[K-1] \end{bmatrix}^{\mathsf{T}}, \tag{4.19}$$

$$\hat{\boldsymbol{s}}' \triangleq \begin{bmatrix} \hat{s}[0] & \hat{s}[1] & \cdots & \hat{s}[K-1] \end{bmatrix}^{\mathsf{T}} \tag{4.20}$$

the vectors comprised of the samples in the target signal $s[k]$ and its estimate $\hat{s}[k]$, respectively. We decompose the estimated signal $\hat{\boldsymbol{s}}'$ into the component $\hat{\boldsymbol{s}}'_{\parallel}$ parallel to $\boldsymbol{s}'$ and the component $\hat{\boldsymbol{s}}'_{\perp}$ perpendicular to it as follows:

$$\hat{\boldsymbol{s}}' = \hat{\boldsymbol{s}}'_{\parallel} + \hat{\boldsymbol{s}}'_{\perp}, \tag{4.21}$$

where

$$\hat{\boldsymbol{s}}'_{\parallel} \triangleq \frac{\hat{\boldsymbol{s}}'^{\mathsf{T}}\boldsymbol{s}'}{\|\boldsymbol{s}'\|_2^2}\boldsymbol{s}', \tag{4.22}$$

$$\hat{\boldsymbol{s}}'_{\perp} \triangleq \hat{\boldsymbol{s}}' - \hat{\boldsymbol{s}}'_{\parallel}. \tag{4.23}$$

Then, the output SNR is defined by

$$\text{output SNR} \triangleq 10\log_{10}\frac{\|\hat{\boldsymbol{s}}'_{\parallel}\|_2^2}{\|\hat{\boldsymbol{s}}'_{\perp}\|_2^2}. \tag{4.24}$$

A higher output SNR means a better overall noise suppression performance.

In order to calculate the NRF and the SDI, we use the output of the noise suppression system when it processes the target or noise component separately. Note here that these components are available for evaluation. Specifically, we denote the output of a noise suppression multichannel filter $\boldsymbol{w}(\tau,\omega)$ in response to the target or noise component by

$$s_{\text{out}}(\tau,\omega) \triangleq \boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{c}(\tau,\omega), \tag{4.25}$$

$$v_{\text{out}}(\tau,\omega) \triangleq \boldsymbol{w}^{\mathsf{H}}(\tau,\omega)\boldsymbol{v}(\tau,\omega), \tag{4.26}$$

where $\boldsymbol{c}(\tau,\omega)$ is the target signal observed at the microphones. The time-domain signals $s_{\text{out}}[k]$ and $v_{\text{out}}[k]$ are computed by the inverse STFT of these time-frequency-domain signals,

whereby the SDI and the NRF are computed as follows [61]:

$$\mathrm{SDI} \triangleq 10\log_{10}\frac{\sum_{k=0}^{K-1}\{s_{\mathrm{out}}[k]-s[k]\}^2}{\sum_{t=0}^{N-1}s^2[k]}, \tag{4.27}$$

$$\mathrm{NRF} \triangleq 10\log_{10}\frac{\sum_{k=0}^{K-1}v_1^2[k]}{\sum_{t=0}^{N-1}v_{\mathrm{out}}^2[k]}, \tag{4.28}$$

where $v_1[k]$ denotes the noise at the first microphone. A higher NRF means a better noise reduction performance, whereas a lower SDI means a better signal preservance performance.

4.3.2 Experimental conditions

Compared methods

We compared the following six methods.

- The MVDR beamformer (2.39) (denoted by MVDR).
- The MVDR beamformer followed by the Wiener post-filter, (2.40), with $\phi_{ss}(\tau,\omega)$ estimated by the proposed unified estimator for each noise model, namely $\hat{\phi}_{ss}^{\mathrm{uncor}}(\tau,\omega)$, $\hat{\phi}_{ss}^{\mathrm{coh}}(\tau,\omega)$, $\hat{\phi}_{ss}^{\mathrm{BND}}(\tau,\omega)$, and $\hat{\phi}_{ss}^{\mathrm{real}}(\tau,\omega)$ (denoted by SV-uncor, SV-coh, SV-BND, and SV-real). For $\hat{\phi}_{ss}^{\mathrm{coh}}(\tau,\omega)$, $\boldsymbol{\Gamma}(\omega)$ was calculated using the cylindrically isotropic noise model (2.52), which generally resulted in a better noise suppression performance in terms of the output SNR than the spherically isotropic noise model (2.51) in a preliminary experiment. For $\hat{\phi}_{ss}^{\mathrm{BND}}(\tau,\omega)$, the 4×4 DFT matrix was used as the diagonalization matrix $\boldsymbol{P}$. Since $\boldsymbol{h}(\omega)$ is calculated by the planewave assumption given the true DOA in this experiment, $\hat{\phi}_{ss}^{\mathrm{uncor}}(\tau,\omega)$ coincides with $\hat{\phi}_{ss}^{\mathrm{Zel}}(\tau,\omega)$ in (2.49).
- The MVDR beamformer followed by the Wiener post-filter, (2.40), with McCowan's estimator $\hat{\phi}_{ss}^{\mathrm{Mc}}(\tau,\omega)$ in (2.56) (denoted by SV-Mc).

$\boldsymbol{h}(\omega)$ for MVDR beamforming and post-filtering was calculated based on the planewave assumption as in (2.28) using the true target DOA. $\boldsymbol{\Phi}_{\boldsymbol{xx}}$ for MVDR beamforming was computed by long-time temporal averaging of $\boldsymbol{x}(\tau,\omega)\boldsymbol{x}^{\mathsf{H}}(\tau,\omega)$ using all time frames. On the

other hand, $\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau, \omega)$ for post-filtering was computed by short-time temporal averaging of $\boldsymbol{x}(\tau, \omega)\boldsymbol{x}^{\mathsf{H}}(\tau, \omega)$ over 16 consecutive frames.

Array fabrication and generation of the observed signals

We fabricated a square array with a diameter of 5 cm, and recorded noise in the following environments in Tokyo:

- square,
- station building,
- train,
- platform.

We used electret-type microphones (SONY ECM-C10) and a multi-channel input board with microphone amplifiers (Tokyo Electron Device TD-BD-8CSUSB). The noise was recorded at the sampling frequency of 44.1 kHz, and then down-sampled to 16 kHz.

The signals observed by the array, denoted by $\boldsymbol{x}[k] \in \mathbb{R}^4$, were generated by summing a target component $\boldsymbol{c}[k]$ and a diffuse noise component $\boldsymbol{v}[k]$ as follows:

$$\boldsymbol{x}[k] = \boldsymbol{c}[k] + \boldsymbol{v}[k]. \tag{4.29}$$

$\boldsymbol{v}[k]$ is the recorded real-world noise. The target component $\boldsymbol{c}[k]$ was simulated under the assumption that the planewave from the target DOA is observed by the 4-element microphone array of the same configuration as the fabricated one. Specifically, $c_m[k]$ is calculated by delaying $c_1[k]$ by δ_m, which is the time it takes for the wave to propagate from the first microphone to the m-th microphone. Here, the subsample delay is realized by applying the phase-shift factor $\mathrm{e}^{-\mathrm{j}\omega\delta_m}$ in the DFT domain. The propagation of each speech interferences was simulated in the same way as the target signal. $\boldsymbol{v}[k]$ was scaled so that the input SNR at the first microphone was 0 dB. Here, the input SNR at the first microphone is defined by the following equation:

$$\text{input SNR} \triangleq 10\log_{10}\frac{\displaystyle\sum_{k=0}^{K-1} c_1^2[k]}{\displaystyle\sum_{k=0}^{K-1} v_1^2[k]} = 10\log_{10}\frac{\displaystyle\sum_{k=0}^{K-1} s^2[k]}{\displaystyle\sum_{k=0}^{K-1} v_1^2[k]}. \tag{4.30}$$

Note that this coincides with the SNR obtained by applying (4.24) to the observed signal at the first microphone, $x_1[k]$, if signal and noise at the microphone are uncorrelated to each other:

$$\sum_{k=0}^{K-1} s[k]v_1[k] = 0. \tag{4.31}$$

The duration of the observed signals was 4 s, and the sampling frequency was 16 kHz.

For each envirionment, we generated 29 samples of observed signals, where the following conditions were varied for each sample:

- the target speech file chosen from 303 files taken from the ATR Japanese speech database [62],
- the target DOA chosen randomly,
- the noise segment chosen from those generated by dividing the recorded noise data (duration: 4 s).

The database consists of 4 (number of environments) × 29 (number of samples of observed signals for each environment) = 116 observed signals in total.

Conditions for processing and evaluation

The observed signals were analyzed by STFT, where the frame length and the frame shift were 512 and 32, respectively and the Hamming window was used. For all methods, the 3 frequency bins at lowest frequencies were removed due to extremely low SNRs. The noise suppression techniques were applied to the observed signals in the STFT domain, and the noise suppression results were converted into the time domain by inverse STFT.

We averaged the performance measures (*i.e.* SNR, NRF, and SDI) over the 29 observed signals in each database to minimize the impact of the unwanted factors.

4.3.3 Experimental results

Tables 4.1–4.3 shows the output SNR, the NRF, and the SDI of the compared methods for different noise environments. In the tables, the bold figures show the best performance

Table 4.1 Output SNR (dB) of the compared methods for different noise environment.

method	MVDR	SV-Mc	SV-uncor	SV-coh	SV-BND	SV-real
square	12.4	13.9	13.2	**14.5**	**14.5**	14.4
station building	14.3	16.5	14.9	**17.0**	**17.0**	16.9
train	12.3	13.6	13.0	**14.8**	**14.8**	**14.8**
platform	13.3	14.4	14.2	**15.3**	**15.3**	15.2

Table 4.2 NRF (dB) of the compared methods for different noise environment.

method	MVDR	SV-Mc	SV-uncor	SV-coh	SV-BND	SV-real
square	13.0	**20.7**	14.0	18.5	18.5	18.5
station building	14.9	**21.3**	15.7	20.7	20.7	20.7
train	12.6	**21.9**	13.4	18.5	18.5	18.5
platform	13.7	**20.9**	14.8	20.4	20.4	20.4

among the methods. The following relationships held betweeen the methods:

$$\text{output SNR: MVDR} \prec \text{SV-uncor} \prec \text{SV-Mc} \prec \text{SV-coh} \simeq \text{SV-BND} \simeq \text{SV-real}, \tag{4.32}$$

$$\text{NRF: MVDR} \prec \text{SV-uncor} \prec \text{SV-coh} \simeq \text{SV-BND} \simeq \text{SV-real} \prec \text{SV-Mc}, \tag{4.33}$$

$$\text{SDI: MVDR} \succ \text{SV-uncor} \succ \text{SV-coh} \simeq \text{SV-BND} \simeq \text{SV-real} \succ \text{SV-Mc}, \tag{4.34}$$

where MVDR $\prec$ SV-uncor means that SV-uncor outperformed MVDR for instance. The rankings of the NRF and the SDI were inverse to each other, so the evaluation of the overall performance by the output SNR is important. All post-filtering techniques outperformed the beamformer in terms of the output SNR. Proposed SV-coh, SV-BND, and SV-real gave higher output SNRs than MVDR, SV-uncor, and SV-Mc. The former three methods gave virtually the identical output SNRs. The SNR gain of these methods compared to the input was 14.4–17.0 dB, and that compared to SV-Mc was 0.4–1.2 dB. Note that some of the noise environments included directional as well as diffuse noise (*e.g.* announcement from loudspeakers in the platform environment). Nevertheless, the proposed method worked well, and this showed that it was robust against the presence of directional noise to a certain degree.

Figures 4.2–4.9 show examples of spectrograms for the noise on a train. Figures 4.2 and 4.3 show the target and the observed spectrograms at the first microphone. The SNR gain of 12.3-14.3 dB by MVDR in Table 4.1 (Fig. 4.4) can mainly be attributed to the removal of the

Table 4.3 SDI (dB) of the compared methods for different noise environment.

method	MVDR	SV-Mc	SV-uncor	SV-coh	SV-BND	SV-real
square	**−28.5**	−15.3	−25.0	−17.2	−17.2	−17.2
station building	**−26.7**	−18.4	−25.3	−19.0	−19.0	−19.0
train	**−30.1**	−14.5	−28.5	−17.7	−17.7	−17.7
platform	**−29.5**	−15.7	−25.9	−16.8	−16.8	−16.8

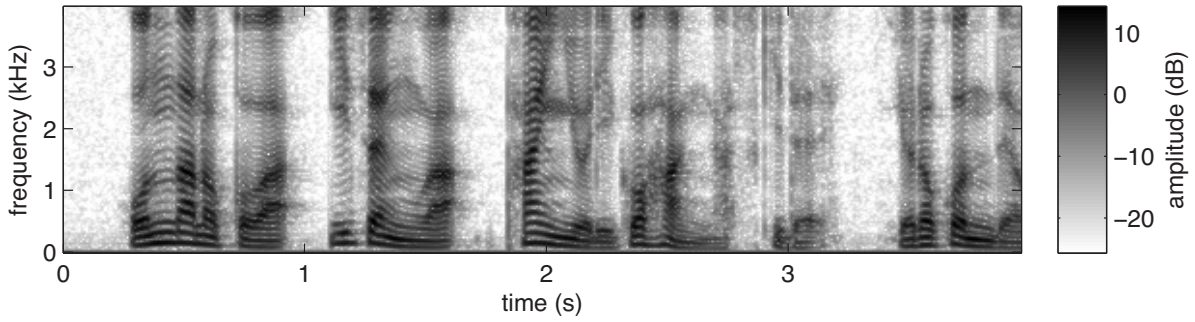


Figure 4.2 Spectrogram of the target signal.

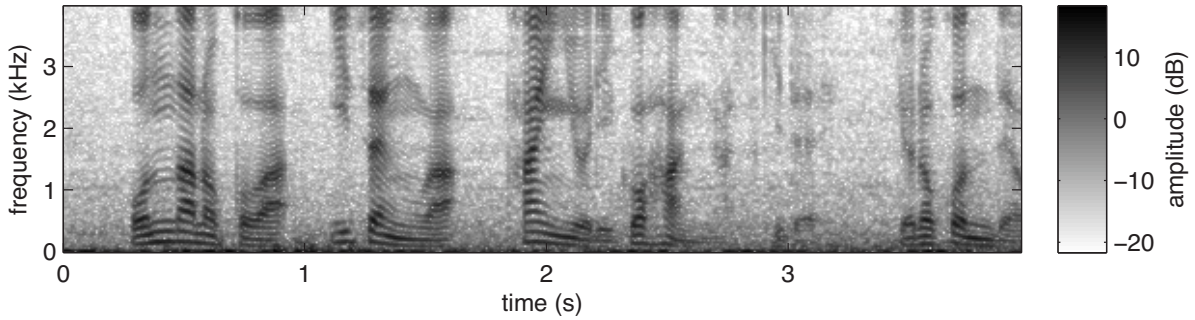


Figure 4.3 Spectrogram of the observed signal at the first microphone.

noisy three low-frequency bins and the suppression of directional noise. However, as seen in this figure, a large amount of noise remains unsuppressed, because real-world noise is diffuse rather than directional. SV-uncor (Fig. 4.6) reduced noise at high frequencies, but low-frequency noise remained, because of high correlation between microphones. By contrast, SV-Mc (Fig. 4.5) reduced noise dramatically at all frequencies. However, it caused much distortion at low frequencies, with the first and the second partials of the speech spectrum sometimes lost. On the other hand, proposed SV-coh, SV-BND, and SV-real suppressed noise effectively without much speech distortion unlike SV-Mc.

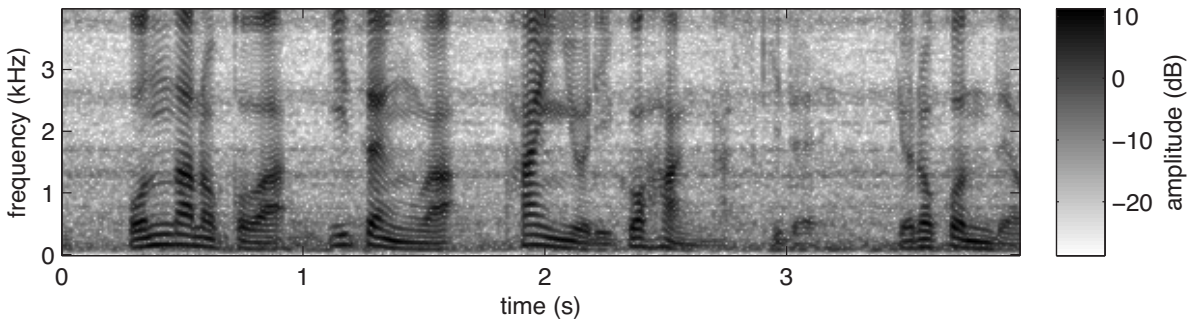


Figure 4.4 Spectrogram of the output of MVDR.

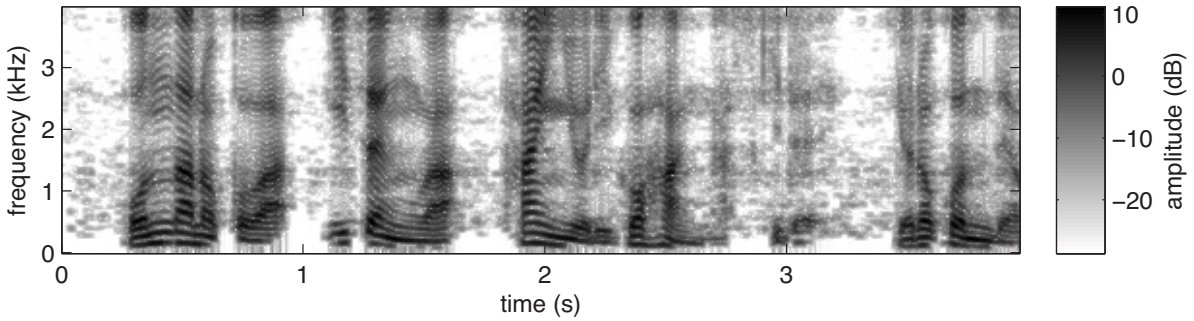


Figure 4.5 Spectrogram of the output of SV-Mc.

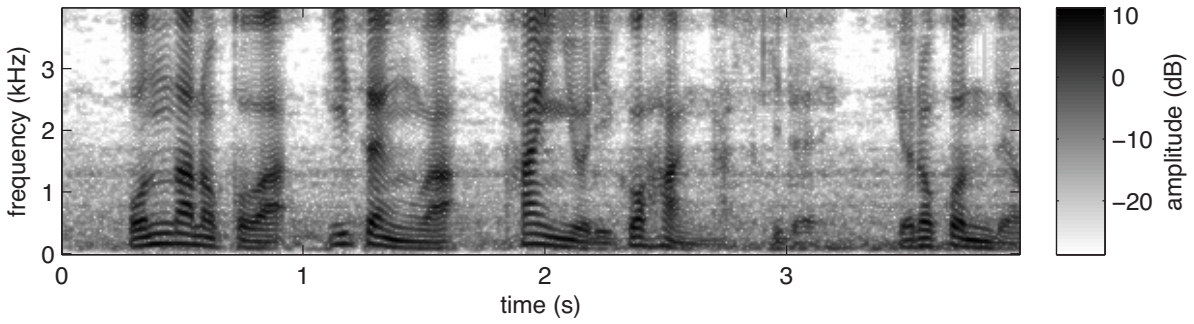


Figure 4.6 Spectrogram of the output of SV-uncor.

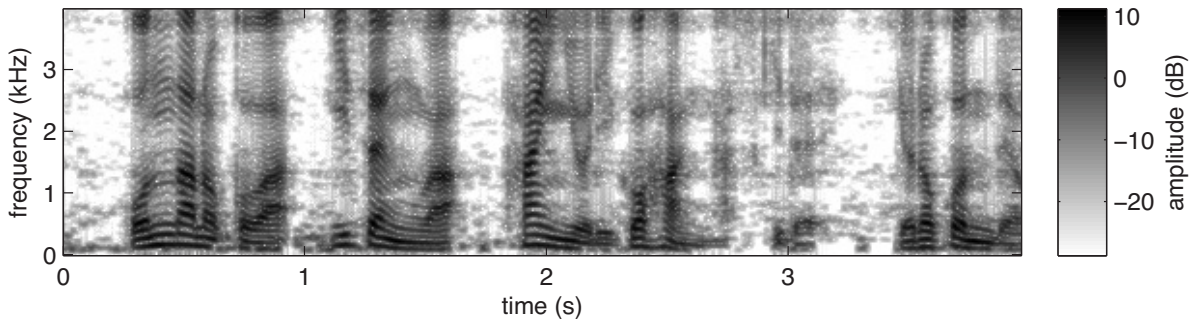


Figure 4.7 Spectrogram of the output of SV-coh.

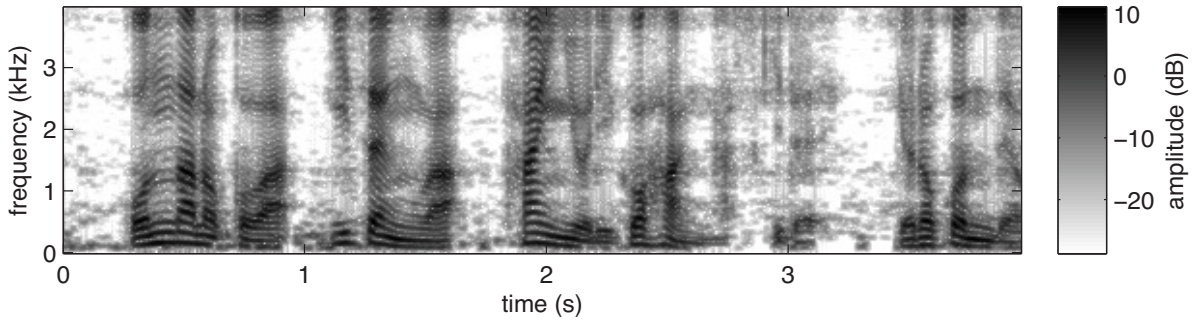


Figure 4.8 Spectrogram of the output of SV-BND.

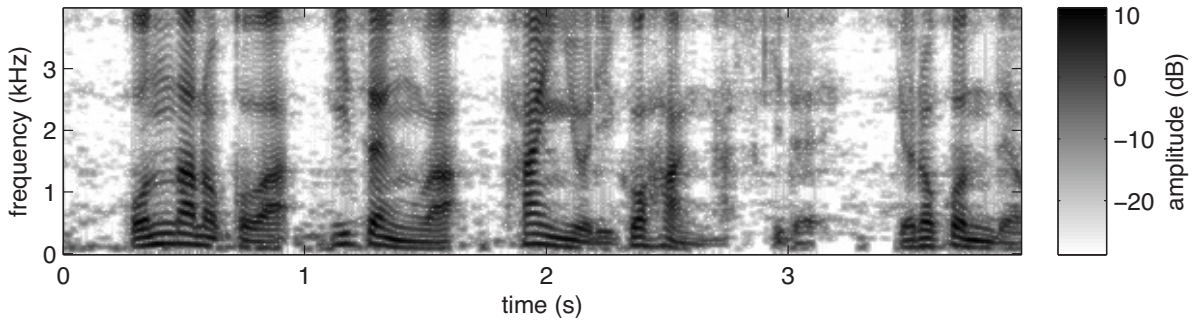


Figure 4.9 Spectrogram of the output of SV-real.

Chapter 5

Noise-Robust Estimation of Directions of Arrival of Target Signals

This chapter has partly been published in [12, 13].

This chapter aims to robustly estimate the DOAs of multiple sounds in diffuse noise. We have seen in Chapter 2 that DOA estimation of multiple sounds can be addressed by MUSIC. In this framework, it is essential to accurately estimate matrix $\boldsymbol{E}(\omega)$ in (2.78), which boils down to the estimation of the signal covariance matrix.

The subspace model implies that we obtain the orthogonal component of $\boldsymbol{\Phi}_{\boldsymbol{cc}}$ from the data as follows:

$$\mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{cc}}] = \mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}]. \tag{5.1}$$

Therefore, the problem becomes one of estimating the underlying matrix $\boldsymbol{\Phi}_{\boldsymbol{cc}}$ from its partial observation in the subspace $\mathcal{V}^{\perp}$, namely $\mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{cc}}]$. In the literature, matrix completion techniques [63, 64, 65, 66] have been proposed, which aims to recover an underlying low-rank matrix from the observation of only part of its entries. It is the low-rankness of the underlying matrix that makes this feasible. We extend the techniques in [63, 64] so that we can recover an underlying low-rank positive semidefinite matrix from its partial observation in a subspace. We present two methods based on low-rank matrix completion and trace norm minimization.

The rest of this chapter is organized as follows. In Sections 5.1 and 5.2, we present the methods based on low-rank matrix completion and trace norm minimization, respectively. In Section 5.3, we assess the proposed algorithms with a large database with various mixture parameters.

5.1 Unified framework based on low-rank matrix completion

The first algorithm is based on the main assumption that an upper bound of the rank of $\boldsymbol{\Phi}_{\boldsymbol{cc}}$ is given. Instead of regarding $\mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}]$ as exactly noise-free, we leave some room for possible errors due to the misestimation of $\boldsymbol{\Phi}_{\boldsymbol{xx}}$ or the imperfection of the noise model. Specifically, of the Hermitian positive semidefinite matrices of rank no greater than R, we seek for the one whose orthogonal projection onto $\mathcal{V}^{\perp}$ is closest to that of $\boldsymbol{\Phi}_{\boldsymbol{xx}}$:

$$\begin{aligned} &\min_{\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}} \ \Psi_{\mathrm{comp}}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}) \triangleq \|\mathcal{P}^{\perp}[\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}] - \mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}]\|_{\mathrm{F}}^2 \\ &\text{s.t. } \hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}\text{: Hermitian positive semidefinite}, \mathrm{rank}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}) \leq R. \end{aligned} \tag{5.2}$$

Here, the positive semidefiniteness constraint is important, because the eigenvectors of $\boldsymbol{\Phi}_{\boldsymbol{cc}}$ belonging to the positive and zero eigenvalues are to be regarded as bases of the signal and noise subspaces, respectively.

Though the constraint in (5.2) may appear complex, we can optimize Ψ_{comp} efficiently based on the following theorem:

Theorem 1. *Let $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(0)}$ be a Hermitian positive-semidefinite matrix, and let us calculate $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k+1)}$ given $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k)}$ for $k = 0, 1, \ldots$ according to the following update rules. Then, $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k)}$ $(k = 1, 2, \cdots)$ is also a Hermitian positive-semidefinite matrix such that rank$(\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k)}) \leq R$, and the obtained sequence of $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k)}$ decreases Ψ_{comp} monotonically: $\Psi_{comp}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(0)}) \geq \Psi_{comp}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(1)}) \geq \cdots$.*

- *Calculate $\boldsymbol{Y}^{(k+1)}$ by*

$$\boldsymbol{Y}^{(k+1)} \triangleq \mathcal{P}[\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k)}] + \mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}]. \tag{5.3}$$

- *Calculate the eigenvalue decomposition of $\boldsymbol{Y}^{(k+1)}$:*

$$\boldsymbol{Y}^{(k+1)} = \boldsymbol{U}^{(k+1)} \boldsymbol{\Sigma}^{(k+1)} \boldsymbol{U}^{H(k+1)}, \tag{5.4}$$

  where $\boldsymbol{U}^{(k+1)}$ is unitary and $\boldsymbol{\Sigma}^{(k+1)}$ is real-valued and diagonal, where the diagonal entries $\sigma_1^{(k+1)}, \ldots, \sigma_M^{(k+1)}$ are arranged in decreasing order: $\sigma_1^{(k+1)} \geq \cdots \geq \sigma_M^{(k+1)}$.

- *Calculate $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k+1)}$ by*

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k+1)} = \boldsymbol{U}^{(k+1)} \boldsymbol{\Sigma}_R^{(k+1)} \boldsymbol{U}^{H(k+1)}, \tag{5.5}$$

where

$$\mathbf{\Sigma}_R^{(k+1)} \triangleq diag\Big(\max\{\sigma_1^{(k+1)},0\},\ldots,\max\{\sigma_R^{(k+1)},0\},\underbrace{0,\ldots,0}_{M-R}\Big), \tag{5.6}$$

with $diag(\alpha_1,\ldots,\alpha_M)$ *denoting the* $M\times M$ *diagonal matrix composed of* $\alpha_1,\ldots,\alpha_M$*, and* $\max\{a,b\}$ *denoting the maximum of* a *and* b*.*

We propose to initialize $\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}$ by $\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}^{(0)} = \mathbf{\Phi}_{\boldsymbol{xx}}$ and iterate the updates (5.3) to (5.5) until a preset maximum number of iteration is reached or

$$\frac{\|\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}^{(k+1)} - \hat{\mathbf{\Phi}}_{\boldsymbol{cc}}^{(k)}\|_{\mathrm{F}}}{\|\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}^{(k)}\|_{\mathrm{F}}} < \epsilon, \tag{5.7}$$

where ϵ is a preset small constant, which means $\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}^{(k)}$ has become almost constant.

This algorithm can be regarded as a generalization of Srebro's algorithm [63]. The generalization is twofold. First, we consider the completion of an arbitrary missing subspace instead of missing entries. Second, we consider optimization of a complex-valued matrix with an Hermitian positive semidefiniteness constraint instead of real-valued matrix without such a constraint.

5.2 Unified framework based on trace-norm minimization

While the algorithm in Section 5.1 requires an upper bound on the rank of $\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}$, now we propose another algorithm that does not require that knowledge. This is advantageous because the the upper bound is not always given in practice due to an unknown number of sources and/or reverberation. We can construct a cost function that favors a solution of lower rank without knowing the upper bound by using a trace norm $\|\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}\|_*$. The trace norm of a matrix is defined as the sum of the singular values (or equivalently the eigenvalues for Hermitian positive semidefinite matrices) of this matrix, and is a convex relaxation of the rank function [64]. Specifically, we consider the following criterion regularized by the trace norm:

$$\begin{aligned}&\min_{\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}}\ \Psi_{\mathrm{trace}}(\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}) \triangleq \frac{1}{2}\|\mathcal{P}^{\perp}[\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}] - \mathcal{P}^{\perp}[\mathbf{\Phi}_{\boldsymbol{xx}}]\|_{\mathrm{F}}^2 + \mu\|\hat{\mathbf{\Phi}}_{\boldsymbol{cc}}\|_*,\\ &\text{s.t. } \hat{\mathbf{\Phi}}_{\boldsymbol{cc}}:\ \text{Hermitian positive semidefinite,}\end{aligned} \tag{5.8}$$

where μ is a positive regularization weight. Compared to (5.2), (5.8) does not have a rank constraint, but a trace norm $\|\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}\|_*$ in the cost function instead.

This optimization problem can be solved efficiently with the following algorithm. This is obtained by generalizing Toh's algorithm [64] to the optimization of a complex-valued matrix subject to a Hermitian positive semidefiniteness constraint.

Algorithm 2.

- *Set* $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(0)} = \hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(-1)} = \boldsymbol{\Phi}_{\boldsymbol{xx}}$.
- *For* $k = 0, 1, \dots,$
  - $\boldsymbol{Z}^{(k)} \triangleq \hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k)} + \frac{t^{(k-1)}-1}{t^{(k)}}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k)} - \hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k-1)})$.
  - $\boldsymbol{Y}^{(k)} \triangleq \mathcal{P}[\boldsymbol{Z}^{(k)}] + \mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}]$.
  - *Calculate the eigenvalue decomposition of* $\boldsymbol{Y}^{(k)}$*:* $\boldsymbol{Y}^{(k)} \triangleq \boldsymbol{U}^{(k)}\boldsymbol{\Sigma}^{(k)}\boldsymbol{U}^{(k)H}$*, where* $\boldsymbol{U}^{(k)}$ *is unitary and* $\boldsymbol{\Sigma}^{(k)}$ *is real-valued and diagonal.*
  - $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k+1)} \triangleq \boldsymbol{U}^{(k)} \max\{\boldsymbol{\Sigma}^{(k)} - \mu\boldsymbol{I}, 0\}\boldsymbol{U}^{(k)H}$.
  - $t^{(k+1)} \triangleq \frac{1+\sqrt{1+4t^{(k)2}}}{2}$.

Here, $\max\{\cdot, 0\}$ denotes the operation of replacing the negative entries of a matrix with zeros.

The following theorem guarantees that Algorithm 2 converges to a global minimum, for the cost function is convex. It can be proven in line with [64].

Theorem 2. *Let* $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k)}$ $(k = -1, 0, 1, \dots)$ *be the sequence generated by Algorithm 2. Then,*

$$|\Psi_{trace}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{(k)}) - \Psi_{trace}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{opt})| \leq \frac{(\|\boldsymbol{\Phi}_{\boldsymbol{xx}}\|_F + \chi)^2}{(k+1)^2}, \tag{5.9}$$

where $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}^{opt}$ *is an optimum solution, and* χ *is an upper bound of the Frobenius norm of the optimal solutions.*

The stopping condition is defined in the same manner as in the algorithm in Section 5.1.

5.3 Large-scale evaluation with real-world noise

In this section, we evaluate the performance of the proposed methods for DOA estimation using a large database with real-world noise.

5.3.1 Created database

We created a database of multichannel speech mixtures in the presence of real-world background noise. We used the real world noise data explained in Section 4.3.2. The reverberant target components at the microphones were simulated and added to the noise data. The dry speech sources were taken from the ATR Japanese database [62]. The source images at the microphones were simulated via the image method [67] (We used a matlab code "roomsim_single.m" written by Vincent [68].), and mixtures were generated by adding the source and the noise components together. The room dimensions were assumed to be $3.3 \times 7.8 \times 2.4$ m. Fig. 5.1 illustrates the geometry of the room and the array. Compared to simultaneously recording the target signals and noise, this is advantageous in controlling the mixture conditions. The velocity of sound was assumed to be 340 m/s. All mixtures were 10 second long and their sampling frequency was 16 kHz.

The database was designed to evaluate the effect of the following 4 parameters:

- the number L of target sources: 2, 4, or 6,
- the angle between adjacent sources: 30°, 60°, or 90°,
- the absorption coefficient of the walls: 0.4, 0.7, or 1.0 (These correspond to the reverberation time RT_{60} (*i.e.* the time it takes for the reverberation to decay by 60 dB) of 186, 79, or 0 ms at 500 Hz.),
- the input SNR: 10, 0, or -10 dB.

Here, the input SNR was defined as the energy ratio between a target component and diffuse noise at the first microphone, where the energies of all sources were set to the same value. Excluding geometrically infeasible combination of 6 sources and 90° angle, we have $(3 \times 3 - 1) \times 3 \times 3 = 72$ mixtures in total.

5.3.2 Methods compared and algorithmic parameters

We compared the performance of the following 10 methods:

- conventional MUSIC based on EVD of the observed covariance matrix (denoted as conv-white),

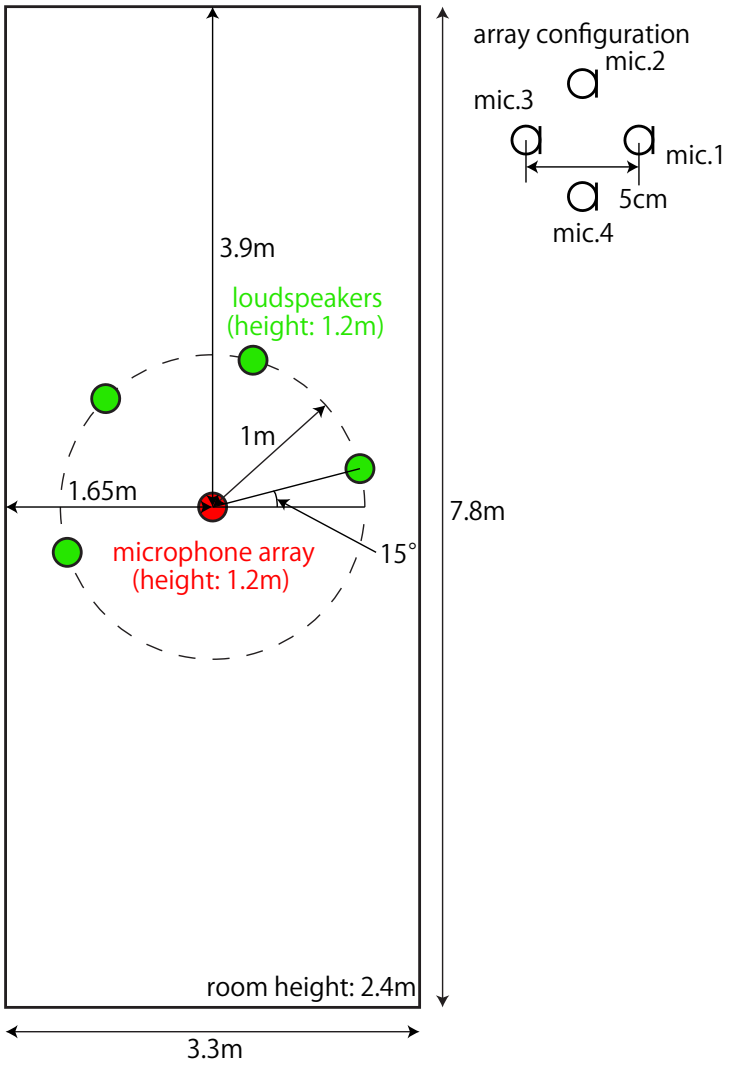


Figure 5.1 The array geometries in the experiment. The case of 4 sources and 60° angle between adjacent sources is shown as an example.

- conventional MUSIC based on the generalized eigenvalue problem $\boldsymbol{Au} = \alpha \boldsymbol{Bu}$, with the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ being the observed covariance matrix and the coherence matrix for cylindrically isotropic noise (denoted as conv-coh),

- MUSIC based on EVD of the signal covariance matrix estimated by the low-rank matrix completion algorithm with the uncorrelated noise model, the fixed noise coherence model, the BND model, and the real-valued noise covariance model (denoted as comp-uncor, comp-coh, comp-BND, and comp-real, respectively),

- MUSIC based on EVD of the signal covariance matrix estimated by the trace norm minimization algorithm with the uncorrelated noise model, the fixed noise coherence

model, the BND model, and the real-valued noise covariance model (denoted as trace-uncor, trace-coh, trace-BND, trace-real, respectively).

We examined the impact of the following two algorithmic parameters:

- The assumed dimension R of the signal subspace: 1, 2, or 3
- The number B of time blocks over which the local angular spectra were calculated: 1, 4, or 16

Assuming L was known, we picked L peaks in the MUSIC spectrum. In the low-rank approximation algorithm, the knowledge of R is used in the $\boldsymbol{\Phi}_{\boldsymbol{cc}}$ estimation and in the derivation of the basis vectors of the noise subspace. On the other hand, in the trace norm minimization algorithm and the conventional MUSIC algorithm, it is used for the latter purpose only.

We set the minimum angular distance Δ defined in Section 2.4.3 to $\Delta = 15°$.

5.3.3 Evaluation metrics

We assess the performance of DOA estimation in terms of *F-measure* and *Root Mean Square Error (RMSE)* [69].

We define a correct peak as one within 5° from a true azimuth. Let C_J denote the number of correct peaks as a function of the number J of selected peaks. Then the precision is defined by

$$P_J \triangleq \frac{C_J}{J} \tag{5.10}$$

and the recall by

$$R_J \triangleq \frac{C_J}{L}. \tag{5.11}$$

The F-measure is defined as this harmonic average as follows:

$$F_J \triangleq \frac{2P_J R_J}{P_J + R_J}. \tag{5.12}$$

Since we assume $J = L$, $P_J = R_J = F_J$. Therefore, of these three metrics, we consider only the F-measure in the following. The RMSE is defined as the square root of the mean square error of the correct peaks.

We extend these criteria to a set of mixtures as the arithmetic average of the criteria across all mixtures.

5.3.4 Experimental results

Table 5.1 shows the F-measure and the RMSE for each method averaged over all mixtures. We set $R = 2$ and $B = 16$. The algorithms based on the fixed noise coherence model (conv-coh, comp-coh, and trace-coh) gave higher F-measures than conv-white by 0.07-0.08. The algorithm trace-BND gave an even higher F-measure than these methods by 0.04-0.05, and also a lower (better) RMSE by 0.3°-0.5°.

The other algorithms (comp-BND, comp-uncor, trace-uncor, comp-real, and trace-real) did not improve (or rather graded) the F-measure compared to conv-white. This implies the inefficiency of the uncorrelated noise model and the real noise covariance model. On the other hand, it is still under investigation why the BND model works well with the trace norm minimization algorithm and does not with the matrix completion algorithm. The above five algorithms shall not be considered in the rest of the experiment.

Figure 5.2 shows examples of the MUSIC spectrum. Fig. 5.2(a) corresponds to a highly reverberant and noisy condition (absorption coefficient: 0.4; SNR: -10 dB). The number of sources was 2, and the true DOAs were 15° and 315° as depicted by the vertical lines in the figure. The largest peaks of conv-white are at 0° and 180°. The other algorithms resulted in less estimation error, and trace-BND resulted in the most accurate peak positions. Fig. 5.2(b) shows another example with a large number of sources. The parameters R and B were increased to $R = 3$ and $B = 16$ to deal with many sources. Only trace-BND managed to accurately localize all of four sources. The DOA estimated by it least accurately was 315°, for which the error was about 7°. On the other hand, the other methods resulted in an error of more than 15° for at least one of the DOAs.

We observed that conv-white tended to have spurious peaks at multiples of 90° in diffuse noise, especially at low SNRs. On the other hand, this effect was less prominent for conv-coh, comp-coh, trace-coh, and trace-BND, which take the correlation of diffuse noise into account.

Here, we would like to give some comments on the comparison of the algorithms based on the fixed noise coherence model: conv-coh, comp-coh, and trace-coh. The algorithm trace-coh generally gave a MUSIC spectrum very similar to that of conv-coh. In comparison, the MUSIC spectrum of comp-coh was similar to conv-coh for $R = 2$, but approached that of conv-white, when $R = 3$ (see Fig. 5.2). This can be understood by considering the ultimate case of $R = M = 4$. In this case, $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}} = \boldsymbol{\Phi}_{\boldsymbol{xx}}$ is a solution to the optimization problem (5.2),

Table 5.1 F-measure and RMSE averaged over all mixtures. We set $R = 2$ and $B = 16$.

method	conventional		completion				trace-norm minimization			
noise model	white	coh	uncor	coh	BND	real	uncor	coh	BND	real
F-measure	0.51	0.58	0.41	0.59	0.42	0.48	0.51	0.58	**0.63**	0.35
RMSE	1.4°	2.1°	**1.1°**	1.9°	1.2°	1.5°	1.6°	2.0°	1.6°	1.8°

and the algorithm stops after one iteration with the output $\hat{\mathbf{\Phi}}_{\boldsymbol{cc}} = \mathbf{\Phi}_{\boldsymbol{xx}}$.

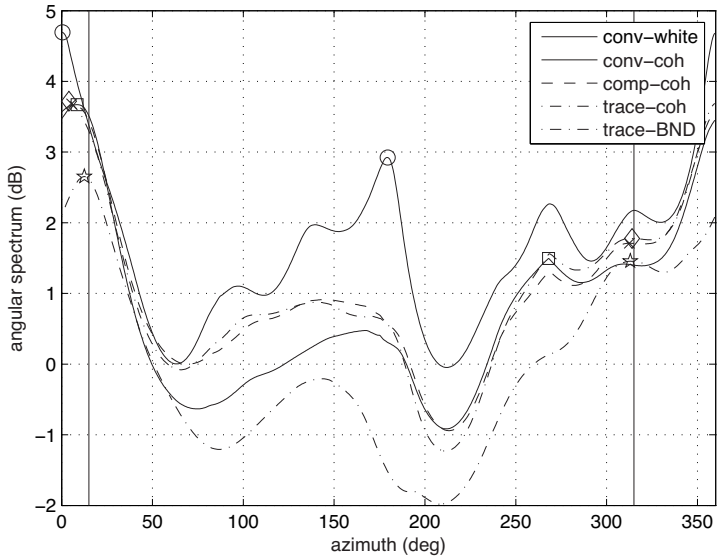
Figs. 5.3-5.6 show the impact of the mixture parameters. The algorithmic parameters were fixed at $R = 2$ and $B = 16$. The database was subdivided into 3 groups according to the value of the mixture parameter of interest. The performance criteria of the algorithms were averaged over the mixtures in each group and plotted as a function of the parameter of interest.

Figure 5.3 shows the impact of the input SNR. The F-measures decreased with decreasing SNR. At the SNR of 10 dB, the F-measure of trace-BND was around 0.7, and comparable to those of conv-coh, comp-coh, and trace-coh. It remained almost unchanged when the SNR decreased from 10 dB to 0 dB, whereas the F-measures of the other three algorithms decreased by about 0.1.

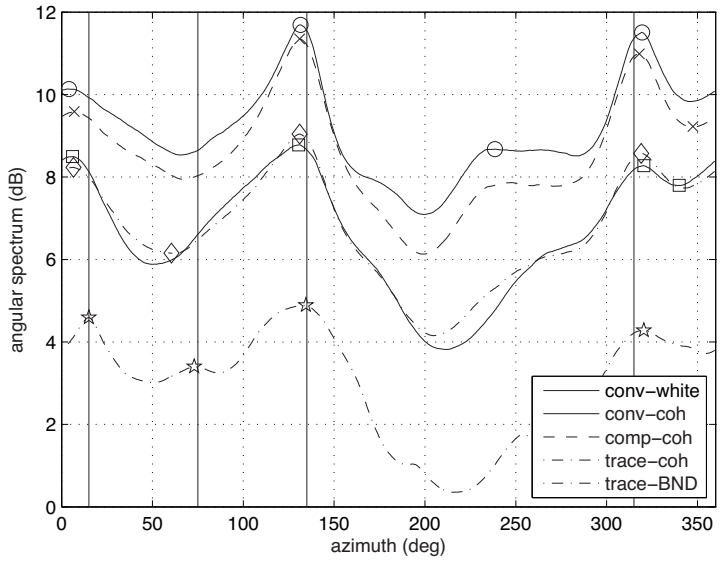
Figure 5.4 shows the measures as a function of L. The algorithm trace-BND gave high F-measures even for very large L (*e.g.* about 0.6 for $L = 6$), whereas those of other algorithms degraded significantly with the increase of L (*e.g.* a decrease by about 0.3 when L increased from 2 to 6). Note however that the F-measure of trace-BND was slightly lower than that of conv-white, for $L = 2$.

Figure 5.5 shows the measures as a function of the angle between adjacent sources. When the angle decreased from 90° to 30°, the F-measure of all methods dropped from approximately 0.8 to 0.4. For the angle of 60°, trace-BND gave a better F-measure (about 0.7), compare to those of conv-coh, comp-coh, and trace-coh (about 0.6) and that of conv-white (about 0.5).

Shown in Figure 5.6 is the impact of the absorption coefficient of the walls, which varies inversely with the reverberation time. The F-measure decreased with a decreasing absorption coefficient (*i.e.* increased amount of reverberation). The F-measure of trace-BND dropped like the other methods when the coefficient decreased from 1 to 0.7, while the decrease was much smaller compared to these methods when it decreased from 0.7 to 0.4. This can be interpreted as follows: the algorithm was affected by the early reflections, which



(a) $L = 2$; angle between adjacent sources: 60°; absorption coefficient: 0.4; SNR: -10 dB; $R = 2$; $B = 1$.

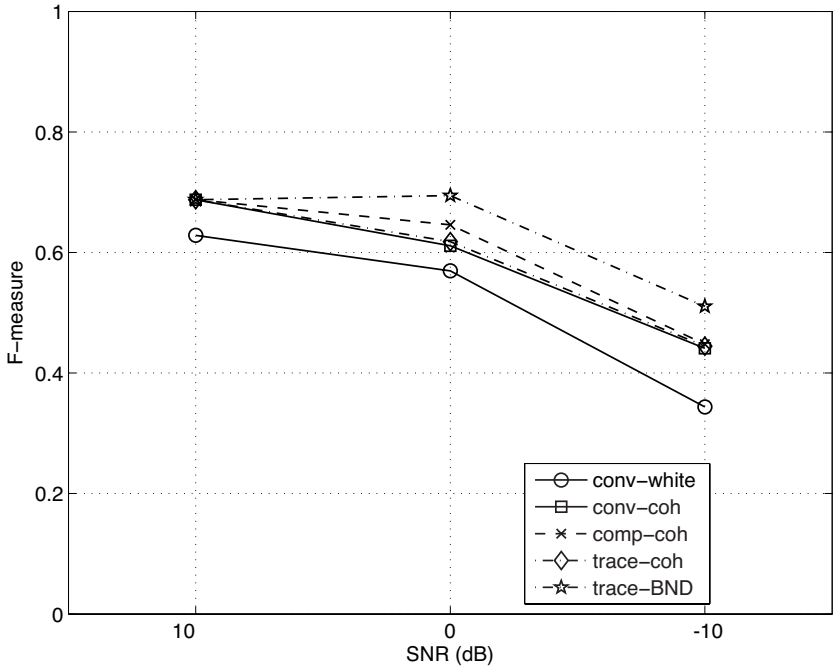


(b) $L = 4$; angle between adjacent sources: 60°; absorption coefficient: 0.4; SNR: 10 dB; $R = 3$; $B = 16$.

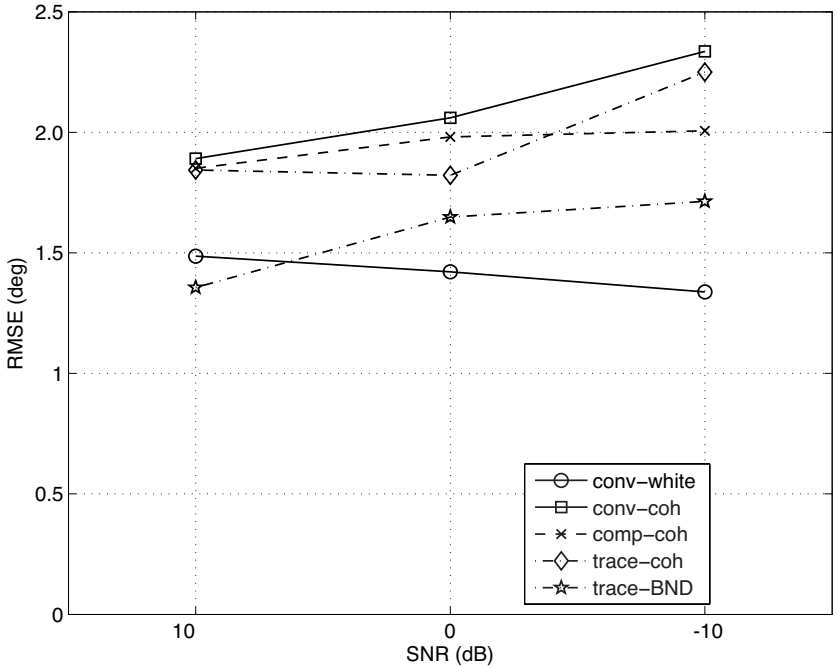
Figure 5.2 Examples of the MUSIC spectrum. The markers are defined in the same way as in Fig. 5.3.

violate the anechoic propagation assumption employed by MUSIC, but was robust against late reverberation, which may be regarded as diffuse noise and well explained by the BND model.

Figures 5.7 and 5.8 shows the effects of the algorithmic parameters. The performance criteria of the results for a fixed value of the algorithmic parameter of interest were averaged over the mixtures, and the averaged criterion was plotted as a function of the parameter. In Fig. 5.7, the impact of R is plotted. The F-measure increased by about 0.1-0.3 with an increased R. Also, as shown in Fig. 5.8, the F-measure increased slightly (by about 0.1 for conv-white and by less than 0.1 for the other methods) when B increased.

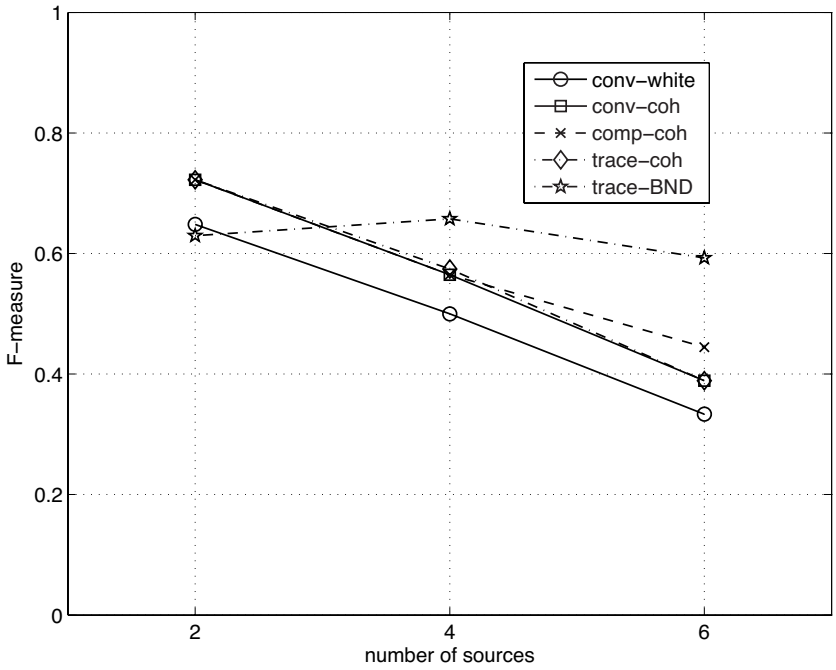


(a) F-measure

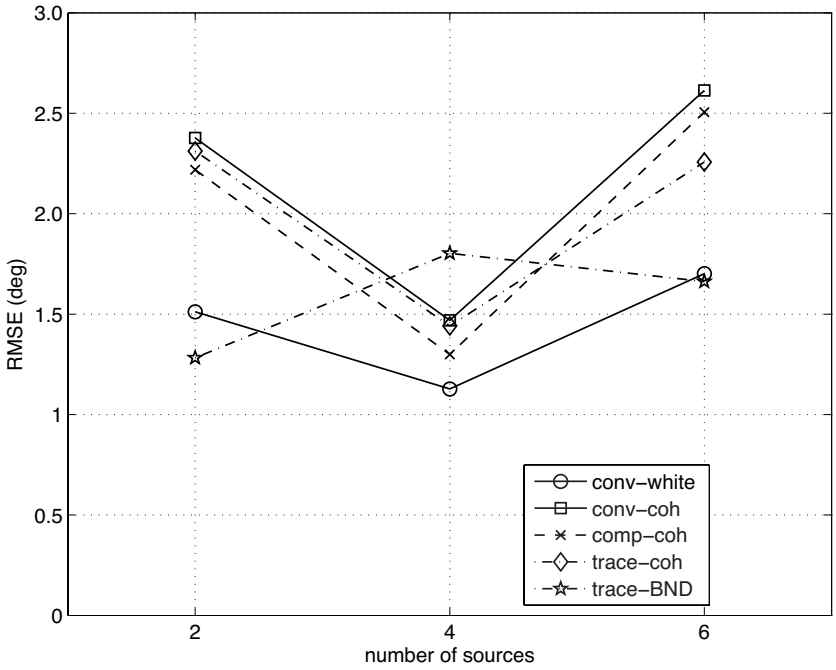


(b) RMSE

Figure 5.3 F-measure and RMSE as a function of the SNR for the conventional and proposed methods for $R = 2$ and $B = 16$.

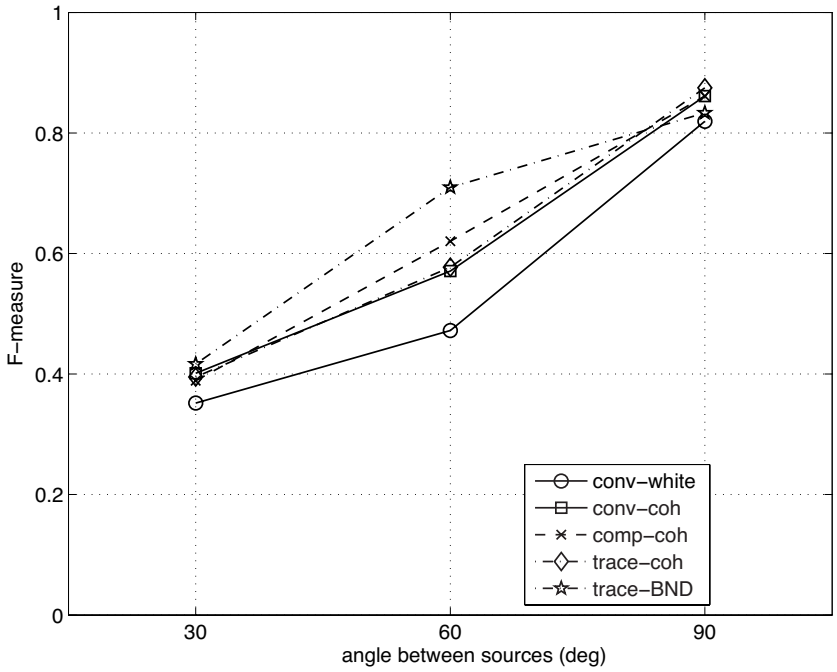


(a) F-measure

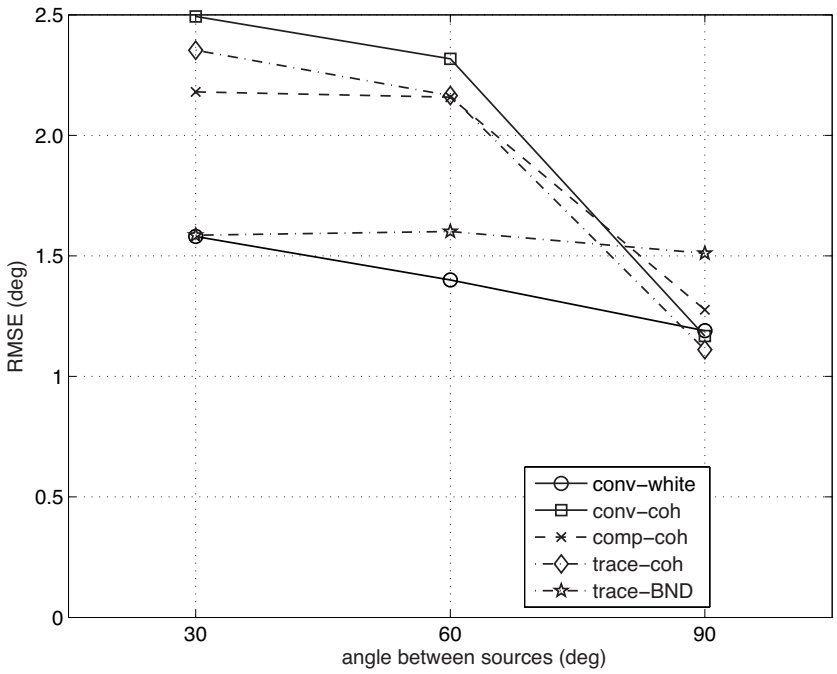


(b) RMSE

Figure 5.4 F-measure and RMSE as a function of the number of sources for the conventional and proposed methods for $R = 2$ and $B = 16$.

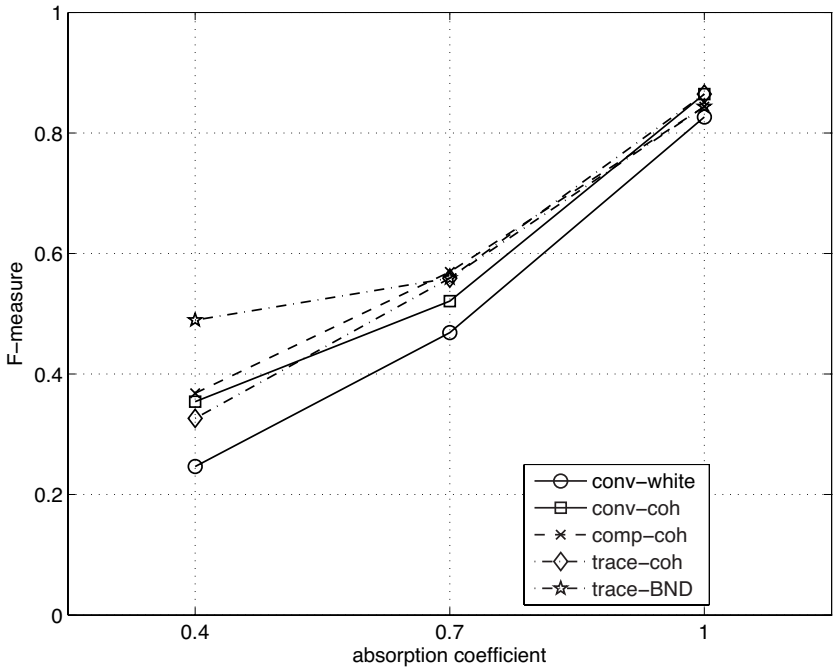


(a) F-measure

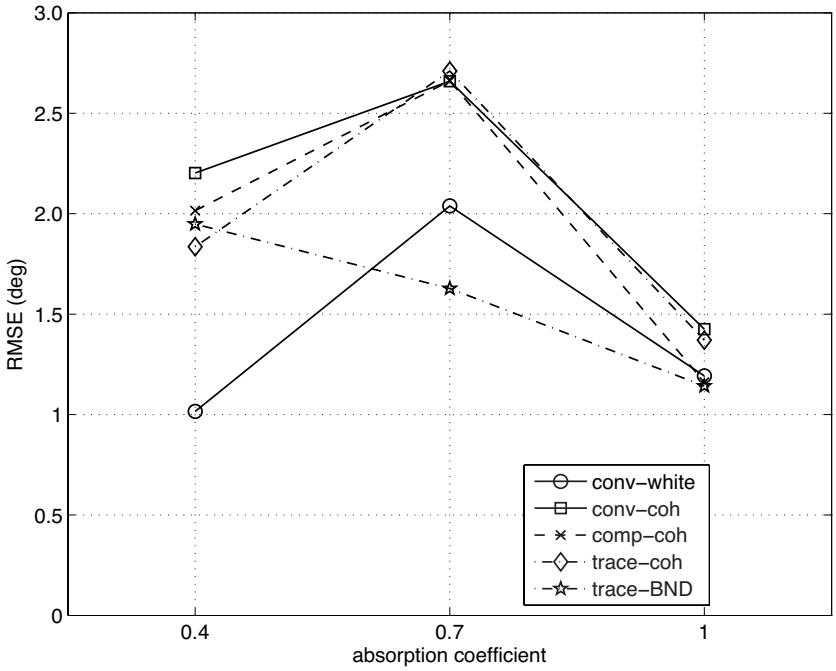


(b) RMSE

Figure 5.5 F-measure and RMSE as a function of the angle between adjacent sources for the conventional and proposed methods for $R = 2$ and $B = 16$.

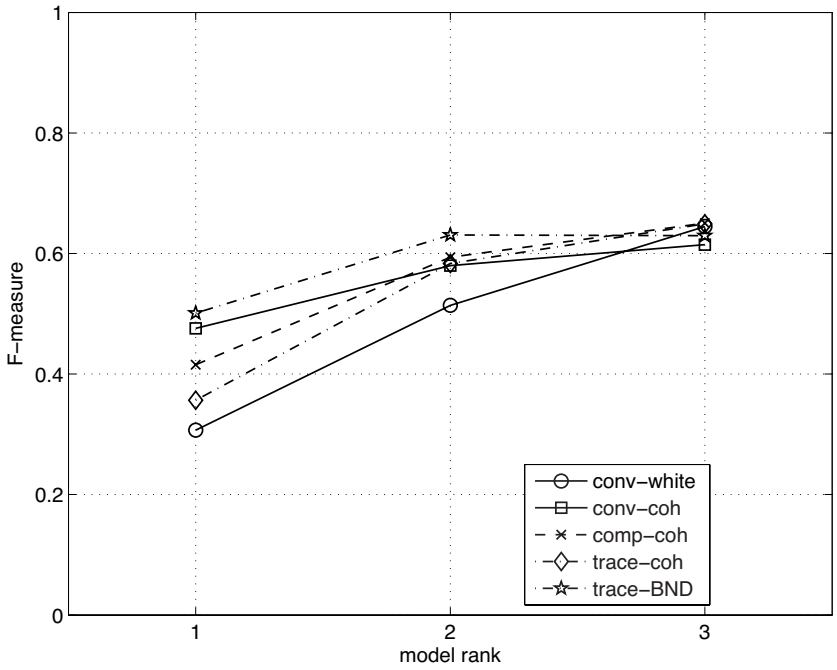


(a) F-measure

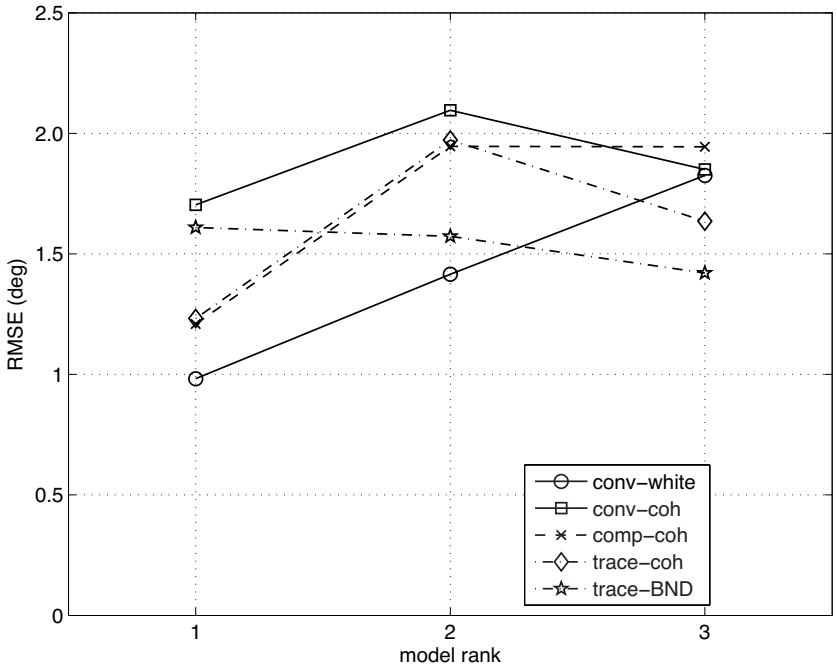


(b) RMSE

Figure 5.6 F-measure and RMSE as a function of the absorption coefficient of the walls for the conventional and proposed methods for $R = 2$ and $B = 16$.

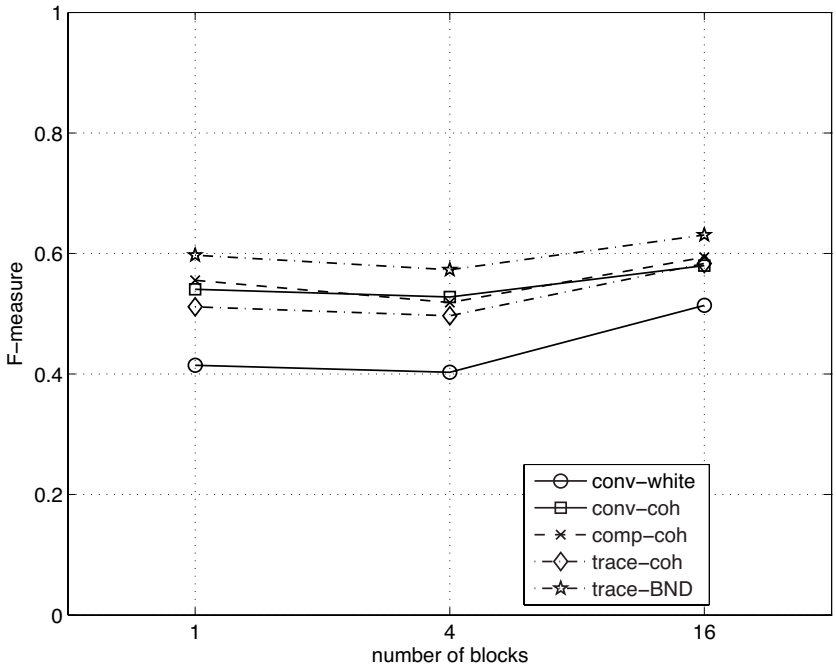


(a) F-measure

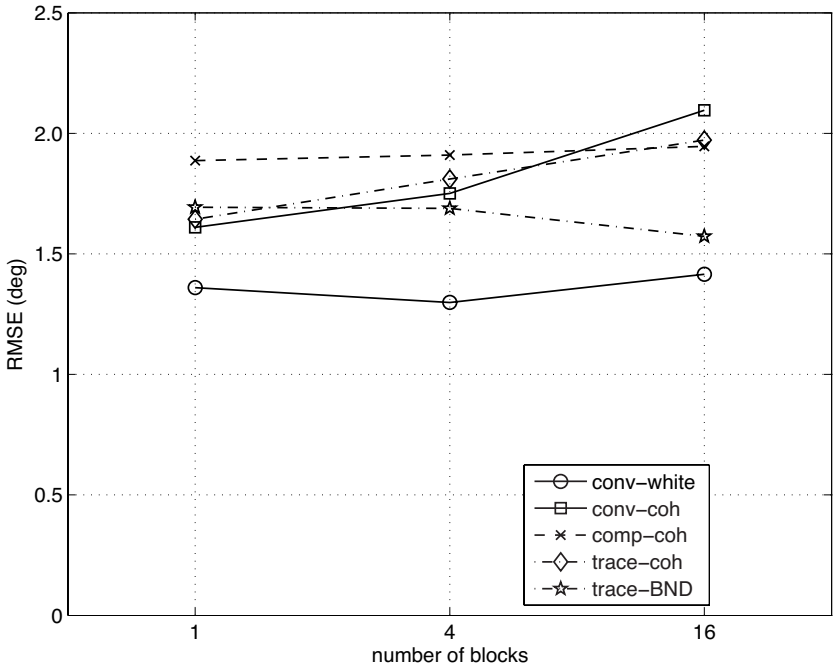


(b) RMSE

Figure 5.7 F-measure and RMSE as a function of the model rank for the conventional and proposed methods for $B = 16$.



(a) F-measure



(b) RMSE

Figure 5.8 F-measure and RMSE as a function of the number of blocks for the conventional and proposed methods for $R = 2$.

Chapter 6

Diffuse Noise Suppression for Target Signal from Unknown Direction

This chapter has partly been published in [8].

In Chapter 4, we presented a method for diffuse noise suppression with a known steering vector. However, the steering vector is not always known in practice, because the target DOA is not necessarily known. Moreover, even if we know the target DOA, this does not necessarily mean that we know the steering vector, because of reflections on the walls and/or diffraction by the rigid mount of the array. Therefore, in order to obtain satisfactory noise suppression performance in the real world, it is important to estimate the steering vector from the observed data. In this chapter, we present a method for estimating the target power spectrogram and the target steering vector jointly from the observed data. These estimates can then be utilized to design the MVDR beamformer and the Wiener post-filter, thereby enabling blind extraction of the target signal from the observed data. Furthermore, we fabricate an icosahedral microphone array, and validate the method through a real-world experiment using the target signal and noise recorded with the fabricated array.

There are several related researches. In a general transfer function generalized sidelobe canceller proposed by Gannot *et al.* [70], normalized transfer functions are estimated on the assumption that noise is stationary for a longer period compared to the target signal. Methods proposed by Benesty *et al.* [71] and Doclo *et al.* [72] manage to avoid the problem by calculating a multichannel noise suppression filter using solely the spatial covariance matrices of the observed signals and noise in order to estimate the spatial images of the target signal. In comparison, we focus on effective diffuse noise suppression in such a realistic scenario of

an unknown steering vector.

6.1 Unified framework for diffuse noise suppression with an unknown steering vector

In Chapter 4, we assumed that the steering vector $\boldsymbol{h}(\omega)$ is given. In this case, $\phi_{ss}(\tau,\omega)$ minimizing the squared error cost (4.3) can be derived by differentiating (4.3) with respect to $\phi_{ss}(\tau,\omega)$. In comparison, here we consider both $\phi_{ss}(\tau,\omega)$ and $\boldsymbol{h}(\omega)$ to be unknown. Since $\boldsymbol{h}(\omega)$ appears in a rather complex term $\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]$ in (4.3), it is difficult to derive the partial derivative of (4.3) with respect to $\boldsymbol{h}(\omega)$. Therefore, instead of directly optimizing (4.3) with respect to $\phi_{ss}(\tau,\omega)$ and $\boldsymbol{h}(\omega)$, we first estimate $\{\phi_{ss}(\tau,\omega)\}_{\tau,\omega}$ and $\{\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\}_{\omega}$. Using the estimated projection $\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]$, we can reconstruct $\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)$ by the low-rank matrix completion technique in Chapter 5. Finally, $\boldsymbol{h}(\omega)$ can be derived from $\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)$ through PCA.

In order to derive the update rules for estimation of $\{\phi_{ss}(\tau,\omega)\}_{\tau,\omega}$ and $\{\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\}_{\omega}$ in the first step, we differentiate (4.3) with respect to $\phi_{ss}(\tau,\omega)$ and $\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]^{*}$. The former leads to (4.4), and the latter gives

$$\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)] = \frac{\sum_{\tau}\phi_{ss}(\tau,\omega)\mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau,\omega)}{\sum_{\tau}\phi_{ss}^{2}(\tau,\omega)}. \tag{6.1}$$

Therefore, we estimate $\{\phi_{ss}(\tau,\omega)\}_{\tau,\omega}$ and $\{\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\}_{\omega}$ by iterating (4.4) and (6.1) alternately. $\{\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]\}_{\omega}$ is initialized using a rough estimation of $\boldsymbol{h}(\omega)$ by some of the conventional techniques. In the experiment in Section 6.3, this rough estimation is performed by Independent Vector Analysis (IVA) [73].

Once $\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]$ has been estimated, we estimate $\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)$ by employing the low-rank matrix completion algorithm in Section 5.1. Let us denote the estimate of $\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]$ by $\boldsymbol{Z}(\omega)$. We would like an estimate $\boldsymbol{W}(\omega)$ of $\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)$ whose rank is no more than 1, and whose projection $\mathcal{P}^{\perp}[\boldsymbol{W}](\omega)$ is close to $\boldsymbol{Z}(\omega)$. This can be formulated as (5.2) with $\hat{\boldsymbol{\Phi}}_{\boldsymbol{cc}}$ replaced by $\boldsymbol{W}$ and $\mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}]$ by $\boldsymbol{Z}(\omega)$ and with $R=1$. Therefore, $\boldsymbol{W}(\omega)$ can be estimated in the same way as in Section 5.1. $\boldsymbol{W}(\omega)$ can be initialized by the minimizer of

$$\sum_{\tau}\|\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)-\hat{\phi}_{ss}(\tau,\omega)\boldsymbol{W}(\omega)\|_{\mathrm{F}}^{2}, \tag{6.2}$$

where $\hat{\phi}_{ss}(\tau,\omega)$ is the estimate obtained in the previous step. The minimizer is given by

$$\frac{\sum_{\tau}\hat{\phi}_{ss}(\tau,\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)}{\sum_{\tau}\hat{\phi}_{ss}^{2}(\tau,\omega)}. \tag{6.3}$$

After that, $\boldsymbol{h}(\omega)$ is obtained as the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{W}(\omega)$. Since we have chosen the first microphone as a reference, $\boldsymbol{h}(\omega)$ should satisfy $h_1(\omega)=1$. Therefore, $\boldsymbol{h}(\omega)$ and $\phi_{ss}(\tau,\omega)$ are scaled accordingly.

The algorithm is summarized as follows:

Algorithm 3.

1. *Initialize* $\hat{\boldsymbol{h}}(\omega)$.
2. $\boldsymbol{Z}(\omega) \leftarrow \mathcal{P}^{\perp}[\hat{\boldsymbol{h}}(\omega)\hat{\boldsymbol{h}}^{H}(\omega)]$.
3. *Iterate the following for a preset time:*
   (a) $\hat{\phi}_{ss}(\tau,\omega) \leftarrow \dfrac{\langle \mathcal{P}_{\omega}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau,\omega), \boldsymbol{Z}(\omega)\rangle}{\|\boldsymbol{Z}(\omega)\|_F^2}$.
   (b) $\boldsymbol{Z}(\omega) \leftarrow \dfrac{\sum_{\tau}\hat{\phi}_{ss}(\tau,\omega)\mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau,\omega)}{\sum_{\tau}\hat{\phi}_{ss}^{2}(\tau,\omega)}$.
4. $\tilde{\boldsymbol{W}}(\omega) \leftarrow \dfrac{\sum_{\tau}\hat{\phi}_{ss}(\tau,\omega)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega)}{\sum_{\tau}\hat{\phi}_{ss}^{2}(\tau,\omega)}$.
5. *Iterate the following until*
$$\frac{\|\tilde{\boldsymbol{W}}(\omega)-\boldsymbol{W}(\omega)\|_F}{\|\boldsymbol{W}(\omega)\|_F} < \epsilon \tag{6.4}$$
is satisfied for a preset positive value ϵ*, or a preset maximum number of iterations is reached.*
   (a) $\boldsymbol{W}(\omega) \leftarrow \tilde{\boldsymbol{W}}(\omega)$.
   (b) $\boldsymbol{Y}(\omega) \leftarrow \mathcal{P}[\boldsymbol{W}(\omega)] + \boldsymbol{Z}(\omega)$.
   (c) *Update* $\boldsymbol{U}(\omega)$ *and* $\boldsymbol{\Sigma}(\omega)$ *according to the eigendecomposition of* $\boldsymbol{Y}(\omega)$*:*
$$\boldsymbol{Y}(\omega) = \boldsymbol{U}(\omega)\boldsymbol{\Sigma}(\omega)\boldsymbol{U}^{H}(\omega), \tag{6.5}$$
where $\boldsymbol{U}(\omega)$ *is unitary and* $\boldsymbol{\Sigma}(\omega)$ *is real-valued and diagonal, where the diagonal entries* $\sigma_1(\omega),\ldots,\sigma_M(\omega)$ *are arranged in decreasing order:* $\sigma_1(\omega)\geq\cdots\geq\sigma_M(\omega)$.

(d) $\tilde{\boldsymbol{W}}(\omega) \leftarrow \max\{\sigma_1(\omega), 0\}\boldsymbol{u}_1(\omega)\boldsymbol{u}_1^H(\omega)$, *where* $\boldsymbol{u}_1(\omega)$ *is the first column of* $\boldsymbol{U}(\omega)$.

6. $\hat{\boldsymbol{h}}(\omega) \leftarrow \dfrac{\boldsymbol{u}_1(\omega)}{u_{11}(\omega)}$. $\hat{\phi}_{ss}(\tau, \omega) \leftarrow \max\{\sigma_1(\omega), 0\}|u_{11}(\omega)|^2$.

6.2 Fabrication of a regular icosahedral array and recording of real-world data

We fabricated a 12-channel icosahedral microphone array with a diameter of 15 cm (see Fig. 6.1). The microphones were mounted on a pair of rigid hemispherical shells that were put together with screws. The array is equipped with female screws so that it can be fixed to a tripod. Electret-type omnidirectional microphones (SONY ECM-C10) were used.

We chose to mount the microphones on a rigid body instead of fixing them in the free field for several reasons. First, this facilitates the accurate fixture of the microphones at the vertices of the icosahedron. Second, it preserves the noise isotropy, whereas for a free-field array, the microphone holders or the wires would perhaps disturb it. Third, this design is also more beautiful.

Using this microphone array, we recorded a target speech and real-world noise to be used for evaluation in an experiment room at the University. In Fig. 6.2, we show the layout of the room, the microphone array, and the loudspeaker used in the experiment.

We used the following devices:

- 16-channel A/D board with microphone amplifiers (Tokyo Electron Device TD-BD-16ADUSB),
- computer
- 4-channel D/A board (M-AUDIO Fast track pro),
- loudspeaker amplifier (BOSE 1705II),
- loudspeaker (BOSE 101MM).

We recorded the signal and noise components separately, and subsequently added them together, because the signal component is needed as a reference in the evaluation. The target signal was a Japanese continuous speech utterance taken from the ATR Japanese speech database [62] and played from the loudspeaker. When recording the signal component, the

Figure 6.1 The fabricated 12-element icosahedral microphone array with the microphones fixed on a spherical shell rigid mount. The diameter is 15cm.

volume of the loudspeaker was set loud enough so that the other sounds from the environment become negligible. As the noise component, we recorded the environmental sound with the windows open. The recorded signal and noise components were downsampled from 48 kHz to 16 kHz, and added together with the appropriate scaling so that the SNR becomes 0 dB. The duration of the mixture was 10 s.

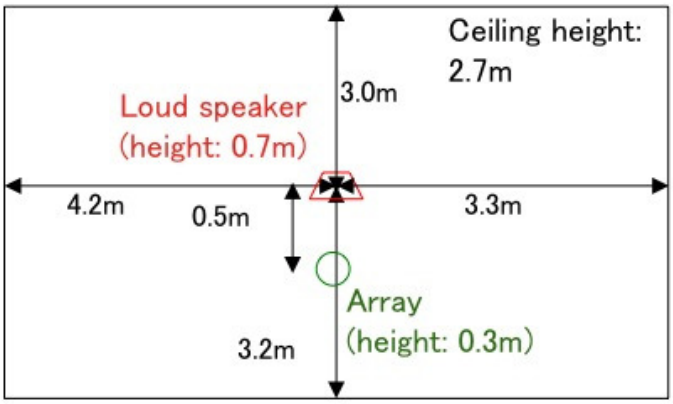


Figure 6.2 The layout of the room, the microphone array, and the loudspeaker in the experiment.

6.3 Real-world validation

6.3.1 Experimental conditions

We conducted an experiment to assess the method for diffuse noise suppression with an unknown steering vector proposed in Section 6.1. The following five methods are applied to the real-world data described in Section 6.2:

- a state-of-the-art blind source separation method called IVA [73] (baseline),
- the MVDR beamformer followed by the Wiener post-filter with the steering vector and the target power spectrogram estimated by the iterative method in Section 6.1 using the uncorrelated noise model, the fixed noise coherence model, the BND model, or the real-valued noise covariance model (denoted by blind-uncor, blind-coh, blind-BND, or blind-real).

The methods were applied to the first 8 s of the observed signals. The observed signal was first analyzed by STFT with a frame length of 2048, a frame shift of 64, and the Hamming window. We discarded the lower 14 frequency bins by setting them to zero, because their

Table 6.1 SNR (dB) of the observed signal, and the target signal estimated by IVA and by the proposed methods.

observed signal	IVA	blind-uncor	blind-coh	blind-BND	blind-real
−0.2	3.0	5.1	7.7	**10.1**	−1.2

SNRs were extremely low. The initial value of $\mathcal{P}^{\perp}[\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)]$ was calculated using $\boldsymbol{h}(\omega)$ derived from the separation matrix estimated by IVA. The observed covariance matrix for the calculation of the target steering vector and the target power spectrogram was computed locally by averaging over 48 consecutive frames. On the other hand, that for the MVDR beamformer was calculated by long-term averaging as in the experiment in Chapter 4. The observed signal was processed by the MVDR beamformer followed by the Wiener post-filter, designed using the estimated steering vector and the power spectrogram. The waveform of the estimated signal was obtained by the inverse STFT.

6.3.2 Experimental results

Table 6.1 shows the SNRs of the observed signal and the outputs of the compared methods for blind signal extraction. The ranking of the performance of the compared methods was as follows:

$$\text{blind-real} \prec (\text{observed signal}) \prec \text{IVA} \prec \text{blind-uncor} \prec \text{blind-coh} \prec \text{blind-BND} \tag{6.6}$$

An SNR enhancement of as much as 10.3 dB from the observed signal was gained by blind-BND.

We also proposed a blind signal extraction method in [8]. Whereas this method is based on the BND model and thus only applicable to the crystal arrays, the method proposed here is applicable to the general subspace noise covariance model. We examined the performance of the former under the same experimental conditions, and the output SNR was 10.3 dB, which is comparable to that of the unified method proposed here (10.1 dB).

Chapter 7

Conclusion

This thesis aimed to propose robust methods for microphone array signal processing against diffuse noise. We proposed a general noise model in the covariance matrix domain, and unified frameworks for the following three tasks based on it: diffuse noise suppression with a known target steering vector, DOA estimation of multiple sources in the presence of diffuse noise, and diffuse noise suppression with an unknown target steering vector. These are all applicable to the general noise model.

In Chapter 3, a unified framework for modeling the noise covariance matrix was proposed, which is based on the notion of linear spaces spanned by Hermitian matrices. We showed that the general model includes the following previous noise models as special cases: the spatially uncorrelated noise model, the fixed noise coherence model, and the BND model. Subsequently, our new more flexible real-valued noise covariance model was introduced. Compared to the previous noise models, it is applicable to an unknown arbitrary array geometry. Finally, these noise models were compared on a database of real-world noise. The real-valued noise covariance model fitted real-world noise best, but it has a significantly higher dimension than the other models. In comparison, the BND model reduced the dimensionality essentially without a significant loss in the fit. The fixed noise coherence model provided a reasonably good fit with only dimension 1, but failed when the microphones were mounted on a rigid mount, which affects the noise coherence. The spatially uncorrelated noise model did not fit the real-world noise well due to the small microphone spacing.

In Chapter 4, we described the application to diffuse noise suppression with a known target steering vector. Based on the general noise modeling in Chapter 3, we derived a general estimator of the target power spectrogram. It is based on the orthogonal projection of the

observed covariance matrix onto the orthogonal complement of the noise model subspace. We also derived the estimators for the specific noise models by applying this general estimator to these models. Finally, the Wiener post-filtering approach with the estimated target power spectrogram was evaluated through an experiment with real-world noise. The best performance was obtained with the proposed framework for the fixed noise coherence model, the BND model, and the real-valued noise covariance model. That is, when the target steering vector is given, these three noise models can be applied to diffuse noise suppression equally effectively. On the other hand, the proposed framework for the spatially uncorrelated noise model, which coincides with Zelinski's method under the experimental condition, resulted in the worst performance among the compared post-filtering methods for spatially correlated real-world noise. McCowan's method, which uses the same fixed noise coherence model as an above-mentioned method but a different estimation scheme, gave higher SNRs than the proposed framework for the spatially uncorrelated noise model, but lower SNRs than that for the other noise models.

In Chapter 5, we described the application to DOA estimation of multiple sources in diffuse noise. We proposed two algorithms based on different matrix completion algorithms: a first approach based on low-rank matrix completion, which uses knowledge of the rank of the signal covariance matrix, and a second one based on trace norm minimization, which does not require that knowledge. Finally, we evaluated and compared the proposed methods for different noise models using a large database we created with various values of the mixture parameters. The proposed trace norm minimization algorithm for the BND model worked best. As in noise suppression in Chapter 4, the spatially uncorrelated noise model did not work well with spatially correlated real-world noise. In addition, in this blind setting, the real-valued noise covariance model failed as well, which has many parameters relative to the number of observations. The estimation performance for the fixed noise coherence model was reasonably high, though lower than that of the trace norm minimization algorithm for the BND model. The BND model did not work well with the low-rank matrix completion algorithm, of which we are to investigate the reason.

In Chapter 6, we described the application to diffuse noise suppression with an unknown target steering vector and validation with real-world data. We proposed a unified method based on rank-1 matrix completion and PCA. We described an icosahedral microphone array we fabricated and the real-world data we recorded using it. We evaluated the proposed method with the recorded data. The proposed framework for the BND model resulted in

the best performance with an SNR enhancement of as much as 10.3 dB from the observed signal. The SNR for the fixed noise coherence model (7.7 dB) was higher than that for the spatially uncorrelated noise model (5.1 dB), which does not take noise correlation into account. The real-valued noise covariance model failed again in this blind scenario.

This work suggests a number of future research directions. In the short term, we plan to conduct additional experiments on the proposed techniques. In the medium term, we plan to investigate new noise models. Finally, in the long term, we would like to study new estimation algorithms.

The additional experiments planned in the short term include evaluation of noise suppression in terms of the ASR performance and evaluation of DOA estimation in terms of source tracking performance. ASR is among the most important applications of noise suppression techniques. We have validated the proposed noise suppression techniques in terms of criteria such as SNR, and we would like to assess these in terms of the ASR performance next. Source tracking is also important in practice, because real-world sound sources are often moving, not static. We plan to apply the proposed DOA estimation techniques to shorter data, and temporally integrate this short-time estimator via Hidden Markov Model (HMM) and/or particle filtering techniques. The theory is not novel, but we aim to achieve a performance breakthrough compared to existing techniques thanks to the use of a more accurate frame-by-frame estimator.

New noise models to be studied in the medium term include models designed for symmetrical arrays other than the crystal arrays and learned/adaptive models. The BND model for crystal arrays worked better than the real-valued noise covariance model for an unknown steering vector (Chapters 5 and 6). We can interpret this as follows: although the latter is applicable to the general array geometry, the former properly reduces the dimensionality by exploiting the symmetry of the geometry, thereby leading to the better performance in this blind setting. Since crystal arrays are not always available, it is beneficial to extend this approach to symmetrical arrays other than the crystal arrays. For example, the standard uniform linear array leads to a Toeplitz covariance matrix for isotropic noise. Although BND is inapplicable to this geometry, the general linear-space modeling enables the exploitation of this symmetry as well. On the other hand, in the proposed unified noise modeling framework, the design of the noise model boils down to the design of the basis of the model subspace. The basis vectors presented in this paper are all fixed, and a better performance would be achieved by leaning them from a database of real-world noise or making them data-adaptive.

In the long term, we would like to study new estimation algorithms such as one for diffuse noise suppression for multiple sources. The proposed noise suppression techniques assume a single target signal. Although experimental results suggested that the method for a known steering vector in Chapter 4 is robust against multiple directional sources to a certain degree, the method is not originally conceived for such scenario. Moreover, the blind method in Chapter 6 is not likely to work for the case of multiple directional sources. A technique for multiple directional sources would enable diffuse noise suppression robust against directional interferers as well.

Acknowledgement

I would like to express my sincere gratitude to a supervisor Professor Shigeki Sagayama (the University of Tokyo) for providing me this precious study opportunity as a Ph.D student in his laboratory. He gave me many useful big-picture suggestions, which guided my research direction. I also learned a lot from him about construction of papers and presentations.

I would like to express my deepest appreciation to a supervisor, Professor Nobutaka Ono (currently with National Institute of Informatics, Japan) for his elaborated guidance, considerable encouragement and invaluable discussion. Especially, his comments are always very insightful, and I learned a lot from his way of thinking.

I am very grateful to Dr. Rémi Gribonval and Dr. Emmanuel Vincent (both at INRIA Rennes - Bretagne Atlantique, France), who contributed to this thesis as official co-supervisors. Especially, I appreciate the precious opportunity of studying under their supervision during my stay at the METISS team of the institution for more than a year. Their contribution covers a large part of the thesis including the mathematical framework in Chapter 3, the DOA estimation algorithms in Chapter 5, the extension of Chapters 4 and 6 to the general noise model, structuring the thesis, extensive proofreading, thoughtful comments, and preparation for the defense. I also would like to thank Dr. Frédéric Bimbot (Team Leader, METISS, France; CNRS Research fellow, France) very much for welcoming me to his team.

I also appreciate variable comments given by Professor Shigeru Ando, Professor Shinji Hara, and Professor Hirokazu Kameoka (the University of Tokyo) at the pre-defense and the defense. These comments helped improve the thesis largely, especially the experiments in Section 3 and the algorithm in Section 6.

I am thankful to Yasuhisa-KOKI Co.,Ltd, Japan for its contribution to the fabrication of the icosahedral array.

This work is supported by Grant-in-Aid for JSPS Fellows 22-6927, Japan, Grant-in-Aid for Young Scientists (B) 21760309 from MEXT, Japan, INRIA under the Associate Team Program VERSAMUS, and JST CREST, Japan.

Bibliography

[1] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[2] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, no. 3, pp. 261–291, Apr. 1995.

[3] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

[4] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[5] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[6] H. Levitt, "Noise reduction in hearing aids: an overview," *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111–121, Jan./Feb. 2001.

[7] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2002.

[8] N. Ito, H. Shimizu, N. Ono, and S. Sagayama, "Diffuse noise suppression using crystal-shaped microphone arrays," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 19, no. 7, pp. 2101–2110, Sept. 2011.

[9] N. Ito, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Crystal-MUSIC: Accurate localization of multiple sources in diffuse noise environments using crystal-shaped microphone arrays," in *Proc. of LVA/ICA, Lecture Notes in Computer Science*, Saint-Malo, France, Sept. 2010, vol. 6365, pp. 81–88.

[10] N. Ito, N. Ono, and S. Sagayama, "Designing the Wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra," Mar. 2010, pp. 2818 – 2821.

[11] N. Ito, N. Ono, and S. Sagayama, "A blind noise decorrelation approach with crystal arrays on designing post-filters for diffuse noise suppression," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Las Vegas, USA, Apr. 2008, pp. 317–320.

[12] N. Ito, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multiple source localization based on matrix completion via trace norm minimization," in *Proc. of ASJ Spring Meeting*, Mar. 2011, pp. 665–666, (in Japanese).

[13] N. Ito, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Diffuse noise robust multiple source localization based on noise reduction in covariance matrix domain," in *IEICE Technical Report*, Dec. 2010, vol. 110, pp. 31–36, (in Japanese).

[14] N. Ito, Y. Kitano, N. Ono, and S. Sagayama, "Instrument separation in reverberant environments using crystal microphone arrays," in *IPSJ Technical Report*, Nov. 2009, vol. 2009-MUS-82, pp. 1–6, (in Japanese).

[15] N. Ito, N. Ono, and S. Sagayama, "Diffuse noise suppression based on the imaginary part of the inter-channel cross-spectrum with a small-sized microphone array," in *Proc. of ASJ Spring Meeting*, Mar. 2009, pp. 705–706, (in Japanese).

[16] N. Ito, N. Ono, and S. Sagayama, "Investigation of real-environmental noise suppression with a crystal array," in *Proc. of ASJ Autumn Meeting*, Sept. 2008, pp. 677–678, (in Japanese).

[17] N. Ito, N. Ono, and S. Sagayama, "Diffuse noise suppression by crystal-array-based post-filter design," in *IEICE Technical Report*, May 2008, vol. EA2008-13, SIP2008-22, pp. 43–46, (in Japanese).

[18] N. Ito, N. Ono, and S. Sagayama, "Diffuse noise suppression based on post-filtering using crystal arrays," in *Proc. of ASJ Spring Meeting*, Mar. 2008, pp. 695–696, (in Japanese).

[19] N. Ito, N. Ono, and S. Sagayama, "A nonlinear beamformer in the isotropic noise field using crystal arrays," in *Proc. of ASJ Autumn Meeting*, Sept. 2007, pp. 619–620, (in Japanese).

[20] N. Ito, "Microphone array signal processing for diffuse noise suppression," *Master's Thesis, The University of Tokyo*, Feb. 2009.

[21] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds., chapter 2, pp. 47–78. Springer, Netherlands, 2007.

[22] T. Nakatani, B. Juang, T. Yoshioka, K. Kinoshita, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, Nov. 2008.

[23] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[24] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice-Hall, Upper Saddle River, NJ, 1993.

[25] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, Nov. 1985.

[26] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, 2008.

[27] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, New York, Apr. 1988, pp. 2578–2581.

[28] K. U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," in *Second Cost 229 Workshop on Adaptive Algorithms in Communications*, Bordeaux, Oct. 1992, pp. 185–194.

[29] S. Fischer and K. U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Commun.*, vol. 20, no. 3–4, pp. 215–227, Dec. 1996.

[30] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., chapter 3, pp. 39–60. Springer-Verlag, Berlin, 2001.

[31] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Commun.*, vol. 34, no. 1–2, pp. 3–12, Apr. 2001.

[32] I. Cohen, "Multichannel post-filtering in nonstationary noise environments," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1149–1160, May 2004.

[33] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[34] J. Li and M. Akagi, "A noise reduction system based on hybrid noise estimation technique and post-filtering in arbitrary noise environments," *Speech Commun.*, vol. 48, no. 2, pp. 111–126, Feb. 2006.

[35] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Commun.*, vol. 49, no. 7–8, pp. 657–666, July–Aug. 2007.

[36] H. L. V. Trees, *Optimum Array Processing*, John Wiley & Sons, New York, 2002.

[37] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 240–259, May 1998.

[38] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson Jr., "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, Nov. 1955.

[39] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Commun.*, vol. 20, no. 3–4, pp. 229–240, Dec. 1996.

[40] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., chapter 4, pp. 61–85. Springer-Verlag, Berlin, 2001.

[41] H. Shimizu, "A theory of array signal processing to orthogonalize isotropic noise field and its application to power spectrum estimation," *Bachelor Thesis, The University of Tokyo*, Feb. 2007, (in Japanese).

[42] H. Shimizu, N. Ono, K. Matsumoto, and S. Sagayama, "Isotropic noise suppression in the power spectrum domain by symmetric microphone arrays," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, Oct. 2007, pp. 54–57.

[43] H. Shimizu, K. Matsumoto, N. Ono, and S. Sagayama, "A theory of array signal processing to orthogonalize isotropic noise field and its application to power spectrum estimation," in *Proc. of ASJ Spring Meeting*, Mar. 2007, pp. 569–570, (in Japanese).

[44] G. A. F. Seber, *A Matrix Handbook for Statisticians*, John Wiley & Sons, Inc., New Jersey, 2008.

[45] N. Ono, N. Ito, and S. Sagayama, "Five classes of crystal arrays for blind decorrelation of diffuse noise," in *Proc. SAM*, Darmstadt, Germany, July 2008, pp. 151–154.

[46] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2149–2157, 2004.

[47] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[48] B. Rafaely, "Phase-mode versus delay-and-sum spherical microphone array processing," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 713–716, Oct. 2005.

[49] M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 3094–3103, Nov. 2005.

[50] L. Zhiyun and D. Ramani, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 15, no. 2, pp. 702–714, Feb. 2007.

[51] H. Sun, S. Yan, and U. P. Svensson, "Robust spherical microphone array beamforming with multi-beam-multi-null steering and sidelobe control," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, Oct. 2009, pp. 113–116.

[52] S. Ando, H. Shinoda, K. Ogawa, and S. Mitsuyama, "A three-dimensional sound localization sensor system based on the spatio-temporal gradient method," *Transactions of the Society of Instrument and Control Engineers*, vol. 29, no. 5, pp. 520–528, 1993, (in Japanese).

[53] S. Ando, "Intelligent three-dimensional vision sensor with ears," *Sensors and Materials*, vol. 7, no. 3, pp. 213–231, 1995.

[54] S. Koyama, T. Kurihara, and S. Ando, "A theory and experiment of instantaneous wave source localization from a wave distribution on a small region," *IEEJ Trans. SM*, vol. 129, no. 10, pp. 350–356, 2009, (in Japanese).

[55] W. Foy, "Position-location solutions by Taylor-series estimation," *IEEE Trans. Aerospace Electro. Sys.*, vol. 12, no. 2, pp. 187–194, Mar. 1976.

[56] J. Smith and J. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661–1669, Dec. 1987.

[57] P. Stoica and J. Li, "Source localization from range-difference measurements," *IEEE. Signal Process. Magazine*, vol. 23, no. 6, pp. 63–69, Nov. 2006.

[58] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[59] M.Wax, T. Shan, and T. Kailath, "Spatio-temporal spectral analysis by eignestructure methods," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 4, pp. 817–827, Aug. 1984.

[60] T. Pham and B. M. Sadler, "Adaptive wideband aeroacoustic array processing," in *Proc. IEEE Signal Process. Workshop on Statist. Sig. Array Process.*, 1996, pp. 295–298.

[61] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, July 2006.

[62] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Commun.*, vol. 9, no. 4, pp. 357–363, Aug. 1990.

[63] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proc. ICML*. AAAI Press, 2003, pp. 720–727.

[64] K. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," Sept. 2010, vol. 6, pp. 615–640.

[65] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *The Journal of the Society for the Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, Apr. 2009.

[66] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proceedings of the 26th International Conference on Machine Learning*, pp. 457–464.

[67] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[68] E. Vincent and D. R. Campbell, "Roomsimove," (2010, Nov. 29). [Online]. Available: http://www.irisa.fr/metiss/members/evincent/software.

[69] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, doi: 10.1016/j.sigpro.2011.09.032.

[70] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[71] J. Benesty, J. Chen, and Y. A. Huang, "A minimum speech distortion multichannel algorithm for noise reduction," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Las Vegas, USA, 2008, pp. 321–324.

[72] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sept. 2002.

[73] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, Oct. 2011.