博士論文（要約）

波動場の特性を陽に利用した ブラインド音源分離

(Blind Source Separation Exploiting Characteristics of Wave Fields)

光藤 祐基

Yuki Mitsufuji

Blind Source Separation Exploiting Characteristics of Wave Fields

概要

Owing to recent progress in sound field reconstruction for captured sound fields with many microphones, sound source separation for a huge number of microphones has become an interesting topic and is, hence, currently actively discussed. In blind source separation (BSS), which does not require previously obtained knowledge, non-negative matrix factorization (NMF) has been studied for several decades. Multichannel NMF (MNMF), an extension of NMF to multiple microphones, has been studied under the assumption of using 2 to 4 channels, but it has rarely been used for signals stemming from a microphone array with 32 channels or more, which is common in sound field reproduction. One of the main difficulties for MNMF is its calculation cost as it scales with $O(A^3)$ where the number of microphones is A. When the spatial covariance matrix (SCM) is updated, an eigenvalue decomposition is performed, resulting in a high computational cost. A number of BSS methods with reduced computational complexity have been proposed so far. Meanwhile, in the field of sound field reproduction, conversion to the spatial frequency domain is often used to improve calculation efficiency. Furthermore, converting time-frequency (TF) signals to the spatial frequency domain makes it possible to identify TF bins that are not required for the sound field reconstruction and that can be excluded from the calculation process. In this thesis, we will solve the problem of the calculation cost for MNMF by introducing the conversion to the spatial frequency used in sound field reproduction, and at the same time, we attempt to achieve performance improvements by explicitly exploiting the characteristics of the wave fields. We propose a unified BSS method using wave field characteristics through conversion to either (i) a signal representation in the wavenumber domain obtained by solving the wave equation in a Cartesian coordinate system or (ii) conversion to a signal representation in the spherical harmonic domain obtained by solving the wave equation in a spherical coordinate system. Specifically, we propose a method to reduce the computational complexity from $O(A^3)$ to $O(A)$ by diagonalizing the SCM included in the MNMF model using a spatial frequency transformation matrix to solve the problem equivalent to the non-negative tensor factorization (NTF). Furthermore, in consideration of not only the use in 32 channels as used in sound field reproduction but also generalization in a small number of microphones such as 2–4 channels, a model using a tridiagonal SCM is proposed and its effectiveness is verified. Finally, we propose a method that can be used in the important setup of using a spherical microphone array and verify its effectiveness.

Keywords Blind Source Separation, Spatial Covariance Matrix, Non-negative Matrix Factorization, Non-negative Tensor Factorization, Wavenumber Domain, Spherical Harmonic Domain

Abstract

近年の音場再現技術の進歩により、膨大な数のマイクロホンによる音源分離が注目を集めるようになり、実現のための議論が活発になされている。事前に得た情報を必要としないブラインド音源分離 (BSS) では非負行列因子分解 (NMF) による手法が長く研究されてきた。複数マイクの拡張である多チャネル NMF (MNMF) は、2–4ch を仮定した状況を題材に研究されてきたが、音場再現で使用されるような 32ch 以上のマイクアレイで得られる信号に用いられることはなかった。主な要因の一つに計算コストが挙げられる。MNMF で必要となる計算コストは、マイクロホンの数が A のとき $O(A^3)$ にも及ぶ。空間相関行列 (SCM) を更新する場合は、更新式に固有値分解が行われるため、非常にコストが大きい。これまでに計算量を削減した数々の BSS の手法が提案されてきた。一方で、音場再現の分野では空間周波数領域への変換を、計算効率の向上のために用いられることが多い。また、空間周波数に変換することで、音場再現に不必要な空間周波数の特定や、計算からの除外を行うことが可能となる。本研究では、音場再現で用いられる空間周波数への変換を、MNMF に導入することで、計算コストの課題を解決するとともに、波動場の特性を陽に利用することで性能向上を達成することに挑戦する。以下の章では、デカルト座標系の波動方程式を解くと得られる波数領域における信号表現への変換、および球座標系の波動方程式を解くと得られる球面調和領域における信号表現への変換を通じて、波動場特性を利用した統一的な BSS の手法を提案する。具体的には、まず MNMF のモデルが含む SCM に対して空間周波数変換行列を用いて対角化することで、非負値テンソル分解 (NTF) と等価な問題を解くことに持ち込み計算量を $O(A)$ に抑える手法を提案する。次に、音場再現で利用されるような 32ch における使用だけではなく、2–4ch 等の少ないマイクロホン数における一般化も考慮し、三重対角成分までを使用したモデルの提案とその有効性を実証する。最後に、球状マイクロホンアレイを用いたユースケースにおいて、実践的に使用可能な手法を提案し、その有効性を検証する。

Keywords Blind Source Separation, Spatial Covariance Matrix, Non-negative Matrix Factorization, Non-negative Tensor Factorization, Wavenumber Domain, Spherical Harmonic Domain

Contents

Chapter 1

Introduction

1.1 Background and Motivation

Owing to recent progress in sound field reconstruction for captured sound fields with many microphones, sound source separation for a huge number of microphones has become an interesting topic and is, hence, currently actively discussed. In blind source separation (BSS) *[1], which does not require previously obtained knowledge, non-negative matrix factorization (NMF) has been studied for several decades. Fig. 1.1 shows the relationship between BSS and sound field reconstruction from the early 2000s up to now. Multichannel NMF (MNMF), an extension of NMF to multiple microphones, has been studied since the early 2010s under the assumption of using 2 to 4 channels, but it has rarely been used for signals stemming from a microphone array with 32 channels or more, which is common in sound field reproduction. One of the main difficulties for MNMF is its calculation cost as it scales with $O(A^3)$ where the number of microphones is A. When the spatial covariance matrix (SCM) is updated, an eigenvalue decomposition is performed, resulting in a high computational cost. A number of BSS methods with reduced computational complexity have been proposed so far.

Meanwhile, in the field of sound field reproduction, conversion to the spatial frequency domain is often used to improve calculation efficiency. Furthermore, converting time-frequency (TF) signals to the spatial frequency domain makes it possible to identify TF bins that are not required for the sound field reconstruction and that can be excluded from the calculation process. For example, when a spherical microphone array is used as a recording device, e.g., the 32-channel Eigenmike®, a spherical harmonic (SH) transform is applied and only a few SH coefficients are often used for reproducing the captured sound field. This framework is called higher-order ambisonics (HOA) and has become a part of the Moving Picture Experts Group (MPEG)-H 3D Audio format [1]. Although, BSS and sound field reconstruction have long been studied assuming multichannel microphones, there was no work until 2014 that bridges these two independent fields.

In this thesis, we solve the problem of the huge calculation cost for MNMF by introducing the conversion to the spatial frequency used in sound field reproduction, and at the same time, we attempt to achieve performance improvements by explicitly exploiting the characteristics of wave fields. We propose a unified BSS method using wave field characteristics through conversion to either (i) a signal representation in the wavenumber domain obtained by solving the wave equation in a Cartesian coordinate system or (ii) conversion to a signal representation in the SH domain obtained by solving the wave equation in a spherical coordinate system. As a consequence, the following improvements

*[1] In this thesis, BSS refers to unsupervised source separation with multi-channel signals without any knowledge about the sources and the mixing process except the employed microphone geometry.

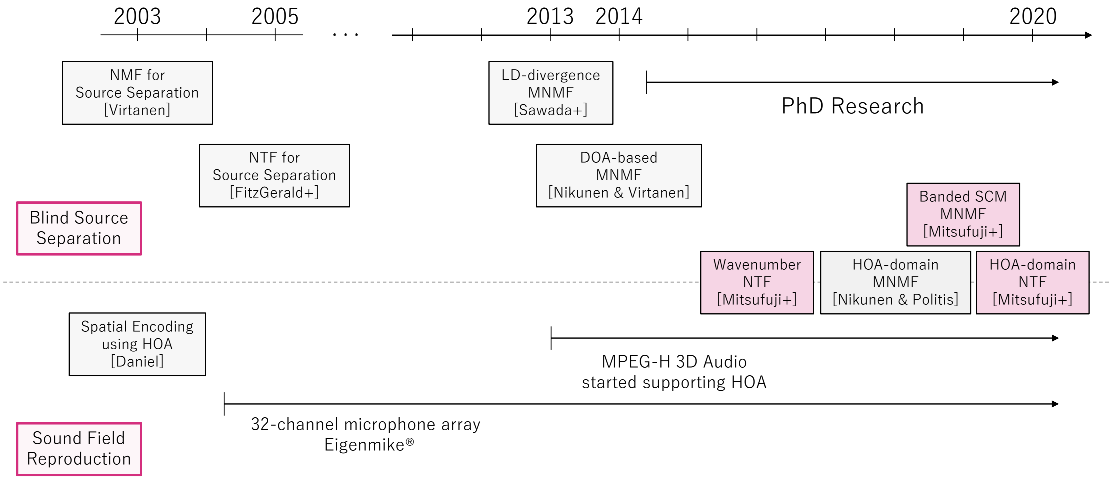

2003
2005
. . .
2013
2014
2020
NMF for Source Separation [Virtanen]
NTF for Source Separation [FitzGerald+]
LD-divergence MNMF [Sawada+]
DOA-based MNMF [Nikunen & Virtanen]
PhD Research
Banded SCM MNMF [Mitsufuji+]
Wavenumber NTF [Mitsufuji+]
HOA-domain MNMF [Nikunen & Politis]
HOA-domain NTF [Mitsufuji+]
Blind Source Separation
Spatial Encoding using HOA [Daniel]
MPEG-H 3D Audio started supporting HOA
32-channel microphone array Eigenmike®
Sound Field Reproduction

Fig. 1.1. Relationship between blind source separation and sound field reproduction. Contributions of thesis are highlighted in red.

can be expected:

- the improvement of separation performance by converting TF signals to sparse domains or exploiting the characteristics of wave fields,
- the improvement of computation speed by simplifying the update rules of NMF parameters achieved by a diagonal or tridiagonal approximation of SCM.

From the following sections, we briefly explain the history of BSS and recent advances in the MNMF approaches, which helps readers understand our motivation and the underlying issues of related works.

1.2 Blind Source Separation

Source separation is a key technology that has enabled major breakthroughs in various fields of audio signal processing, such as automatic speech recognition and music transcription. The techniques for solving the source separation problem can be classified into two groups: non-blind methods, e.g., beamforming [2, 3] and BSS. The separation quality can generally be enhanced by incorporating prior knowledge, which is mainly provided as a property of the target source. For example, the incorporation of a spatial cue has the potential to improve applications such as Sound Zoom by enabling the extraction of target sounds that might be located in a certain direction. In the 1990s, a new BSS technique called independent component analysis (ICA) was developed that automatically finds the directions of the sources in a mixture, thereby enabling the extraction of a target source [4, 5]. One limitation of an ICA-based method is that it cannot handle the under-determined case where the number of sources is higher than the number of microphones A. Another approach, called the degenerate unmixing estimation technique (DUET) [6, 7, 8], which is capable of de-mixing from 2ch-stereo signals, requires the sources to be sparse with respect to the spectrogram bins [9, 10]. In recent years, owing to the advent of deep learning, non-blind methods based on training spectro-temporal information of audio signals have been proven to yield notable results [11, 12, 13, 14, 15, 16]. In contrast, BSS, where training data are not available, still remains a challenging and open problem.

1.3 Approaches Based on Matrix Factorization

In the last few decades, NMF [17, 18, 19, 20] has become one of the most prevalent techniques to tackle the underdetermined source separation problem where the number of sources is greater than or equal to the number of observations. NMF is based on the idea that a mixture is a composite of a number of object basis elements, each of which represents an underlying characteristic of the sources. Estimation is carried out by simple matrix factorization, with all the elements being non-negative. NMF also eliminates another problem with ICA, namely, that the number of microphones should be greater than or equal to the number of target sources. A large number of related techniques have been developed so far [21, 22, 23, 24, 25, 26].

A cost function for NMF estimation has long been investigated by several researchers [17, 27, 28]. In particular, the Itakura–Saito (IS) divergence is known to be an appropriate cost function for approximation of audio spectra due to its scale-invariant nature. To complete NMF-based source separation, clustering of the decomposed basis elements follows the factorization to properly classify them to corresponding sources.

In the last decade, the local Gaussian model has gained much attention as one of the most promising unsupervised approaches exploiting multichannel coherence. In 2005, the local Gaussian model was first applied to multichannel source separation [29, 30], in which the spectrum of each TF bin is modeled as an instantaneous mixture of complex multivariate Gaussians. To cope with more complex mixing conditions such as convolutive environments, the model was further extended to incorporate a full-rank model, and a generalized expectation–maximization (GEM) algorithm was employed to derive the update rules to obtain the model parameters [31, 32].

Ozerov and Févotte applied a low-rank factorization in this framework for modeling source amplitudes of TF bins [33]. Their approach can be regarded as the multichannel extension of NMF [34]. This original approach was limited to a rank-1 matrix, which was later generalized to the full-rank case so that the algorithm can also be used under reverberant conditions [35, 36, 37]. The separation quality of MNMF was highlighted in the literature as outperforming other existing methods, such as l_1-norm minimization [38], l_p-norm minimization [39], and binary clustering [40].

Its huge computational cost and slow convergence were identified as two major problems to be tackled in the future. In detail, the computational cost explodes with increasing the number of microphones, A, as $O(A^3)$ owing to the multiple matrix inversions during the parameter [41]. For the convergence of MNMF, GEM-based parameter updates were shown to be much slower than multiplicative updates by comparison with non-negative tensor factorization (NTF) [6]. The convergence speed of MNMF was increased by incorporating multiplicative updates into the M-step of the source parameter updates [42]. A different update method was proposed in [43, 44] consisting solely of multiplicative updates with the MM algorithm [45]. While the convergence problem was mitigated by such means, the authors reported that the separation performance was prone to local minima resulting from the initialization of the model parameters.

Nikunen and Virtanen developed a DOA-based method to overcome the initialization problem [46]. The algorithm provides a series of DOA kernels enabling the spatial properties observed in the multichannel signals to be encoded. A DOA kernel is composed of outer products of steering vectors. The weights of the DOA kernels are obtained by multiplicative update rules. In [46], it was reported that the method yielded robust results even in the case of initializing parameters with random values. To reduce the number of parameters to be updated, the DOA-based method was further enhanced by splitting a

DOA kernel into two parts: a fixed kernel consisting of phase covariances and an updatable kernel consisting of amplitude covariances [47].

1.4 Approaches Based on Tensor Factorization

As a generalized NMF technique, NTF extends the NMF idea to tensors. An n-way tensor is a generalization of the mathematical concepts of scalar, vector, and matrix (e.g., a two-way tensor is a matrix). Specifically, a three-way tensor, which can be regarded as a collection of multichannel spectrograms, is now being investigated for use in NTF [48, 49, 50, 51, 52, 53, 54]. Extension to the third dimension provides another matrix that describes the energy distribution of each base component on every channel, which can also be regarded as spatial information. NTF can also be regarded as a simplified approach to MNMF because the model extracts the diagonal part of a SCM while discarding the off-diagonal part that contains information of interchannel phase differences. Since NTF assumes that the original sources are mixed instantaneously, exploiting only the diagonal part is often not sufficient to model more realistic mixing conditions. However, thanks to its low computational cost of order $O(A)$ [6, 53] per TF bin, NTF is being rigorously investigated for many types of applications [49, 55], and a number of variants have been proposed [52, 56]. In addition to NTF modeling of the diagonal part, an NMF-based treatment of off-diagonal elements has recently been proposed [57].

1.5 Incorporation of Prior Knowledge

Taking advantage of prior knowledge for the purpose of enhancing the performance of NMF has been widely investigated. Smaragdis et al. have attempted to make use of a user-guided humming for the extraction of melodies in a mixture [58]. Diknen et al. investigated the Bayesian NMF model assigning different prior distributions for tonal and percussive signals [59]. Ewert et al. presented an extended approach that uses additional score information to guide the NMF process [60]. Since NMF which separates a single-channel signal does not produce any spatial information during the separation process, it has been difficult to associate a spatial cue until the emergence of MNMF and NTF. The NTF extension enables the NMF approach to accept a spatial cue [61]. In addition, since a spatial cue indicates which bins of the tensor spectrogram are important, it is possible to improve the quality of an approximation to the specific bins of the tensor by giving more weight to bins where the target is likely to exist and less weight to the others [62]. The text in [62] can be found in Appendix.

1.6 Applications to Source Detection

In source detection problems, some target sound events of a given class need to be identified in a complex mix. A number of source detection methods based on NMF have been proposed so far. Weninger et al. incorporate NMF-based sound event detection into a speech recognition framework to remove non-linguistic events from speech recordings [63]. There are several other sound event detection systems based on NMF, customized for realistic situations in multi-source environments [64, 65]. When dealing with detection problems in multichannel signals, NTF presents a significant advantage over NMF after downmixing, because it fully exploits the lower masking of targets within the individual channels. An extension to enhance detection results of NMF is to incorporate modeling of temporal-spectral pattern as a basis, as done with non-negative matrix deconvolution (NMD, [66, 67, 68]). This approach has been extended to tensors in [69, 70], with a

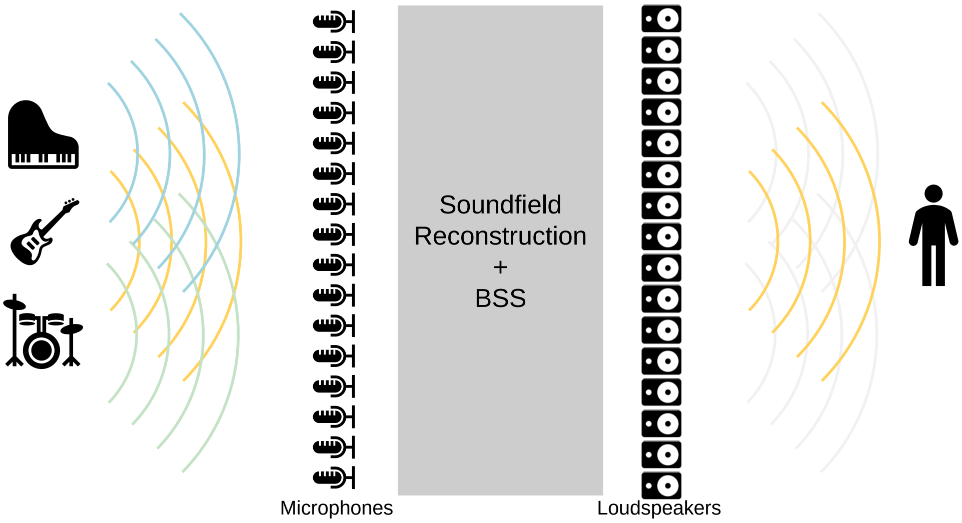


Fig. 1.2. General concept of this thesis

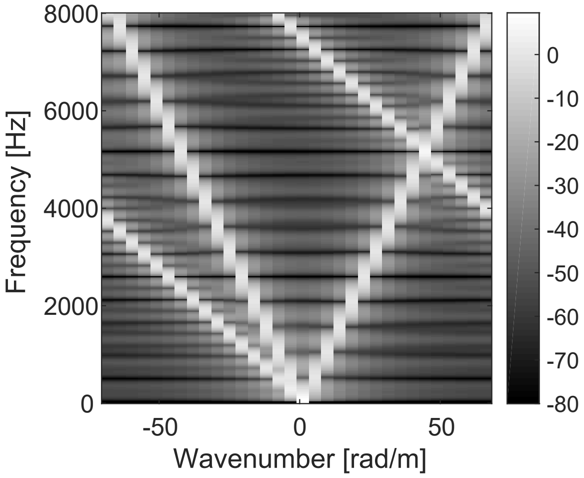


Fig. 1.3. Virtual plane waves represented in the wavenumber domain, originating from $\pi/20$, $8\pi/20$, and $14\pi/20$.

different formalism, leading to non-negative tensor deconvolution (NTD).

1.7 Relation to Sound Field Reconstruction

In the field of sound field reconstruction, in which a large number of microphones and loudspeakers are used, signal representation in the spatial frequency domain is regarded as an essential technique for reducing computational cost [71, 72]. This transformation allows essential information to be compressed and computational complexity to be reduced from order $O(A^3)$ to $O(A)$. Koyama et al. proposed a MAP estimation method to derive both spatial basis components and their weights, given the position of the primary source, so that spherical waves could be modeled with less spatial aliasing [73]. However, unlike NMF, the method does not take into account the source properties. In this thesis, we assume many microphones and loudspeakers are used as shown in Fig. 1.2, and the idea

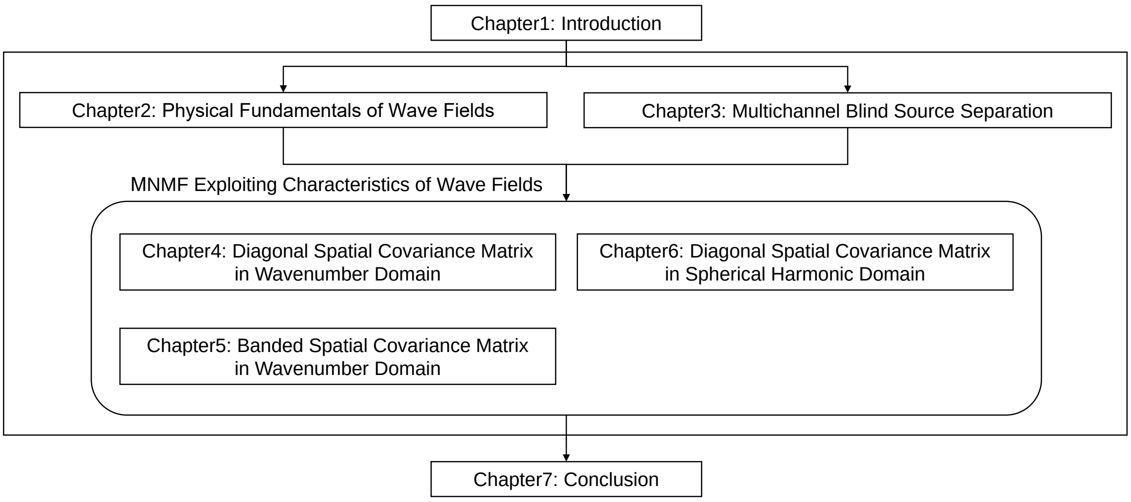

Chapter1: Introduction
Chapter2: Physical Fundamentals of Wave Fields
Chapter3: Multichannel Blind Source Separation
MNMF Exploiting Characteristics of Wave Fields
Chapter4: Diagonal Spatial Covariance Matrix in Wavenumber Domain
Chapter6: Diagonal Spatial Covariance Matrix in Spherical Harmonic Domain
Chapter5: Banded Spatial Covariance Matrix in Wavenumber Domain
Chapter7: Conclusion

Fig. 1.4. Outline of this thesis

of using a spatial transform was applied to MNMF. Fig. 1.3 shows plane waves in a wavenumber domain, originating from three different directions. It is clear that the plane waves can be represented with little overlapping of the spectrograms, which is a great advantage for NMF-based BSS. Furthermore, we can derive an approximated modeling for SCM in the spatial domain, showing that SCM can be diagonalized and its inverse is efficiently calculated. This leads to faster implementation of NMF-based multichannel BSS with a large number of microphones, compared to the conventional MNMF.

1.8 Outline and Notations

The organization of this thesis is given in Fig. 1.4. In Chapter 2, mathematical representations of a sound field is presented as preliminaries. In Chapter 3, MNMF and its variant are introduced. In Chapter 4, to tackle the problem of MNMF, we introduce NTF in the wavenumber domain, which reduces the cost to the order $O(A)$. It transforms microphone signals into the wavenumber domain, a technique that is commonly used for soundfield reconstruction. In Chapter 5, we introduce a more robust algorithm approximating the tridiagonal part of the spatial covariance matrix, which requires a complexity of $O(A^2)$ for the inversion by applying the Thomas algorithm. To remove the ad-hoc integration of post clustering after the decomposition, we also examine a self-clustering algorithm. In Chapter 6, we model near-field sources by estimating the model parameters of NTF in the SH domain assuming that microphone signals can be obtained with a rigid spherical array and converted to the SH domain. We also examine a masking scheme to exclude noisy regions in the SH domain from the NTF cost function. Finally, Chapter 7 concludes this thesis.

The following notations are used throughout this thesis: $\mathbf{x}$ denotes a column vector and $\mathbf{X}$ a matrix, where $\mathbf{I}$ is the identity matrix. The conjugate operator, trace, determinant, matrix transpose, conjugate matrix transpose, Moore–Penrose pseudoinverse, Euclidean vector norm, and Frobenius matrix norm are denoted by $(.)^*$, $\mathrm{tr}(.)$, $\det(.)$, $(.)^T$, $(.)^H$,

$(.)^\dagger$, $\|.\|$, and $\|.\|_F$, respectively. $\mathbf{X} \succ \mathbf{0}$, $\mathbf{X} \succeq \mathbf{0}$ means that $\mathbf{X}$ is symmetric and positive definite / semi-definite. Furthermore, $\text{tridiag}\{\mathbf{X}\}$ returns a matrix of the same size as $\mathbf{X}$ that contains the tridiagonal part of $\mathbf{X}$.

Chapter 2

Physical Fundamentals of Wave Fields

2.1 Wave Equation in Cartesian Coordinate System

In the following, we briefly summarize the wave equation in Cartesian coordinates and its solution. Interested readers are referred to [74] for more detailed a treatment.

Let $p(x, y, z, \tau)$ be the sound pressure that satisfies the acoustic wave equation

$$\nabla^2 p - \frac{1}{c^2}\frac{\partial^2 p}{\partial \tau^2} = 0, \tag{2.1}$$

assuming that the sound pressure has an infinitesimal deviation from its equilibrium value. In (2.1), c denotes the speed of sound which is, e.g., $c = 343$ m/s in air and $c = 1481$ m/s in water at 20°. The right side of the equation shows that there is no sound source in the region when the equation holds. ∇^2 denotes the Laplace operator and in Cartesian coordinates, it is given by,

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{2.2}$$

A Fourier transform on the wave equation leads to the Helmholtz equation

$$\nabla^2 P + \bar{k}^2 P = 0, \tag{2.3}$$

where $\bar{k} = \omega/c$ denotes the wavenumber of the sound, the frequency is given by $2\pi f = \omega$, and P is a function of (x, y, z, ω). The general solution of the Helmholtz equation in three dimensions is given by,

$$p(\omega) = \rho(\omega)e^{\mathrm{i}(\bar{k}_x x + \bar{k}_y y + \bar{k}_z z)} \tag{2.4}$$

where $\rho(\omega)$ and i denote an arbitrary constant and the imaginary unit, respectively. This equation satisfies the above equation in the case of $\bar{k}^2 = \bar{k}_x^2 + \bar{k}_y^2 + \bar{k}_z^2$. Since $\bar{k}$ is a constant, the three wavenumbers are not independent. The coefficients in the wavenumber domain, $P(\bar{k}_x, y, z, \tau)$, can be obtained by the Fourier transform in the x axis using the orthogonality of $e^{-\mathrm{i}\bar{k}_x x}$:

$$P(\bar{k}_x, y, z, \tau) = \int_{-\infty}^{\infty} p(x, y, z, \tau)e^{-\mathrm{i}\bar{k}_x x}dx. \tag{2.5}$$

2.2 Wave Equation in Spherical Coordinate System

Let us now consider the wave equation in spherical coordinates. Again, interested readers are referred to [74] for a more detailed treatment.

The wave equation including the time dependent term in the spherical coordinate system is given below:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial p}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial p}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 p}{\partial\phi^2}-\frac{1}{c^2}\frac{\partial^2 p}{\partial\tau^2}=0. \tag{2.6}$$

where r, ϕ, and θ denotes the radius, the azimuth, and the elevation, respectively. The separation of variables $p(r,\theta,\phi,\tau)=\mathbb{R}(r)\Theta(\theta)\Phi(\phi)\mathbb{T}(\tau)$ leads to four ordinary differential equations:

$$\frac{d^2\Phi}{d\phi^2}+m^2\Phi=0, \tag{2.7}$$

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)+[n(n+1)-\frac{m^2}{\sin^2\theta}]\Theta=0, \tag{2.8}$$

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\mathbb{R}}{dr}\right)+\bar{k}^2\mathbb{R}-\frac{n(n+1)}{r^2}\mathbb{R}=0, \tag{2.9}$$

$$\frac{1}{c^2}\frac{d^2\mathbb{T}}{d\tau^2}+\bar{k}^2\mathbb{T}=0. \tag{2.10}$$

The solutions to (2.10) and (2.7) are given, respectively by,

$$\mathbb{T}(\omega)=\mathbb{T}_1e^{\mathrm{i}\omega\tau}+\mathbb{T}_2e^{-\mathrm{i}\omega\tau}, \tag{2.11}$$

$$\Phi(\phi)=\Phi_1e^{im\phi}+\Phi_2e^{-im\phi}. \tag{2.12}$$

The orthogonality of the azimuth function can be expressed by the following equation:

$$\int_0^{2\pi}e^{-\mathrm{i}m'\phi}e^{\mathrm{i}m\phi}d\phi=2\pi\delta_{mm'}, \tag{2.13}$$

where $\delta_{mm'}$ denotes the Kronecker delta function defined as

$$\delta_{mm'}=\begin{cases}1\ (m=m')\\0\ (m\neq m')\end{cases}. \tag{2.14}$$

The solution of (2.8) is obtained by transforming the variables. With $\eta=\cos\theta$ (i.e., $-1\leq\eta\leq1$), the differential equation for Θ is

$$\frac{d}{d\eta}\left[(1-\eta^2)\frac{d\Theta}{d\eta}\right]+\left[n(n+1)-\frac{m^2}{1-\eta^2}\right]\Theta=0. \tag{2.15}$$

The solution of the above equation is given by Legendre functions of the first and second kind, and can be written as

$$\Theta(\theta)=\Theta_1P_{nm}(\cos\theta)+\Theta_2Q_{nm}(\cos\theta). \tag{2.16}$$

Since the function of the second kind, Q_{nm}, diverges at the pole of $\eta=\pm1$, this solution is not adopted ($\Theta_2=0$). The first kind of the function diverges at $\cos\theta=1$ (i.e.,$\theta=0$) unless n is limited to an integer. Furthermore, when n is an integer, $P_{nm}(\eta)=0$ in $m>n$.

Finally, the solution of the differential equation (2.9) in the radial direction is given by,

$$\mathbb{R}(r)=\mathbb{R}_1j_n(\bar{k}r)+\mathbb{R}_2y_n(\bar{k}r), \tag{2.17}$$

where j_n and y_n are spherical Bessel functions of the first and second kind, respectively. This solution can also be expressed as follows:

$$\mathbb{R}(r)=\mathbb{R}_3h_n^{(1)}(\bar{k}r)+\mathbb{R}_4h_n^{(2)}(\bar{k}r), \tag{2.18}$$

where $h_n^{(1)}$ and $h_n^{(2)}$ are spherical Hankel functions of the first and second kind, respectively.

2.3 Signal Representation in Spherical Harmonic Domain

An interior sound field in spherical coordinates can be expressed as a weighted sum of spherical Bessel functions $j_n(\cdot)$ and SH functions, i.e., $Y_{nm}(\cdot)$:

$$X(\bar{k}, r, \Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_{nm}(\bar{k}) j_n(\bar{k} r) Y_{nm}(\Omega), \tag{2.19}$$

where $X(\bar{k}, r, \Omega)$ denotes an arbitrary narrowband sound field at radius r and angle Ω. The weight of the basis functions $\alpha_{nm}(\bar{k})$ denotes the interior SH coefficient. The SH function is defined with $\Omega = \{\theta, \phi\}$ [74] as

$$Y_{nm}(\Omega) \equiv \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_{nm}(\cos\theta) e^{\mathrm{i}m\phi}, \tag{2.20}$$

Given a sound field on the boundary of a sphere, the SH coefficient $\alpha_{nm}(\bar{k})$ can be obtained by exploiting the orthonormality of the SH function:

$$\alpha_{nm}(\bar{k}) = \frac{1}{j_n(\bar{k} R_{\mathrm{b}})} \int_{\Omega} X(\bar{k}, R_{\mathrm{b}}, \Omega) Y_{nm}^{*}(\Omega) d\Omega, \tag{2.21}$$

with

$$\int_{\Omega} Y_{nm}(\Omega) Y_{n'm'}^{*}(\Omega) d\Omega = \delta_{nn'} \delta_{mm'}, \tag{2.22}$$

where R_{b} denotes the radius of the boundary.

2.4 Evanescent Waves

In the Cartesian coordinate system, the wavenumber $\bar{k}$ fulfills the relationship that we introduced in Sec. 2.1, i.e., $\bar{k}^2 = \bar{k}_x^2 + \bar{k}_y^2 + \bar{k}_z^2$, and the equation can be rearranged in the following way assuming that $\bar{k}$ is non-negative:

$$\bar{k}_x = \begin{cases} \pm\sqrt{\bar{k}^2 - \bar{k}_y^2 - \bar{k}_z^2} & \text{for } \bar{k}^2 \geq \bar{k}_y^2 + \bar{k}_z^2 \\ \pm\mathrm{i}\sqrt{\bar{k}_y^2 + \bar{k}_z^2 - \bar{k}^2} & \text{for } \bar{k}_y^2 + \bar{k}_z^2 \geq \bar{k}^2 \end{cases}. \tag{2.23}$$

The first case in the above equation represents propagating waves while the second case represents evanescent waves. An evanescent wave decays exponentially and this property can be understood by substituting the second case (2.23) into (2.4), i.e.,

$$p(\omega) = \rho(\omega) e^{\mp\sqrt{\bar{k}_y^2 + \bar{k}_z^2 - \bar{k}^2}} e^{\mathrm{i}(\bar{k}_y y + \bar{k}_z z)}. \tag{2.24}$$

The left plot of Fig. 2.1 shows the relationship between propagating waves and evanescent waves when $\bar{k}_y = 0$ and $\bar{k}_z = 0$. The bold line $\bar{k} = \bar{k}_x$ splits the region into two groups, i.e., propagating waves and evanescent waves. Evanescent waves, highlighted in Fig. 2.1 as the region below the border line, decay exponentially and, hence, are often prone to measurement noise of microphones [*2].

[*2] An evanescent wave is often introduced in the context of loudspeakers and is referring there to the part of the emitted sound from the loudspeakers that does not reach far, i.e., the evanescent component of sound. However, the colored region in Fig. 2.1 is also referred to as evanescent waves in the literature of near-field acoustic holography with microphone arrays, e.g., [74].

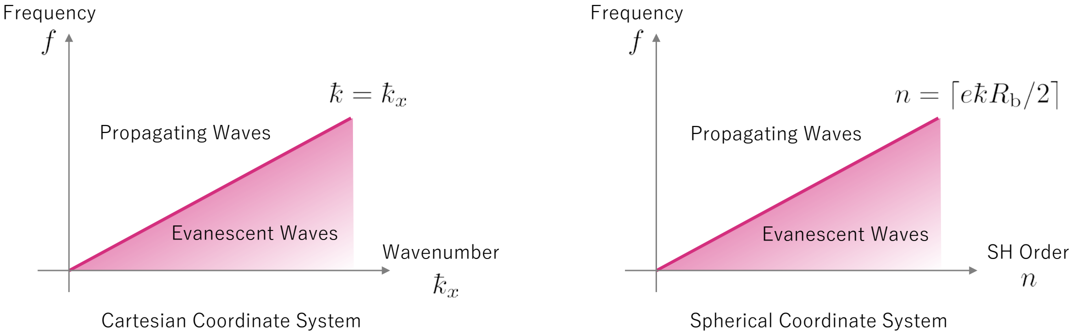


Fig. 2.1. Relationship between propagating and evanescent waves in the Cartesian (left) and spherical coordinate systems (right). The wavenumbers of y and z axes are set to 0 for the left plot.

In the spherical coordinate system, it is known that $n = \lceil e\hbar R_\mathrm{b}/2 \rceil$ can split the region into two groups owing to the characteristics of the spherical Hankel function [75] as shown in the right plot of Fig. 2.1. We investigate in Chapter 6 whether the use of the evanescent region is beneficial for source separation.

Chapter 3

Multichannel Blind Source Separation

3.1 Local Gaussian Model

This section provides an introduction to the model underlying MNMF and its variant, DOA-based MNMF [46]. The model assumes that an A-channel vector of a short-time Fourier transform (STFT) bin for ith source can be modeled as a multivariate complex Gaussian, i.e.,

$$\mathbf{s}_{ft}^{i} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}_{ft}^{i}\right), \tag{3.1}$$

where $\mathbf{s}_{ft}^{i} \in \mathbb{C}^{A}$ denotes the spatial image of the ith source in the STFT domain, $\mathbf{R}_{ft}^{i} = \mathbb{E}\left[\mathbf{s}_{ft}^{i}\mathbf{s}_{ft}^{i}{}^{H}\right] \in \mathbb{C}^{A\times A}$ denotes the covariance matrix of the complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}$, f is the frequency bin index that corresponds to ꝁ, and t is the time frame index.

The spatial image of a mixture of multiple sources $\mathbf{x}_{ft} \in \mathbb{C}^{A}$ is represented as a sum of complex Gaussians, i.e.,

$$\mathbf{x}_{ft} = \sum_{i} \mathbf{s}_{ft}^{i} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}_{ft}\right), \tag{3.2}$$

where $\mathbf{R}_{ft} \in \mathbb{C}^{A\times A}$ denotes the SCM. Assuming that the sources are mutually independent, the SCM of the mixture $\mathbf{R}_{ft}$ is given by the sum of the SCMs of all sources, i.e.,

$$\mathbf{R}_{ft} = \mathbb{E}\left[\mathbf{x}_{ft}\mathbf{x}_{ft}^{H}\right] = \sum_{i} \mathbf{R}_{ft}^{i}. \tag{3.3}$$

The log-likelihood of the spatial image $\mathbf{x}_{ft}$ for the model parameters $\boldsymbol{\varphi}$ under the assumption of the local Gaussian model (3.2) is given by

$$\log \mathbb{P}(\mathbf{x}|\boldsymbol{\varphi}) = \sum_{ft} \log \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{ft}|\mathbf{0}, \hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi})\right), \tag{3.4}$$

where the SCM of the mixture is modeled by $\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi})$ and the parameter $\boldsymbol{\varphi}$ will be defined later in Secs. 3.2 and 3.3. The maximization of this likelihood can be interpreted as the minimization of the log-determinant divergence [76] between the empirical SCM, $\tilde{\mathbf{R}}_{ft} = \mathbf{x}_{ft}\mathbf{x}_{ft}^{H}$, and the estimated SCM, $\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi}) \in \mathbb{C}^{A\times A}$:

$$\begin{aligned} C(\boldsymbol{\varphi}) &= \sum_{ft} D_{\mathrm{LD}}\left(\tilde{\mathbf{R}}_{ft}|\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi})\right) \\ &\equiv \sum_{ft} \operatorname{tr}\left(\tilde{\mathbf{R}}_{ft}\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi})^{-1}\right) + \log\det\left(\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi})\right), \end{aligned} \tag{3.5}$$

where $C(\boldsymbol{\varphi})$ can be seen as a cost function that we want to minimize with respect to the model parameters $\boldsymbol{\varphi}$. The log-determinant divergence, D_{LD}, is often preferred for minimizing SCMs owing to its scale-invariant nature [37, 44].

3.2 Multichannel Non-negative Matrix Factorization

In the framework of MNMF proposed in [44] where $\boldsymbol{\varphi} = \{\mathbf{Q}_{fk}, w_{fk}, h_{kt}\}$, the SCM $\hat{\mathbf{R}}_{ft}$ is assumed to be a superposition of time-invariant normalized SCMs $\mathbf{Q}_{fk} \in \mathbb{C}^{A \times A}$ coupled with a scale value that represents the power spectral density. The scale value is decomposed into a non-negative frequency weight w_{fk} and a non-negative activation h_{kt},

$$\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi}) = \sum_{k} \mathbf{Q}_{fk} w_{fk} h_{kt}, \tag{3.6}$$

where the NMF component index is denoted by k. Even though the performance of MNMF exceeds that of other multichannel based approaches, such as independent vector analysis (IVA) [77], it is known to be sensitive to the initialization of the parameters. The drawback can be mitigated by employing other multichannel methods as an initializer of MNMF [78]. Another major problem of MNMF is its computational cost because the complexity increases with the number of microphones, A. In particular, in the case of MNMF based on the log-determinant divergence (3.5), numerous matrix inversions in the update rules of w_{fk} and h_{kt} and eigendecompositions in the update of $\mathbf{Q}_{fk}$ result in a very high computational cost of order $O(A^3)$ [41].

3.3 MNMF with Fixed DOA Kernels

In the extended approach of MNMF proposed in [46] where $\boldsymbol{\varphi} = \{z_{ko}, w_{fk}, h_{kt}\}$, the time-invariant normalized SCM is further decomposed into a set of fixed DOA kernels $\mathbf{J}_{fo} \in \mathbb{C}^{A \times A}$ and the corresponding directional weights $z_{ko} \in \mathbb{R}_{+}^{K \times O}$,

$$\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi}) = \sum_{k} \sum_{o} \mathbf{J}_{fo} z_{ko} w_{fk} h_{kt}, \tag{3.7}$$

where $\mathbf{Q}_{fk} = \sum_{o} \mathbf{J}_{fo} z_{ko}$ results in the same equation as (3.6) and o denotes the index of steering directions for the DOA kernels. By using a fixed basis of DOA kernels throughout the separation process, the stability of the performance is improved even when the parameters are initialized with random values. Furthermore, there is also a major advantage in terms of computational cost owing to the elimination of the update of the normalized SCM $\mathbf{Q}_{fk}$.

Opposed to MNMF, which blindly estimates the normalized SCM, the DOA-based approach requires that the geometry of the array is known in order to compute the DOA kernels. In the case of a uniform linear array, the DOA kernel is composed of an outer product of steering vectors $\mathbf{q}_{fo} \in \mathbb{C}^{A}$, where each steering vector is represented as a function of the time delay determined by the steering direction and the microphone distance, i.e.,

$$\mathbf{J}_{fo} = \mathbf{q}_{fo} \mathbf{q}_{fo}^{H}, \tag{3.8}$$

with

$$\mathbf{q}_{fo}^{T} = \begin{bmatrix} 1 & e^{\mathrm{i}\omega_f \tau_o} & \cdots & e^{\mathrm{i}\omega_f (A-1)\tau_o} \end{bmatrix}, \tag{3.9}$$

where ω_f denotes the frequency and τ_o denotes the time delay between each microphone and the array center.

Although the overall computational cost is greatly reduced by fixing the DOA kernels, the presence of the matrix inversions in the updates of z_{ko}, w_{fk}, and h_{kt} prevents us from applying the algorithm to the scenario where a large number of microphones are required, i.e., where A is large.

3.4 Derivation of Update Rules

To minimize the cost function in (3.5) with (3.7), we employ the majorization–minimization (MM) algorithm [45] to reduce the cost monotonically. The upper bound of the cost function (3.5) together with a model (3.7) can be constructed by applying two inequalities to the convex part and the concave part, respectively [79]. This yields

$$C^{+}(\boldsymbol{\varphi}, \mathbf{T}_{ftko}, \mathbf{U}_{ft}) = \sum_{ft} \Bigg(\sum_{ko} \frac{\operatorname{tr}\left(\tilde{\mathbf{R}}_{ft} \mathbf{T}_{ftko}^{H} \left(\mathbf{J}_{fo}\right)^{-1} \mathbf{T}_{ftko}\right)}{z_{ko} w_{fk} h_{kt}} + \log\det \mathbf{U}_{ft} + \operatorname{tr}\left(\mathbf{U}_{ft}^{-1} \hat{\mathbf{R}}_{ft}\right) - A \Bigg), \quad (3.10)$$

where $\mathbf{T}_{ftko}$ and $\mathbf{U}_{ft}$ are auxiliary variable matrices that satisfy $\sum_{ko} \mathbf{T}_{ftko} = \mathbf{I}, \mathbf{T}_{ftko} \succeq 0$, and $\mathbf{U}_{ft} \succeq 0$. The equality of the auxiliary function and the cost function holds only when the auxiliary variables satisfy

$$\mathbf{T}_{ftko} = \left(\mathbf{J}_{fo} z_{ko} w_{fk} h_{kt}\right) \left(\hat{\mathbf{R}}_{ft}\right)^{-1}, \quad (3.11a)$$

$$\mathbf{U}_{ft} = \hat{\mathbf{R}}_{ft}. \quad (3.11b)$$

The partial derivatives with respect to z_{ko}, w_{fk}, and h_{kt} are derived by minimizing the upper bound function $C^{+}(\boldsymbol{\varphi}, \mathbf{T}_{ftko}, \mathbf{U}_{ft})$. The derivatives of $C^{+}(\boldsymbol{\varphi}, \mathbf{T}_{ftko}, \mathbf{U}_{ft})$ with respect to the model parameters are

$$\frac{\partial C^{+}}{\partial z_{ko}} = \sum_{ft} \Bigg(- \frac{\operatorname{tr}\left(\tilde{\mathbf{R}}_{ft} \mathbf{T}_{ftko}^{H} \left(\mathbf{J}_{fo}\right)^{-1} \mathbf{T}_{ftko}\right)}{z_{ko}^{2} w_{fk} h_{kt}} + \operatorname{tr}\left(\mathbf{U}_{ft}^{-1} \mathbf{J}_{fo}\right) w_{fk} h_{kt} \Bigg), \quad (3.12a)$$

$$\frac{\partial C^{+}}{\partial w_{fk}} = \sum_{to} \Bigg(- \frac{\operatorname{tr}\left(\tilde{\mathbf{R}}_{ft} \mathbf{T}_{ftko}^{H} \left(\mathbf{J}_{fo}\right)^{-1} \mathbf{T}_{ftko}\right)}{z_{ko} w_{fk}^{2} h_{kt}} + \operatorname{tr}\left(\mathbf{U}_{ft}^{-1} \mathbf{J}_{fo}\right) z_{ko} h_{kt} \Bigg), \quad (3.12b)$$

$$\frac{\partial C^{+}}{\partial h_{kt}} = \sum_{fo} \Bigg(- \frac{\operatorname{tr}\left(\tilde{\mathbf{R}}_{ft} \mathbf{T}_{ftko}^{H} \left(\mathbf{J}_{fo}\right)^{-1} \mathbf{T}_{ftko}\right)}{z_{ko} w_{fk} h_{kt}^{2}} + \operatorname{tr}\left(\mathbf{U}_{ft}^{-1} \mathbf{J}_{fo}\right) z_{ko} w_{fk} \Bigg). \quad (3.12c)$$

The multiplicative update rules for the model parameters z_{ko}, w_{fk}, and h_{kt} are obtained by equating the partial derivatives to zero and can be simplified by substituting (3.11a) and (3.11b):

$$z_{ko} \leftarrow z_{ko} \sqrt{\frac{\sum_{ft} \operatorname{tr}\left(\hat{\mathbf{R}}_{ft}^{-1} \tilde{\mathbf{R}}_{ft} \hat{\mathbf{R}}_{ft}^{-1} \mathbf{J}_{fo}\right) w_{fk} h_{kt}}{\sum_{ft} \operatorname{tr}\left(\hat{\mathbf{R}}_{ft}^{-1} \mathbf{J}_{fo}\right) w_{fk} h_{kt}}}, \tag{3.13}$$

$$w_{fk} \leftarrow w_{fk} \sqrt{\frac{\sum_{on} \operatorname{tr}\left(\hat{\mathbf{R}}_{ft}^{-1} \tilde{\mathbf{R}}_{ft} \hat{\mathbf{R}}_{ft}^{-1} \mathbf{J}_{fo}\right) z_{ko} h_{kt}}{\sum_{on} \operatorname{tr}\left(\hat{\mathbf{R}}_{ft}^{-1} \mathbf{J}_{fo}\right) z_{ko} h_{kt}}}, \tag{3.14}$$

$$h_{kt} \leftarrow h_{kt} \sqrt{\frac{\sum_{fo} \operatorname{tr}\left(\hat{\mathbf{R}}_{ft}^{-1} \tilde{\mathbf{R}}_{ft} \hat{\mathbf{R}}_{ft}^{-1} \mathbf{J}_{fo}\right) z_{ko} w_{fk}}{\sum_{fo} \operatorname{tr}\left(\hat{\mathbf{R}}_{ft}^{-1} \mathbf{J}_{fo}\right) z_{ko} w_{fk}}}. \tag{3.15}$$

These formulations are simpler variants of an algorithm proposed by Higuchi and Kameoka [80]. As can be seen from the update rules, the matrix inversions of the estimated SCMs, which are of order $O(A^3)$, make the algorithm intractable for a large number of microphones (e.g., 32).

Chapter 4

Diagonal Spatial Covariance Matrix in Wavenumber Domain

4.1 Motivation

To overcome the weakness of the high computational cost in MNMF explained in Chapter 3, several authors have attempted to apply different types of orthogonal transforms to an SCM, enabling the energy of the SCM to be concentrated in the diagonal part. As a result, a matrix inversion in the update can be replaced with element-wise diagonal divisions, thus reducing the high computational cost. The authors in [81] leveraged a steering matrix to convert the SCM into the so-called *beamspace* domain. The method was further generalized in the framework called *PROJET* [82]. There have been several works on the iterative estimation of the projection matrix by IVA [77], either jointly with NMF updates [78, 83] or independently followed by the NMF approach [84]. In this chapter, we first propose the use of a fast Fourier transform (FFT) to project signals into the wavenumber domain and to model only the band elements of an SCM, condition that a uniform linear array is used as a recording device. The conversion can be achieved more efficiently than by other projection-based methods while making use of the property that plane waves can be sparsely represented in the wavenumber domain.

4.2 Model and Derivation of Update Rules

4.2.1 Wavenumber Transform of SCMs

The STFT multichannel signals are converted into the wavenumber domain, where the underlying probability model is based on the zero-mean complex Gaussian distribution,

$$\mathbf{F}^H \mathbf{x}_{ft} = \mathbf{F}^H \sum_i \mathbf{s}_{ft}^i \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \mathbf{R}_{ft}^{\mathrm{SP}} \right), \tag{4.1}$$

with

$$\mathbf{R}_{ft}^{\mathrm{SP}} = \mathbf{F}^H \tilde{\mathbf{R}}_{ft} \mathbf{F}, \tag{4.2}$$

where $\mathbf{R}_{ft}^{\mathrm{SP}} \in \mathbb{C}^{A \times A}$ denotes the SCM in the wavenumber domain and $\mathbf{F} \in \mathbb{C}^{A \times A}$ denotes the discrete Fourier transform (DFT) matrix.

Given $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}} = \mathbf{F}^H \hat{\mathbf{R}}_{ft} \mathbf{F}$, the cost function for the local Gaussian model in (3.5) can then be modified by replacing the SCM in the TF domain with the matrix in the wavenumber domain,

$$C_{\mathrm{sp}}(\boldsymbol{\varphi}) = \sum_{ft} D_{\mathrm{LD}}\left(\mathbf{R}^{\mathrm{SP}}_{ft} | \hat{\mathbf{R}}^{\mathrm{SP}}_{ft}\right). \tag{4.3}$$

Note that $C_{\mathrm{sp}}(\boldsymbol{\varphi})$ is equivalent to $C(\boldsymbol{\varphi})$ in (3.5) as can be seen from

$$\begin{aligned}
C_{\mathrm{sp}}(\boldsymbol{\varphi}) &= \sum_{ft} D_{\mathrm{LD}}\left(\mathbf{F}^H \tilde{\mathbf{R}}_{ft}\mathbf{F} | \mathbf{F}^H \hat{\mathbf{R}}_{ft}\mathbf{F}\right) \\
&= \sum_{ft} \mathrm{tr}\left(\mathbf{F}^H \tilde{\mathbf{R}}_{ft}\mathbf{F}\mathbf{F}^{-1}\hat{\mathbf{R}}_{ft}^{-1}\mathbf{F}^{-H}\right) \\
&\quad + \log\det\left(\mathbf{F}^H \hat{\mathbf{R}}_{ft}\mathbf{F}\right) \\
&= \sum_{ft} \mathrm{tr}\left(\mathbf{F}^H \mathbf{F}\mathbf{F}^{-1}\mathbf{F}^{-H}\tilde{\mathbf{R}}_{ft}\hat{\mathbf{R}}_{ft}^{-1}\right) \\
&\quad + \log\det\left(\hat{\mathbf{R}}_{ft}\right) - \log\det\left(\mathbf{F}\right) + \log\det\left(\mathbf{F}\right) \\
&= C(\boldsymbol{\varphi}).
\end{aligned} \tag{4.4}$$

Since w_{fk} and h_{kt} are not dependent on the channel dimension, the spatial transform of the estimated SCMs, supposed to be performed in every iteration, can be replaced with a single spatial transform of the fixed DOA kernels $\mathbf{J}_{fo}$ at the initialization stage,

$$\begin{aligned}
\hat{\mathbf{R}}^{\mathrm{SP}}_{ft} &= \mathbf{F}^H \hat{\mathbf{R}}_{ft}\mathbf{F} \\
&= \sum_k \sum_o \mathbf{J}^{\mathrm{SP}}_{fo} z_{ko} w_{fk} h_{kt},
\end{aligned} \tag{4.5}$$

with

$$\mathbf{J}^{\mathrm{SP}}_{fo} = \mathbf{F}^H \mathbf{J}_{fo}\mathbf{F}. \tag{4.6}$$

4.2.2 Sparseness of SCMs

Fig. 4.1 shows an example of a comparison between an SCM for $A = 32$ in the TF domain and the converted matrix in the wavenumber domain. It is clear from the figure that the SCM in the wavenumber domain has strong peaks in the band elements, whereas the one in the TF domain has quasi-uniformly distributed values. A comparison for $A = 4$ is shown in Fig. 4.2. Again, the SCM in the wavenumber domain exhibits a small number of strong peaks. To numerically assess the sparseness of the SCMs, we computed the percentage of elements that are smaller in magnitude than 20% of the maximum element. For the case of $A = 32$ (Fig. 4.1), we obtain 6.1% in (a) and 99.5% in (b). Also, for the case of $A = 4$ (Fig. 4.2), we obtain 0% in (a) and 81.3% in (b). Thus, even for $A = 4$, more than 80% of all SCM elements in the wavenumber domain are smaller than the threshold whereas it was none in the TF domain. Finally, to observe the effect of room reverberation on the sparseness of the SCM, we conducted a simulation based on the image method [85]. The simulated room size is 7.0 m × 12.0 m × 3.0 m and the reverberation time is 400 ms. The setup is the same as for Figs. 4.1 and 4.2, i.e., the three sources are placed at the corresponding directions at the radius of 3 m. Fig. 4.3 shows the resultant wavenumber-domain SCMs obtained in this simulation. In the reverberant scenario, the sparseness of the SCMs are 91.8% for $A = 32$ and 68.8% for $A = 4$. Although the values decrease gradually compared with the anechoic case, the sparse nature in the wavenumber domain still remains. Furthermore, the concentration on the diagonal elements can still be seen in Fig. 4.3, as in Figs. 4.1 and 4.2. The high power regions in the off-diagonal parts correspond to the reflections from the walls, and omitting those regions can be considered as a loss of information in general. However, the evaluations in this chapter proves that our proposed method is still effective in improving performance in the case of near-field sources under reverberant conditions.

4.2.3 Wiener Filtering in Wavenumber Domain

Given the estimated model parameters, the STFT coefficients of each source can be recovered by a multichannel Wiener filter, i.e., a minimum mean squared error (MMSE) estimator [86]. Since $\hat{\mathbf{R}}_{ft} = \mathbf{F}\hat{\mathbf{R}}^{\mathrm{SP}}_{ft}\mathbf{F}^H$ and $\mathbf{J}_{fo} = \mathbf{F}\mathbf{J}^{\mathrm{SP}}_{fo}\mathbf{F}^H$ also hold owing to $\mathbf{F}\mathbf{F}^H = \mathbf{I}$ where $\mathbf{I}$ denotes an identity matrix, the MMSE estimator can be given in the wavenumber domain by

$$\begin{aligned}\hat{\mathbf{s}}^i_{ft} &= \left(\sum_{k\in K_i}\sum_{o}\mathbf{J}_{fo}z_{ko}w_{fk}h_{kt}\right)\left(\hat{\mathbf{R}}_{ft}\right)^{-1}\mathbf{x}_{ft}\\ &= \left(\sum_{k\in K_i}\sum_{o}\mathbf{F}\mathbf{J}^{\mathrm{SP}}_{fo}\mathbf{F}^H z_{ko}w_{fk}h_{kt}\right)\left(\mathbf{F}\hat{\mathbf{R}}^{\mathrm{SP}}_{ft}\mathbf{F}^H\right)^{-1}\mathbf{x}_{ft}\\ &= \mathbf{F}\left(\sum_{k\in K_i}\sum_{o}\mathbf{J}^{\mathrm{SP}}_{fo}z_{ko}w_{fk}h_{kt}\right)\left(\hat{\mathbf{R}}^{\mathrm{SP}}_{ft}\right)^{-1}\mathbf{F}^H\mathbf{x}_{ft}, \qquad (4.7)\end{aligned}$$

where K_i denotes the set of components that belong to the ith source. The set of NMF components K_i can be determined by a clustering method, such as LPC-based [87] or Mel-spectrum-based clustering [88].

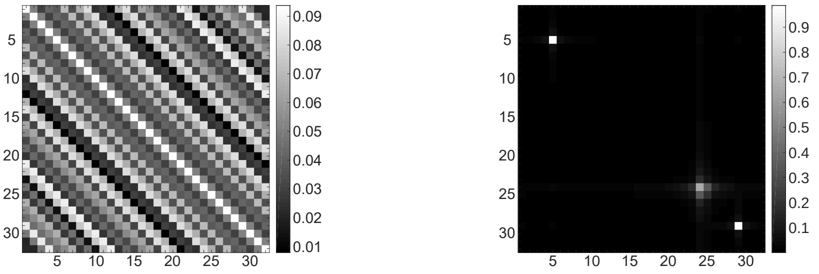
4.2.4 Diagonal Approximation

If we assume that $\mathbf{F}^H\hat{\mathbf{R}}_{ft}\mathbf{F}$ and $\mathbf{F}^H\mathbf{J}_{fo}\mathbf{F}$ are diagonal matrices (see Fig. 4.1), then $\mathbf{F}^H\hat{\mathbf{R}}_{ft}\mathbf{F}$ and $\mathbf{F}^H\mathbf{J}_{ft}\mathbf{F}$ can be well approximated by considering only their diagonal elements,

$$\mathbf{F}^H\hat{\mathbf{R}}_{ft}\mathbf{F} \approx \hat{\mathbf{R}}^{\mathrm{Diag}}_{ft} = \begin{pmatrix} \hat{\gamma}_{1ft} & 0 & \dots & 0\\ 0 & \hat{\gamma}_{2ft} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \hat{\gamma}_{Aft}\end{pmatrix}, \qquad (4.8)$$

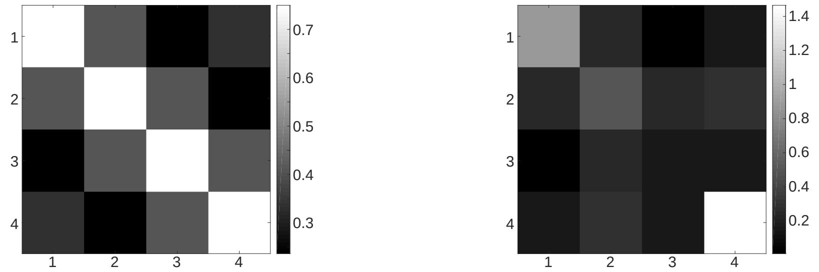
$$\mathbf{F}^H\mathbf{J}_{fo}\mathbf{F} \approx \mathbf{J}^{\mathrm{Diag}}_{fo} = \begin{pmatrix} \beta_{1fo} & 0 & \dots & 0\\ 0 & \beta_{2fo} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \beta_{Afo}\end{pmatrix}. \qquad (4.9)$$

Thus, the cost function (4.3) can also be approximated by focusing on the diagonal



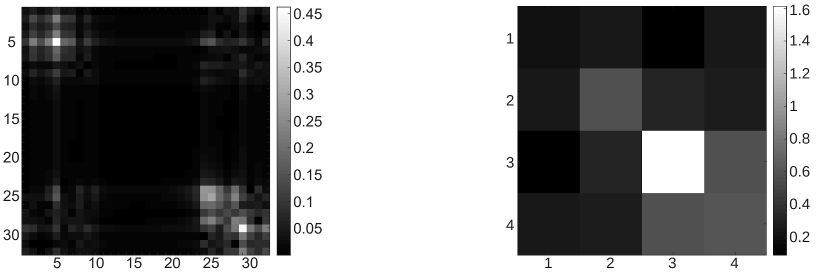
(a) $|\mathbf{R}_{ft}|$ (b) $|\mathbf{F}^H\mathbf{R}_{ft}\mathbf{F}|$

Fig. 4.1. SCM of mixed plane waves (2 kHz) originating from $\pi/20$, $8\pi/20$, and $14\pi/20$. The number of microphones, A, is 32. $|.|$ denotes the element-wise absolute value of the matrix.



(a) $|\mathbf{R}_{ft}|$ (b) $|\mathbf{F}^H\mathbf{R}_{ft}\mathbf{F}|$

Fig. 4.2. SCM of mixed plane waves (2 kHz) originating from $\pi/20$, $8\pi/20$, and $14\pi/20$. The number of microphones, A, is 4. $|.|$ denotes the element-wise absolute value of the matrix.



(a) $|\mathbf{F}^H\mathbf{R}_{ft}\mathbf{F}|$ (b) $|\mathbf{F}^H\mathbf{R}_{ft}\mathbf{F}|$

Fig. 4.3. SCM of three point sources (2 kHz) originating from $\pi/20$, $8\pi/20$, and $14\pi/20$, simulated in the reverberant scenario with the image method (7.0 m × 12.0 m × 3.0 m, T60 = 400 ms). The number of microphones, A, is 32 for (a), and 4 for (b). $|.|$ denotes the element-wise absolute value of the matrix.

elements,

$$\begin{aligned} C_{\text{sp}}(\boldsymbol{\varphi}) &\approx \sum_{ft} D_{\text{LD}}\left(\mathbf{R}^{\text{SP}}_{ft}|\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right) \\ &= \sum_{ft} D_{\text{LD}}\left(\mathbf{R}^{\text{SP}}_{ft}|\sum_{k}\sum_{o}\mathbf{J}^{\text{Diag}}_{fo} z_{ko} w_{fk} h_{kt}\right). \end{aligned} \tag{4.10}$$

4.2.5 Derivation of Update Rules

The update rules reflecting the approximation can be derived as in 3.4 such that they only contain the inversions of diagonal matrices, allowing the algorithm to run at a computational cost of order $O(A)$ in each iteration,

$$z_{ko} \leftarrow z_{ko}\sqrt{\frac{\sum_{ft}\operatorname{tr}\left(\left(\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right)^{-1}\mathbf{R}^{\text{SP}}_{ft}\left(\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right)^{-1}\mathbf{J}^{\text{Diag}}_{fo}\right) w_{fk}h_{kt}}{\sum_{ft}\operatorname{tr}\left(\left(\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right)^{-1}\mathbf{J}^{\text{Diag}}_{fo}\right) w_{fk}h_{kt}}}, \tag{4.11a}$$

$$w_{fk} \leftarrow w_{fk}\sqrt{\frac{\sum_{on}\operatorname{tr}\left(\left(\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right)^{-1}\mathbf{R}^{\text{SP}}_{ft}\left(\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right)^{-1}\mathbf{J}^{\text{Diag}}_{fo}\right) z_{ko}h_{kt}}{\sum_{on}\operatorname{tr}\left(\left(\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right)^{-1}\mathbf{J}^{\text{Diag}}_{fo}\right) z_{ko}h_{kt}}}, \tag{4.11b}$$

$$h_{kt} \leftarrow h_{kt}\sqrt{\frac{\sum_{fo}\operatorname{tr}\left(\left(\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right)^{-1}\mathbf{R}^{\text{SP}}_{ft}\left(\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right)^{-1}\mathbf{J}^{\text{Diag}}_{fo}\right) z_{ko}w_{fk}}{\sum_{fo}\operatorname{tr}\left(\left(\hat{\mathbf{R}}^{\text{Diag}}_{ft}\right)^{-1}\mathbf{J}^{\text{Diag}}_{fo}\right) z_{ko}w_{fk}}}. \tag{4.11c}$$

The cost function can also be written in a similar manner to the cost function for NTF with the IS divergence denoted by D_{IS}, i.e.,

$$\begin{aligned} C_{\text{sp}}(\boldsymbol{\varphi}) &\approx \sum_{aft} D_{\text{IS}}\left(\gamma_{aft}|\hat{\gamma}_{aft}\right) \\ &= \sum_{aft} D_{\text{IS}}\left(\gamma_{aft}|\sum_{k}\sum_{o}\beta_{afo} z_{ko} w_{fk} h_{kt}\right), \end{aligned} \tag{4.12}$$

under the assumption that $\mathbf{F}^{H}\hat{\mathbf{R}}_{ft}\mathbf{F} = \operatorname{diag}(\hat{\gamma}_{1ft}, ..., \hat{\gamma}_{aft})$.

To minimize the cost function in (4.12), we employ the MM algorithm [45] to reduce the cost monotonically. The auxiliary function can be constructed by applying Jensen's inequality to the convex part of the cost function and the tangent line to the concave part:

$$C^{+}_{\text{sp}}(\boldsymbol{\varphi}, \eta_{aftko}, \psi_{aft}) = \sum_{aft}\left(\sum_{ko}\eta^{2}_{aftko}\frac{\gamma_{aft}}{\beta_{afo} z_{ko} w_{fk} h_{kt}} + \log\psi_{aft} + \frac{\hat{\gamma}_{aft} - \psi_{aft}}{\psi_{aft}}\right), \tag{4.13}$$

where η_{aftko} and ψ_{aft} are auxiliary variables that satisfy $\sum_{ko}\eta_{aftko} = 1, \eta_{aftko} \geq 0$ and $\psi_{aft} \geq 0$. The equality of the auxiliary function and the cost function holds only when the auxiliary variables satisfy

$$\eta_{aftko} = \frac{\beta_{afo} z_{ko} w_{fk} h_{kt}}{\hat{\gamma}_{aft}}, \tag{4.14}$$

$$\psi_{aft} = \hat{\gamma}_{aft}. \tag{4.15}$$

The partial derivatives with respect to z_{ko}, w_{fk}, and h_{kt} are derived by minimizing the auxiliary function:

$$\frac{\partial C_{\mathrm{sp}}^{+}}{\partial z_{ko}} = \sum_{aft} \left(-\eta_{aftko}^{2} \frac{\gamma_{aft}}{\beta_{afo} z_{ko}^{2} w_{fk} h_{kt}} + \frac{\beta_{afo} w_{fk} h_{kt}}{\psi_{aft}} \right), \tag{4.16}$$

$$\frac{\partial C_{\mathrm{sp}}^{+}}{\partial w_{fk}} = \sum_{ato} \left(-\eta_{aftko}^{2} \frac{\gamma_{aft}}{\beta_{afo} z_{ko} w_{fk}^{2} h_{kt}} + \frac{\beta_{afo} z_{ko} h_{kt}}{\psi_{aft}} \right), \tag{4.17}$$

$$\frac{\partial C_{\mathrm{sp}}^{+}}{\partial h_{kt}} = \sum_{afo} \left(-\eta_{aftko}^{2} \frac{\gamma_{aft}}{\beta_{afo} z_{ko} w_{fk} h_{kt}^{2}} + \frac{\beta_{afo} z_{ko} w_{fk}}{\psi_{aft}} \right). \tag{4.18}$$

The update rules can be derived by equating the derivative to zero and can be simplified by substituting (4.14) and (4.15):

$$z_{ko} \leftarrow z_{ko} \sqrt{\frac{\sum_{ft} \sum_{a} \frac{\gamma_{aft}}{\hat{\gamma}_{aft}^{2}} \beta_{afo} w_{fk} h_{kt}}{\sum_{ft} \sum_{a} \frac{1}{\hat{\gamma}_{aft}} \beta_{afo} w_{fk} h_{kt}}}, \tag{4.19a}$$

$$w_{fk} \leftarrow w_{fk} \sqrt{\frac{\sum_{on} \sum_{a} \frac{\gamma_{aft}}{\hat{\gamma}_{aft}^{2}} \beta_{afo} z_{ko} h_{kt}}{\sum_{on} \sum_{a} \frac{1}{\hat{\gamma}_{aft}} \beta_{afo} z_{ko} h_{kt}}}, \tag{4.19b}$$

$$h_{kt} \leftarrow h_{kt} \sqrt{\frac{\sum_{fo} \sum_{a} \frac{\gamma_{aft}}{\hat{\gamma}_{aft}^{2}} \beta_{afo} z_{ko} w_{fk}}{\sum_{fo} \sum_{a} \frac{1}{\hat{\gamma}_{aft}} \beta_{afo} z_{ko} w_{fk}}}. \tag{4.19c}$$

It is clear from the equations that they no longer contain matrix inversions. The proposed diagonal algorithm is given in Algorithm 1.

4.3 Evaluation

An evaluation was conducted by using the BSS Eval Toolbox, which calculates the source-to-distortion ratio (SDR), the source-to-interferences ratio (SIR), and the sources-to-artifacts ratio (SAR) [89]. Algorithm 1 (Diag) was compared with other BSS methods in two scenarios: anechoic and reverberant. The sound samples from SiSEC (Signal Separation Evaluation Campaign) database 2008[*1] were used in combination with room impulse responses (RIRs) associated with source positions [90]. The test conditions (Table 4.1) were the same for both experiments.

[*1] http://sisec2008.wiki.irisa.fr/tiki-index.html

Algorithm 1 Diagonal approximation approach

```
Require: Mixture $\mathbf{x}_{ft}$

  Compute $\tilde{\mathbf{R}}_{ft} = \mathbf{x}_{ft}\mathbf{x}_{ft}^H$
  Apply spatial transform to $\tilde{\mathbf{R}}_{ft}$ with (4.2)
  Apply spatial transform to $\mathbf{J}_{fo}$ with (4.6)
  Extract diagonal part of $\mathbf{F}^H\mathbf{J}_{fo}\mathbf{F}$ with (4.9)
  Initialize $z_{ko}, w_{fk}, h_{kt}$ with randomized values
  Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}}$ with (4.5)

  for $\jmath = 0$ to MM iteration do
    $z_{ko} \leftarrow$ (4.19a)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}}$ with (4.5)
    $w_{fk} \leftarrow$ (4.19b)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}}$ with (4.5)
    $h_{kt} \leftarrow$ (4.19c)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}}$ with (4.5)
  end for

  Cluster NMF components based on [87, 88]
  Apply Wiener filtering with (4.7)

Ensure: Estimates $\hat{\mathbf{s}}_{ft}^i$
```

Table 4.1. Experimental setup

Number of sources	3
Number of channels	$A = 32$
Sampling rate	16 kHz
STFT frame size	1024
STFT frame shift	512
Number of iterations	100

4.3.1 Anechoic Scenario

For the anechoic scenario, Algorithm 1 (Diag) was compared to IVA [77] and NTF [53]. Multichannel observations for a uniform linear array were created simply by summing all the sources together with the addition of proper delays in the frequency domain. Separated signals for IVA were reconstructed by applying the projection back [91]. The distance between microphones was set to 0.384 m. The number of components for NTF and Algorithm 1 (Diag) was 18. The angle resolution for DOA kernels was limited to 10° in the range 0–180° due to computational cost. It should be noted that the approximation resulting from the extraction of the diagonal elements of SCMs is not correct for a uniform linear array because an SCM cannot be a circulant matrix. However, the approximation error can be mitigated by increasing the length of the array [74, 92].

The average improvement in SDR per file (Fig. 4.4) and the average improvement in SDR per angle between the two nearest sources (Fig. 4.5) show that Algorithm 1 (Diag) outperformed the other two methods for all three files at all source distances. The results

Table 4.2. SDR, SIR, and SAR results

	DS with oracle DOAs			Proposed Method (Diag)		
	SDR	SIR	SAR	SDR	SIR	SAR
Hi-hat	-12.77	-9.88	24.30	7.43	-0.49	2.72
Snare	-3.36	-9.01	10.54	0.68	-7.04	0.70
Bass	3.28	15.93	19.38	16.56	15.13	19.51

Table 4.3. Computation time for 1 iteration [s]

IVA	NTF	Proposed Method (Diag)	DOANMF
0.3	1.2	16.1	1093.4

for different distances show the same tendency as that in [46], namely, that the larger the source distance is, the better the DOA-based method performs with respect to non DOA-based methods. In addition, we assume that Algorithm 1 (Diag) has an advantage over NTF due to the sparse representation of propagated waves in the wavenumber domain, where the sources are less superposed. Unfortunately, due to the extreme computational complexity of MNMF, the comparison for 32 channels could not be completed in a reasonable amount of time. It took 1024 times longer than it took for Algorithm 1 (Diag) ($32^3/32 = 1024$).

4.3.2 Reverberant Scenario

Evaluations for a reverberant scenario were conducted on Algorithm 1 (Diag) and on a delay-and-sum (DS) beamformer with oracle DOAs. The azimuths of the three sources were $\pi/20$, $8\pi/20$, and $14\pi/20$ radians. An image-source method was used to obtain reverberant RIRs [93]. The reverberation time, T_{60}, was 0.5 s for a room 7.68 m $\times$ 14 m $\times$ 6 m in size. A uniform linear array of microphones was assumed to be in the center of the room. The distance between microphones was set to 0.046 m. The other parameters were the same as those in Table 4.1.

The SDR and SIR results (Table 4.2) show that Algorithm 1 (Diag) outperformed the DS beamformer in the reverberant scenario, even with oracle DOAs for the beamformer. This is probably due to the fact that Algorithm 1 (Diag) is capable of modeling full-rank SCMs, even though a circulant matrix is approximated by a large Toeplitz matrix, whereas the DS beamformer can only steer in a single direction.

4.3.3 Computation Time

The computation time required for 1 iteration for each method is listed in Table 4.3. The parameter settings for the experiment are the same as those of Sec. 4.3.1. The computation time was measured by using MATLAB codes. We ran the programs on a Xeon E5-2690 v2 CPU where each core has 3.00 GHz CPU capability. Although there is unexpected overhead in real implementation, Algorithm 1 (Diag) is still greatly faster than MNMF with fixed DOA kernels described in Sec. 3.3 (DOANMF), confirming the advantage of Algorithm 1 (Diag) in computational efficiency.

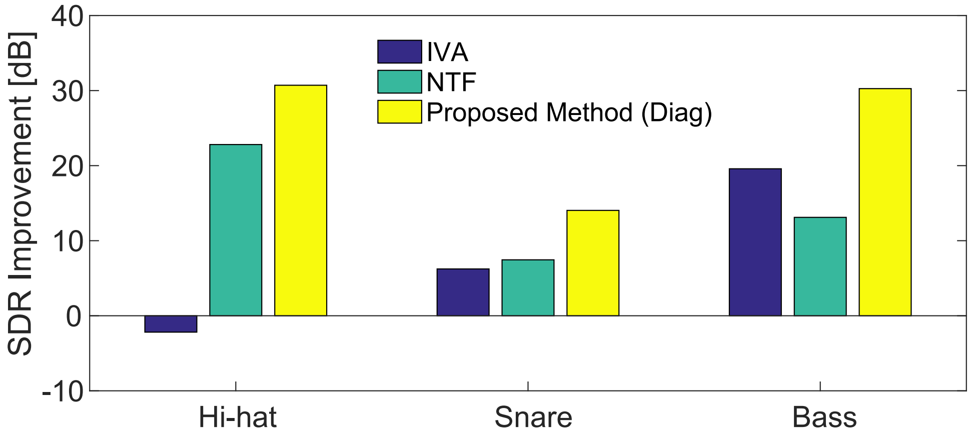


Fig. 4.4. Average improvement in SDR per source for hi-hat, snare drums, and bass guitar.

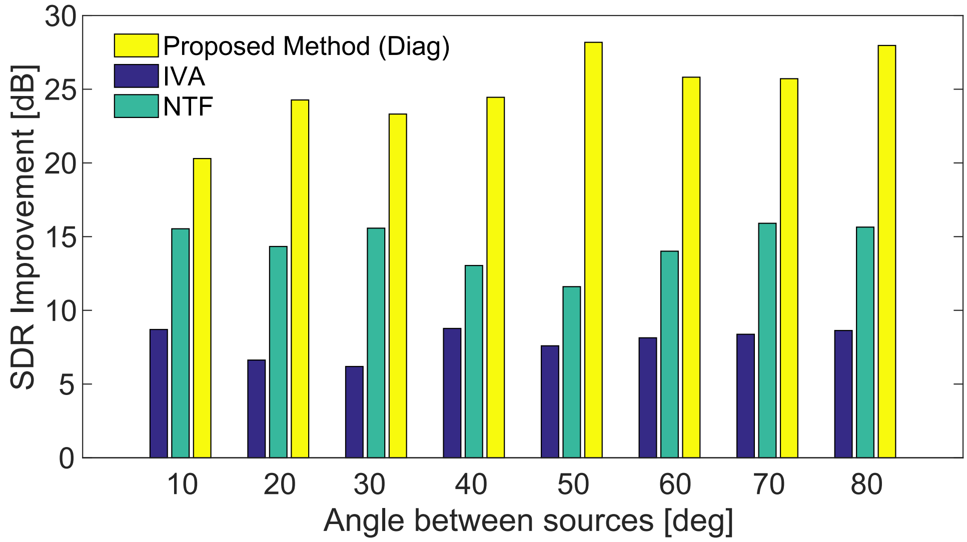


Fig. 4.5. Average improvement in SDR per direction, with the x-axis being the angle between the two nearest sources.

4.4 Conclusion

This chapter describes the use of NTF in the wavenumber domain to reduce the computational cost of BSS for a large number of channels. The technique is based on the approximation of SCMs, which are transformed into the wavenumber domain in advance. With this approximation, the cost of running the algorithm is of order $O(A)$, whereas it is of order $O(A^3)$ for MNMF. An evaluation conducted for anechoic and reverberant scenarios showed that Algorithm 1 (Diag) yielded good separation quality.

Chapter 5

Banded Spatial Covariance Matrix in Wavenumber Domain

5.1 Motivation

The diagonal approximation of an SCM in the wavenumber domain was investigated in Chapter 4, in which only the scenario of using a uniform linear array with a large number of microphones was evaluated. However, it can be assumed that the performance gradually degrades as the number of microphones is reduced because the SCM of projected signals cannot contain sufficient information in the diagonal part when the number of microphones is small. Evaluating the robustness to such a scenario is one of the focuses of this chapter. Secondly, to further increase the robustness, we also devised a tridiagonal approximation approach, where the algorithm not only takes into account the diagonal part but also exploits the adjacent lower and upper bands of the matrix. It relies on the assumption that the tridiagonal part contains more spatial information than the diagonal part. A difficulty exists in the extraction step of the tridiagonal elements because a simple truncation of off-diagonal elements cannot maintain the positive semi-definitiveness of the matrix, which is required to ensure the convergence of the iterative update rules. The inversions of tridiagonal matrices, which appear a few times in the multiplicative updates, can be efficiently computed by means of the Thomas algorithm [94], which can achieve a matrix inversion of order $O(A^2)$. Finally, we examine the incorporation of self-clustering to remove the dependency of clustering methods after the decomposition, which was reported in [95] to yield a quality improvement. The summary of this chapter are as follows:

- an in-depth evaluation of the FFT-based projection method described in Chapter 4 and [96] in the case of a small number of microphones,
- an extension to tridiagonal SCM approximation, where we ensure the positive semi-definitiveness of the matrix, followed by an efficient inverse calculation based on the Thomas algorithm, and,
- the incorporation of self-clustering to iteratively group NMF components into several source types.

5.2 Model and Derivation of Update Rules

5.2.1 Tridiagonal Approximation

Owing to the finite length of the DFT matrix, the SCM cannot be perfectly diagonalized and off-diagonal elements cannot be avoided. This “smear” becomes more dominant in the case of a uniform linear array because the matrix cannot be regarded as a circulant matrix for which diagonalization can be perfectly achieved. Different from the method explained in Chapter 4, where we exploited only the diagonal part, we also make use of the upper and lower adjacent bands, i.e., the tridiagonal part. The matrix entries outside the tridiagonal part are not taken into account in the optimization process.

It should be emphasized that simply discarding the off-tridiagonal part does not ensure the convergence of the multiplicative update rules because setting these entries to zero does not guarantee the positive semi-definitiveness of the resulting tridiagonal SCM, which is an essential assumption throughout the update of model parameters in MNMF. To maintain the positive semi-definitiveness of the matrix during the tridiagonalization process, we must solve the optimization problem

$$\begin{aligned}\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}} &= \arg\min_{\mathbf{V}} \left\|\hat{\mathbf{R}}_{ft}^{\mathrm{SP}} - \mathrm{tridiag}\{\mathbf{V}\}\right\|_F^2 \\ \text{subject to} &\quad \mathrm{tridiag}\{\mathbf{V}\} \succeq \mathbf{0}.\end{aligned} \tag{5.1}$$

The operation tridiag$\{\mathbf{V}\}$ returns a tridiagonal matrix in which elements that are not on the main diagonal and on the diagonal above/below are set to zero. The optimization problem (5.1) is a *nearest-matrix* problem where we solve for the positive semi-definite, tridiagonal matrix $\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}} \in \mathbb{C}^{A\times A}$ that is nearest to $\mathbf{R}_{ft}^{\mathrm{SP}}$ using the Frobenius norm. We solve the semi-definite programming problem (5.1) by using YALMIP [97] with SeDuMi [98].

Although the optimization (5.1) in order to obtain $\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}$ is costly, it can be replaced with solving the same problem for the DOA kernel $\mathbf{J}_{fo}$, which only must be carried out once in the initialization process as the DOA kernels are fixed throughout the update iterations.

$$\begin{aligned}\hat{\mathbf{J}}_{fo}^{\mathrm{Tri}} &= \arg\min_{\mathbf{V}} \left\|\mathbf{J}_{fo}^{\mathrm{SP}} - \mathrm{tridiag}\{\mathbf{V}\}\right\|_F^2 \\ \text{subject to} &\quad \mathrm{tridiag}\{\mathbf{V}\} \succeq \mathbf{0}.\end{aligned} \tag{5.2}$$

$$\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}} = \sum_k \sum_o \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}} z_{ko} w_{fk} h_{kt}. \tag{5.3}$$

Using the approximated tridiagonal SCM, the cost function in (4.3) can be modified to

$$\begin{aligned}C_{\mathrm{sp}}(\boldsymbol{\varphi}) &\approx \sum_{ft} D_{\mathrm{LD}}\left(\mathbf{R}_{ft}^{\mathrm{SP}} | \hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right) \\ &= \sum_{ft} D_{\mathrm{LD}}\left(\mathbf{R}_{ft}^{\mathrm{SP}} | \sum_k \sum_o \mathbf{J}_{fo}^{\mathrm{Tri}} z_{ko} w_{fk} h_{kt}\right).\end{aligned} \tag{5.4}$$

To minimize the log-determinant divergence between two SCMs, we again employ the MM algorithm [45] to reduce the cost monotonically.

The multiplicative update rules for the model parameters z_{ko}, w_{fk}, and h_{kt} are given by

$$z_{ko} \leftarrow z_{ko} \sqrt{\frac{\sum_{ft} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right)^{-1} \mathbf{R}_{ft}^{\mathrm{SP}} \left(\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) w_{fk} h_{kt}}{\sum_{ft} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) w_{fk} h_{kt}}}, \tag{5.5a}$$

Algorithm 2 Extension to tridiagonal approximation

```
Require: Mixture $\mathbf{x}_{ft}$

  Compute $\tilde{\mathbf{R}}_{ft} = \mathbf{x}_{ft}\mathbf{x}_{ft}^H$
  Apply spatial transform to $\tilde{\mathbf{R}}_{ft}$ with (4.2)
  Apply spatial transform to $\mathbf{J}_{fo}$ with (4.6)
  Obtain tridiagonal approximation $\hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}$ with (5.2)
  Initialize $z_{ko}, w_{fk}, h_{kt}$ with randomized values
  Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}$ with (5.3)

  for $\jmath = 0$ to MM iteration do
    $z_{ko} \leftarrow$ (5.5a)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}$ with (5.3)
    $w_{fk} \leftarrow$ (5.5b)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}$ with (5.3)
    $h_{kt} \leftarrow$ (5.5c)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}$ with (5.3)
  end for

  Cluster NMF components based on [87, 88]
  Apply Wiener filtering with (4.7)

Ensure: Estimates $\hat{\mathbf{s}}_{ft}^i$
```

$$w_{fk} \leftarrow w_{fk} \sqrt{\frac{\sum_{on} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right)^{-1} \mathbf{R}_{ft}^{\mathrm{SP}} \left(\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) z_{ko} h_{kt}}{\sum_{on} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) z_{ko} h_{kt}}}, \tag{5.5b}$$

$$h_{kt} \leftarrow h_{kt} \sqrt{\frac{\sum_{fo} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right)^{-1} \mathbf{R}_{ft}^{\mathrm{SP}} \left(\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) z_{ko} w_{fk}}{\sum_{fo} \mathrm{tr}\left(\left(\hat{\mathbf{R}}_{ft}^{\mathrm{Tri}}\right)^{-1} \hat{\mathbf{J}}_{fo}^{\mathrm{Tri}}\right) z_{ko} w_{fk}}}. \tag{5.5c}$$

Even in the case of a huge number of microphones, A, the inversion of the estimated SCM is not costly when employing the Thomas algorithm [94], which has a computational complexity of $O(A^2)$. The proposed tridiagonal algorithm is given in Algorithm 2.

5.2.2 Extension to Self-Clustering

Inspired by the prior works [44, 53] in which the models have clustering capability, we extended (4.12) as follows to possess a grouping factor g_{ik}:

$$\hat{\gamma}_{aft} = \sum_i \sum_k \sum_o \beta_{afo} z_{io} g_{ik} w_{fk} h_{kt}. \tag{5.6}$$

The update rules are derived in a similar way to in Sec. 4.2.5 and are given by

Algorithm 3 Extension to self-clustering

```
Require: Mixture $\mathbf{x}_{ft}$

  Compute $\tilde{\mathbf{R}}_{ft} = \mathbf{x}_{ft}\mathbf{x}_{ft}^H$
  Apply spatial transform to $\tilde{\mathbf{R}}_{ft}$ with (4.2)
  Apply spatial transform to $\mathbf{J}_{fo}$ with (4.6)
  Extract diagonal part of $\mathbf{F}^H\mathbf{J}_{fo}\mathbf{F}$ with (4.9)
  Initialize $z_{io}, g_{ik}, w_{fk}, h_{kt}$ with randomized values
  Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}}$ with (4.5)

  for $\jmath = 0$ to MM iteration do
    $z_{io} \leftarrow$ (5.7a)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}}$ with (4.5)
    $g_{ik} \leftarrow$ (5.7b)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}}$ with (4.5)
    $w_{fk} \leftarrow$ (5.7c)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}}$ with (4.5)
    $h_{kt} \leftarrow$ (5.7d)
    Compute $\hat{\mathbf{R}}_{ft}^{\mathrm{SP}}$ with (4.5)
  end for

  Apply Wiener filtering with (4.7)

Ensure: Estimates $\hat{\mathbf{s}}_{ft}^i$
```

$$z_{io} \leftarrow z_{io}\sqrt{\frac{\sum_{kft}\sum_a \frac{\gamma_{aft}}{\hat{\gamma}_{aft}^2}\beta_{afo}g_{ik}w_{fk}h_{kt}}{\sum_{kft}\sum_a \frac{1}{\hat{\gamma}_{aft}}\beta_{afo}g_{ik}w_{fk}h_{kt}}}, \tag{5.7a}$$

$$g_{ik} \leftarrow g_{ik}\sqrt{\frac{\sum_{oft}\sum_a \frac{\gamma_{aft}}{\hat{\gamma}_{aft}^2}\beta_{afo}z_{io}w_{fk}h_{kt}}{\sum_{oft}\sum_a \frac{1}{\hat{\gamma}_{aft}}\beta_{afo}z_{io}w_{fk}h_{kt}}}, \tag{5.7b}$$

$$w_{fk} \leftarrow w_{fk}\sqrt{\frac{\sum_{iot}\sum_a \frac{\gamma_{aft}}{\hat{\gamma}_{aft}^2}\beta_{afo}z_{io}g_{ik}h_{kt}}{\sum_{iot}\sum_a \frac{1}{\hat{\gamma}_{aft}}\beta_{afo}z_{io}g_{ik}h_{kt}}}, \tag{5.7c}$$

$$h_{kt} \leftarrow h_{kt}\sqrt{\frac{\sum_{ifo}\sum_a \frac{\gamma_{aft}}{\hat{\gamma}_{aft}^2}\beta_{afo}z_{io}g_{ik}w_{fk}}{\sum_{ifo}\sum_a \frac{1}{\hat{\gamma}_{aft}}\beta_{afo}z_{io}g_{ik}w_{fk}}}. \tag{5.7d}$$

In [44], it was shown that incorporating self-clustering into the iterations improved their performance. We compare self-clustering with various other post-clustering approaches in Sec. 5.3.4 and observe similar behavior for the average performance over all instruments. Note that the self-clustering extension can also be applied to the tridiagonal case. The proposed self-clustering extension can be found in Algorithm 3.

Table 5.1. Experimental setup

Sampling rate	16 kHz
STFT frame size	1024
STFT hop size	512

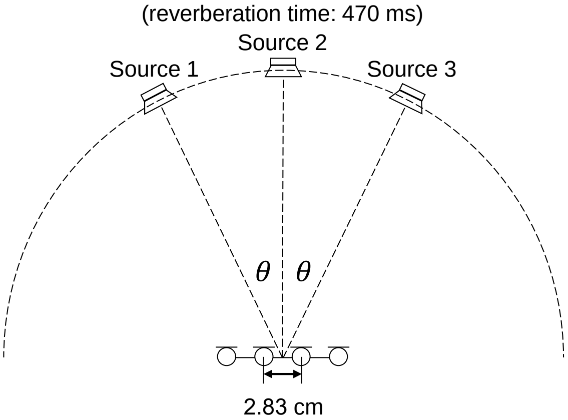


Fig. 5.1. Relationship between source positions with respect to the center of the microphone array.

5.3 Evaluation

5.3.1 Experimental Conditions

To evaluate the proposed algorithms, Algorithms 1, 2, and 3, we conducted various experiments in the case of a small number of microphones. The experimental conditions are listed in Table 5.1. As an evaluation metric, we employed the SDR improvement, which can be computed by subtracting the outputs of BSS Eval Toolbox [89] for the original mixture from the BSS Eval values of the separations. Furthermore, the SIR improvement and the SAR are also given for the experiment in Sec. 5.3.2. Please note that we did not compute SAR improvements as the input SAR is infinity and, therefore, the improvement is not computable. Moreover, we omit the SIR improvement and the SAR for the other experiments because the tendencies of these metrics are the same as that of the corresponding SDR improvement, as can be seen in Sec. 5.3.2.

5.3.2 Comparison with Other Methods under Reverberant Conditions

Reverberant room conditions were simulated to compare the proposed method with various other methods under realistic conditions. To answer the issue addressed in Sec. 5.1 regarding the case of imperfect diagonalization of SCM for small A, the number of microphones was set to $A = 4$ to observe our proposed method in the case of a small number of microphones. Note that the case of many microphones was investigated in [96], where we studied an example with $A = 32$ microphones and showed the effectiveness of our diagonal approximation in this scenario. Three instrument signals of 10 seconds each were taken from the SiSEC database 2008 for the task of underdetermined speech and music mixtures. The angle between adjacent sources and the array center is denoted as θ, as shown in Fig. 5.1. To create reverberant signals, the RWCP impulse response [99] was used and convolved with the source signals. The reverberation time of the selected room response was 470 ms, which corresponds to the reverberation of a standard conference room. The angle θ was varied from 10 to 30° to verify the robustness to angle differences. To exclude the dependence of random initializations, the average over 10 trials per angle was computed and the result is labeled as "mean" in Fig. 5.3. The maximum value over 10 trials is also shown to observe the potential of each algorithm and the corresponding bar is labeled as "max" in Fig. 5.3. The locations of the three sources were rotated three times to avoid the bias of spatial locations. In addition to the proposed Algorithm 1 (Diag) and Algorithm 2 (Tridiag) with $O = 5$ kernels, five different algorithms were evaluated as baselines: minimum variance distortionless response (MVDR) beamformer, IVA [77], independent low-rank matrix analysis (ILRMA) [78], full-rank MNMF [44], and MNMF with fixed DOA kernels ($O = 5, 8$), described in Sec. 3.3 (DOANMF). For MVDR beamformer, the DOAs of all three sources as well as the oracle SCM of interferences were fed to the algorithm as prior knowledge. The cost function for ILRMA, MNMF, and DOANMF was based on the log-determinant divergence. The number of iterations for all the iterative methods was set to 100, empirically determined based on the convergence plot shown in Fig. 5.2. The NMF clustering method in [88] was carried out for MNMF, DOANMF, and the proposed algorithms. The SDR improvements, SIR improvements, and raw SARs for the three different angles between the sources are shown in Fig. 5.3. For all three angles, both the diagonal approach and the tridiagonal approach consistently outperformed the other methods. A more rigorous comparison between the proposed algorithms is described in Sec. 5.3.3. For the conventional methods, DOANMF exhibits inferior performance to MNMF, in contrast to our expectation that a DOA-based method should be robust against random initializations as reported in [46]. The performance of DOANMF was improved as the number of DOA kernels was increased from $O = 5$ to $O = 8$ (the best-performance case), but it cannot reach the performance of the proposed methods. Although ILRMA exhibits superior performance to IVA owing to the NMF-based source model on top of the IVA spatial model, it did not achieve as good results as our approaches. This is due to the fact that one of the algorithmic assumptions of ILRMA, i.e., the spatially rank-1 property for each source, does not hold under our simulated reverberant conditions. In this simulation, the STFT frame size was set to 1024 points, corresponding to a length of 64 ms, which is much shorter than the reverberation time of 470 ms and thus, no valid time-invariant demixing matrices exist in ILRMA and IVA. For MVDR beamformer, although the above mentioned prior knowledge was given to the system, it does not perform as good as ILRMA. It is natural to observe such results because MVDR beamformer's prior information on the target source is only the direct-wave direction (steering vector) without taking the reverberant components into account. Thus, in the reverberant condition, the target source component has much leakage. On the other hand, ILRMA can estimate the optimal separation matrix, which consists of multiple beamformers' weights to cancel each of interferences with their reverberant components, resulting in less leakage (the detailed mechanism has been reported in [100]).

5.3.3 Robustness to Different Angles

To evaluate the effectiveness of adding upper and lower bands to the diagonal elements in the tridiagonal approach, we further conducted an experiment comparing Algorithm 1 (Diag) and Algorithm 2 (Tridiag) by changing the angle between the sources. We assume that the tridiagonal approach is more robust against changes to the angle because it contains additional information in the lower and upper bands. The source

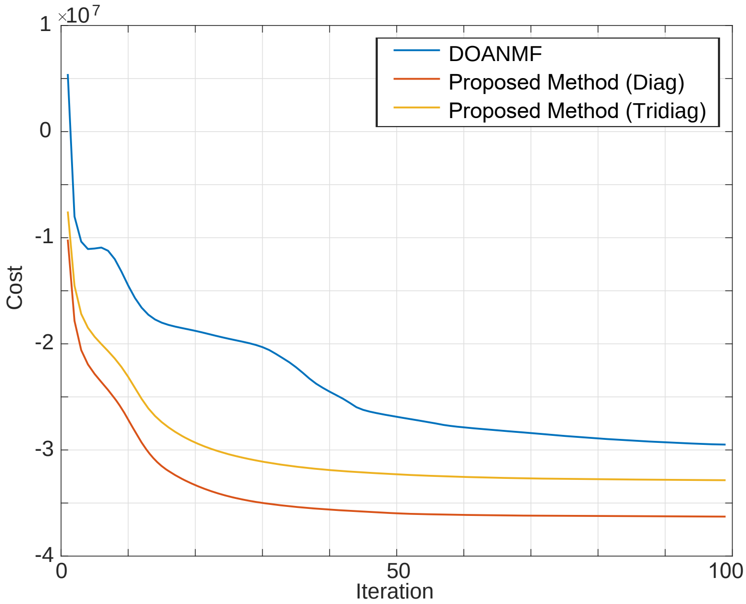


Fig. 5.2. Convergence curves for DOANMF, the proposed Algorithm 1 (Diag), and Algorithm 2 (Tridiag). The costs for DOANMF, Algorithm 1 (Diag), and Algorithm 2 (Tridiag) are computed by (3.5), (4.10), and (5.4), respectively.

Table 5.2. Maximum SDR improvements among 10 trials for different source angles under anechoic condition

Approach	10 [deg]	20 [deg]	30 [deg]	40 [deg]	50 [deg]	60 [deg]	70 [deg]	80 [deg]	Average over angles
Tridiag	**16.50**	**17.49**	**17.35**	**17.65**	**17.79**	**18.09**	**18.67**	**17.51**	**17.63**
Diag	16.15	16.78	16.29	16.32	17.43	17.45	17.43	16.80	16.83

Table 5.3. Averaged SDR improvements over 10 trials for different source angles under anechoic condition

Approach	10 [deg]	20 [deg]	30 [deg]	40 [deg]	50 [deg]	60 [deg]	70 [deg]	80 [deg]	Average over angles
Tridiag	**14.23**	13.45	**13.58**	**13.52**	**10.52**	12.87	**12.13**	8.91	**12.40**
Diag	13.88	**14.16**	12.64	7.58	10.41	**12.95**	11.55	**11.71**	11.86

angle was varied from 10 to 80°. To minimize the overlap with the previous experiment, this experiment was performed under anechoic conditions and the number of DOA kernels was set to $O = 8$. Other than the angle variation and the room conditions, the experimental settings listed in Table 5.1 were retained in this experiment. The SDR improvements are shown in Tables 5.2 and 5.3. The better result for each angle is highlighted in bold. In Table 5.2, the maximum SDR improvement among 10 trials with different initializations is given for each source angle θ. Regardless of the source angle, the results show consistent improvements upon adding band elements to the diagonal matrix. In contrast, Table 5.3 did not show the clear superiority of the tridiagonal approach over the diagonal approach. We assume that the reasons for this are twofold. First, the Thomas algorithm used in the tridiagonal approach is probably not as stable as diagonal division [94]. Second, since the tridiagonal approach involves the approximation of the SCM, the error resulting from the optimization of (5.1) is not negligible.

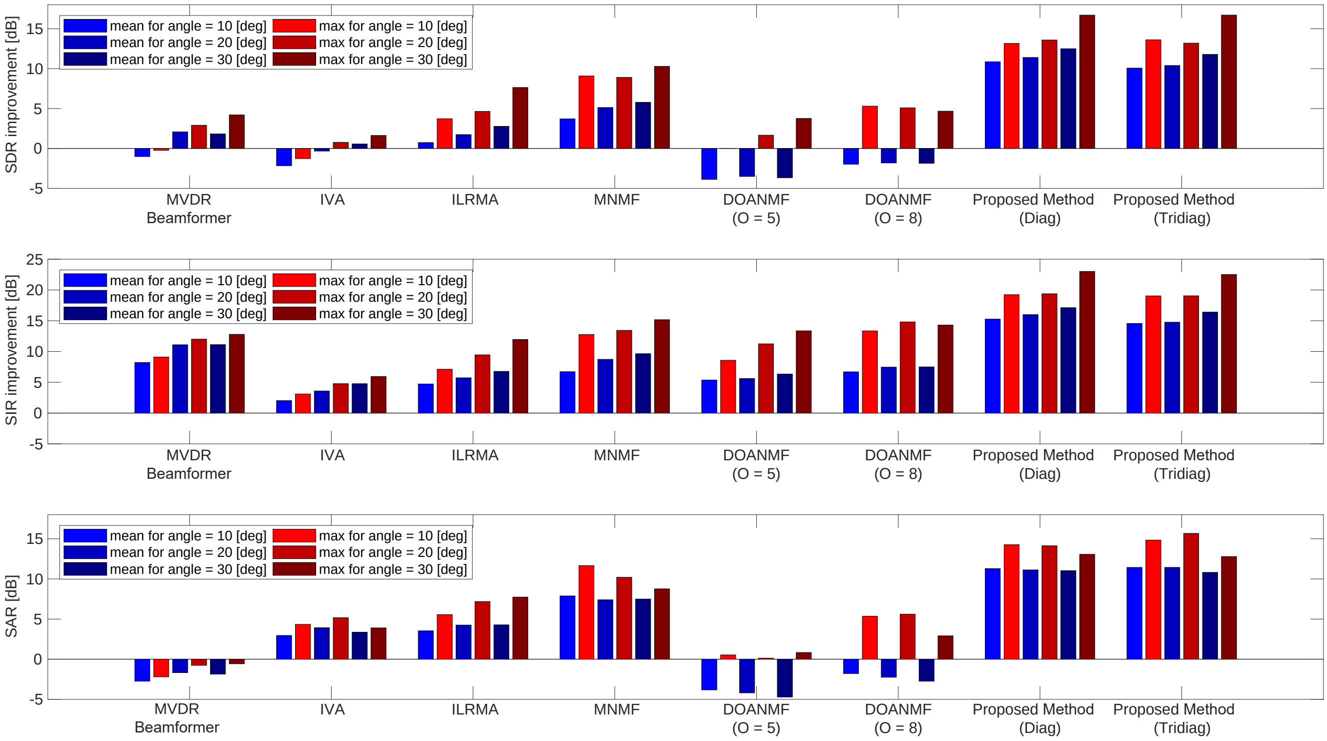


Fig. 5.3. SDR improvements, SIR improvements, and SARs under the reverberant conditions shown in Fig. 5.1, where the reverberation time is 470 ms.

5.3.4 Different Clustering Algorithms

To evaluate the self-clustering algorithm for the diagonal approach, we incorporated the state-of-the-art LPC-based clustering algorithm [87] into Algorithm 1 (Diag). Two other clustering methods [88] were also compared as baselines. The three clustering algorithms are denoted as, LPC k-medoid, MFCC k-means, and Mel-NMF. The SDR improvements for the different sources are shown in Fig. 5.4. On average, we can see that self-clustering has the best performance among the four clustering algorithms. Note that the SDR improvements for each instrument do not show a clear trend in the performance for different algorithms.

5.3.5 Underdetermined Case with Real Recordings

To take into account data obtained from more realistic acoustic conditions, we recorded music sources consisting of four musical instruments emitted by four loudspeakers with box enclosures. Three omni-directional microphones are placed with a distance of 0.45 m. The four-second long sources of musical instruments were obtained from the songKitamura dataset[*2]. Garritan Personal Orchestra 4 was chosen as MIDI source as it is considered more realistic than the other provided MIDI sources. More details about the dataset can be found in [101]. Four loudspeakers are placed in a circle with an angle θ clockwise with respect to the microphone array. The relationship between the musical instruments and the positions of the loudspeakers are listed in Table 5.4. The rough size of the room

[*2] http://d-kitamura.net/dataset.htm

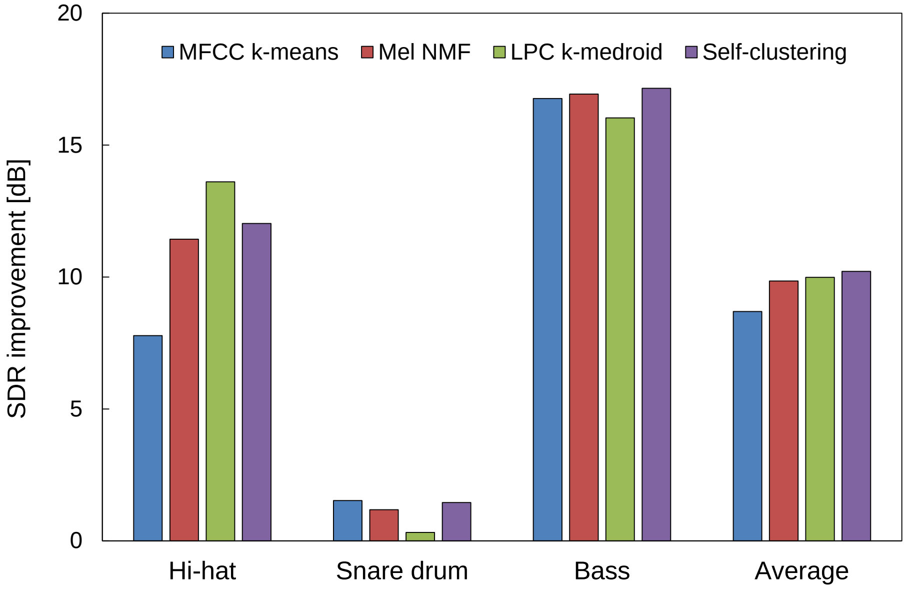


Fig. 5.4. Comparison of SDR improvement for various clustering methods. Three instrument signals (Hi-hat, Snare drum, Bass) obtained from SiSEC database were used to evaluate clustering methods.

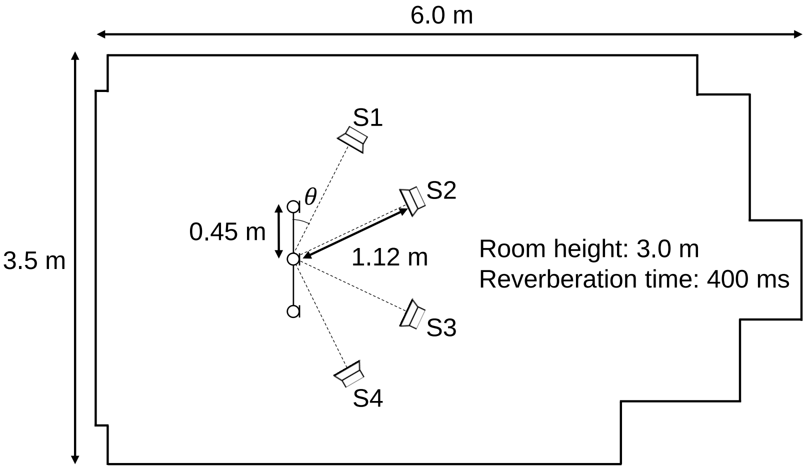


Fig. 5.5. The bird-view geometry of the common room used for the real data recordings in Sec. 5.3.5. Three microphones were linearly placed with a distance of 0.45 m while four loudspeakers were placed in a circle configuration with a radius of 1.12 m.

was 3.5 m × 6.0 m × 3.0 m and the reverberation time was 400 ms. The shape of the room can be found in Fig. 5.5. We chose four best performing methods in terms of SDR improvements shown in Sec. 5.3.2, i.e., ILRMA, MNMF, Algorithm 1 (Diag), and Algorithm 2 (Tridiag). The number of iterations was set to 500 to ensure convergence of the methods. Table 5.5 shows mean and maximum values of the SDR improvements if all four methods are run ten times. Algorithm 2 (Tridiag) outperformed the other three methods in both mean and maximum values. ILRMA does not perform as good as in Sec. 5.3.2 as it is not designed for such an underdetermined case.

Table 5.4. Relationship between musical instruments and loudspeaker positions

Musical instrument	Oboe	Flute	Piano	Trombone
Loudspeaker position	S1	S2	S3	S4
Angle θ [deg]	26.6	63.4	116.6	153.4

Table 5.5. SDR improvements under the conditions shown in Fig. 5.5

Approach	Mean	Max
ILRMA [78]	-1.44	0.47
MNMF [44]	2.65	4.83
Proposed Method (Diag)	3.63	5.17
Proposed Method (Tridiag)	**3.90**	**5.21**

Table 5.6. Computation time

Approach	Time [s]
MVDR beamformer	2.9
IVA [77]	8.7
MNMF [44]	785.2
ILRMA [78]	22.9
DOANMF	2223.3
Proposed Method (Diag)	68.0
Proposed Method (Tridiag)	1447.7

5.3.6 Computation Time

The computation times for all the methods compared in Sec. 5.3.2 were measured to investigate the efficiencies of the proposed algorithms. The measurement was carried out by inserting MATLAB time commands in the space before inputting the STFT signals and after outputting the separated time-domain signals. We ran the programs on a Xeon E5-2620 v4 CPU where each core has 2.1 GHz CPU capability. The number of DOA kernels was set to $O = 5$ as in Sec. 5.3.2. The results are listed in Table 5.6. MVDR beamformer, the non-iterative method, has the shortest computation time. By comparing Algorithm 1 (Diag) of order $O(A)$, Algorithm 2 (Tridiag) of order $O(A^2)$, and DOANMF of order $O(A^3)$, the efficiency of the proposed algorithms can be clearly observed. The speed of Algorithm 2 (Tridiag) can be further improved if a specialized matrix product that assumes that one of the two matrices is sparse is used instead of a standard matrix product. We can expect that the difference in efficiency between the algorithms will be even more significant when the number of microphones, A, is large. The diagonal approach and ILRMA have similar results because both algorithms are based on diagonal approximations of SCMs computed in the projected space. ILRMA is faster than the diagonal approach because the signals are projected to a more compact space, i.e., source space. In this experiment, the dimension of the projected space for ILRMA is equal to the number of sources, 3, whereas that for the diagonal approach is equal to the size of spatial DFT, 4.

5.4 Conclusion

In this chapter, we devised another extension that makes use of the upper and lower bands in addition to the diagonal elements to increase the accuracy of the matrix approximation. The inverse of the tridiagonal matrix can be computed with a computational complexity of $O(A^2)$ by applying the Thomas algorithm. To ensure that the tridiagonal matrix is positive semi-definite while truncating the off-diagonal elements, a semi-definite problem was solved. Moreover, the diagonal algorithm discussed in Chapter 4 was further extended by incorporating the self-clustering framework. The experimental results show that our two approximations consistently gave better results in terms of SDR improvement than various unsupervised methods. The computation time measured by MATLAB time commands showed the greater efficiency of the tridiagonal approach than MNMF in the case of four channels, and we expect that this tendency will become more significant when the number of microphones is large.

Chapter 6

Diagonal Spatial Covariance Matrix in Spherical Harmonic Domain

Please note that the content of Chapter 6 is currently under review as a journal paper. Hence, this part is omitted in the abridged version.

Chapter 7

Conclusions

In this thesis, we focused on NMF-based BSS for a large number of microphones to address the issues of existing methods, e.g., the computational costs of MNMF. To explain the background and our motivation to work on this topic, we introduced in Chapter 1 several methods including MNMF, NTF, and variants of them, and showed their application to source detection. In Chapter 2, we introduced the mathematical representation of sound fields, which is necessary to understand the concept of converting signals to the spatial frequency domain. We also introduced in Chapter 3 the model used in the well-known MNMF model and its variants. We derived in detail the update rules that are obtained by the MM algorithm as this approach is used throughout this thesis. In Chapter 4–6, we described the details of our models exploiting the characteristics of wave fields. First, we applied a DFT transform to multichannel STFT vectors to convert them to the wavenumber domain. By focusing on the diagonal parts of a SCM, the cost function of DOANMF in the wavenumber domain was simplified to a NTF cost function, which can reduce the computational cost from $O(A^3)$ to $O(A)$. The experiments were performed for the case of 32 channels, and the approach was proven to be effective in both improving the separation quality while at the same time reducing the computational complexity. Second, to generalize our method to the case where a small number of microphones are used, we attempted devising a method exploiting that the SCM has a tridiagonal structure. The extension to self-clustering was also examined in this chapter. The evaluations including captured signals by a real recording were carried out to see the performance of our methods, and they were shown to outperform other methods, e.g., ILRMA. Third, to examine the robustness of our approach to spherical arrays, the algorithm was modified to work in the SH domain. To tackle the problem of boosting the higher-order SH coefficients in the low frequency regime when the inverse radial function is applied, a masking scheme was applied as a weighting for the cost function. The model was further simplified by omitting the frequency dependence, enabling the system to become more efficient. The experiments showed that our proposed method can deal with various settings.

In summary, we proposed in this thesis a unified theory for multichannel BSS that is applicable to various arrays generally used in sound field reproductions. The conversion to the spatial frequency domain enables MNMF-based methods to be simplified in terms of the computational complexity. It also open doors to exploit the characteristics of wave fields in BSS, which bridges a gap between two different signal processing fields.

Acknowledgement

I would first of all like to thank my supervisors, Prof. Hiroshi Saruwatari and Dr. Shoichi Koyama, for all the guidance that they have provided. I would also like to thank Norihiro Takamune for helping me conduct experiments. Your advice was always spot on and it helped me move forward with my project.

Next, I would like to thank my colleague, Dr. Stefan Uhlich at Sony, for giving me a lot of valuable comments. I would also like to thank all the other people at Sony, especially Dr. WeiHsiang Liao and Keiichi Osako, who sometimes took over my part when I was too busy.

Last but not the least, I want to thank my wife, my lovely daughters, and my parents for always being there for me and supporting me during six years of my PhD.

本論文は、筆者が東京大学大学院 情報理工学系研究科 システム情報学専攻 猿渡・小山研究室に後期博士課程として在学した3年間および満期退学後の3年間の合計6年間の成果をまとめたものです。

Bibliography

[1] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio—the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, 2015.

[2] L. J. Griffiths, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Antennas and Propagation Society*, vol. 30, no. 1, pp. 22–34, 1982.

[3] H. Cox, R. M. Zeskind, and M. M. Owen, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, no. 10, pp. 22–34, 1987.

[4] A. Hyvarinen, "Survey on independent component analysis," *Neural Computing Surveys 2*, pp. 94–128, 1999.

[5] K. Torkkola, "Blind separation for audio signals - are we there yet?" in *Proc. Workshop on Independent Component Analysis and Blind Signal Separation*, 1999.

[6] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* Wiley, 2009.

[7] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, 2004.

[8] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, 2007.

[9] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proc. of International Conference on Digital Audio Effects (DAFx)*, 2003, pp. 209–213.

[10] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals : Demixing N sources from 2 mixture," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 2000, pp. 2985–2988.

[11] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 2135–2139.

[12] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[13] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 261–265.

[14] S. I. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.

[15] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band DenseNets for audio source separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 21–25.

[16] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 106–110.

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 556–562.

[18] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct.

[19] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proc. International Computer Music Conference (ICMC)*, 2003.

[20] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[21] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proc. International Conference on Digital Audio Effects (DAFx)*, 2009.

[22] R. Jaiswal, D. Fitzgerald, D. Barry, E. Coyle, and S. Rickard, "Clustering NMF basis functions using shifted NMF for monaural sound source separation," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 245–248.

[23] J. M. Becker, M. Spiertz, and V. Gnann, "A probability-based combination method for unsupervised clustering with application to blind source separation," in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012, pp. 99–106.

[24] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012, pp. 322–329.

[25] Z. Duan, G. Mysore, and P. Smaragdis, "Online PLCA for real-time semi-supervised source separation," in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012, pp. 34–41.

[26] T. O. Virtanen, "Monaural sound source separation by perceptually weighted non-negative matrix factorization," 2007.

[27] C. Févotte, *Itakura–Saito nonnegative factorizations of the power spectrogram for music signal decomposition.* IGI Global Press, Aug. 2010, ch. 11.

[28] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with beta-divergence," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2010, pp. 283 –288.

[29] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 78–81.

[30] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 775–782.

[31] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," in *Proc. IEEE Workshop*

on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2009, pp. 129–132.

[32] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[33] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[34] D. Lee, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[35] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. International Conference on Information Science, Signal Processing and Their Applications (ISSPA)*, 2010, pp. 1–4.

[36] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *Proc. Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 33–40.

[37] ——, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[38] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[39] E. Vincent, "Complex nonconvex l_p norm minimization for underdetermined source separation," in *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 430–437.

[40] A. Jourjine, S. Rickard, and Ö. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, pp. 2985–2988.

[41] S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares.* Cambridge University Press, 2018.

[42] A. Ozerov, C. Févotte, R. Blouet, and J. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 257–260.

[43] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura–Saito nonnegative matrix factorization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2012, pp. 261–264.

[44] ——, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[45] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[46] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[47] J. J. Carabias-Orti, J. Nikunen, T. Virtanen, and P. Vera-Candeas, "Multichannel blind sound source separation using spatial covariance model with level and time differences and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio,*

Speech and Language Processing, vol. 26, no. 9, pp. 1512–1527, 2018.

[48] D. FitzGerald, M. Cranitch, and E. Coyle, “Non-negative tensor factorisation for sound source separation,” in *Proc. Irish Signals and Systems Conference (ISSC)*, 2005.

[49] A. Shashua and T. Hazan, “Non-negative tensor factorization with applications to statistics and computer vision,” in *Proc. International Conference on Machine Learning (ICML)*, 2005, pp. 792–799.

[50] D. FitzGerald, M. Cranitch, and E. Coyle, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” in *Proc. Irish Signals and Systems Conf. (ISSC)*, 2005.

[51] A. Ozerov, C. Févotte, R. Blouet, and J. Durrieu, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 257–260.

[52] D. FitzGerald, M. Cranitch, and E. Coyle, “Extended nonnegative tensor factorisation models for musical sound source separation,” *Computational Intelligence and Neuroscience*, 2008.

[53] C. Févotte and A. Ozerov, “Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues,” in *Proc. International Symposium on Computer Music Modeling and Retrieval (CMMR)*, 2010, pp. 102–115.

[54] J. Nikunen, T. Virtanen, and M. Vilermo, “Multichannel audio upmixing based on non-negative tensor factorization representation,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 33–36.

[55] F. Cong, A. H. Phan, Q. Zhao, A. K. Nandi, V. Alluri, P. Toiviainen, H. Poikonen, M. Huotilainen, A. Cichocki, and T. Ristaniemi, “Analysis of ongoing EEG elicited by natural music stimuli using nonnegative tensor factorization,” in *Proc. European Signal Processing Conference (EUSIPCO)*, 2012, pp. 494–498.

[56] N. D. Stein, “Nonnegative tensor factorization for directional blind audio source separation,” *CoRR*, vol. abs/1411.5010, 2014.

[57] M. Fakhry, P. Svaizer, and M. Omologo, “Audio source separation in reverberant environments using β-divergence-based nonnegative factorization,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 7, pp. 1462–1476, 2017.

[58] P. Smaragdis and G. J. Mysore, “Separation by ”humming”: User-guided sound extraction from monophonic mixtures,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009.

[59] O. Dikmen and A. T. Cemgil, “Unsupervised single-channel source separation using bayesian nmf,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009.

[60] S. Ewert and M. Müller, “Using score-informed constraints for nmf-based source separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[61] Y. Mitsufuji and A. Roebel, “Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[62] ——, “On the use of a spatial cue as prior information for stereo sound source separation based on spatially weighted non-negative tensor factorization,” *EURASIP Journal on Advances in Signal Processing*, vol. 2014, p. 40, 2014.

[63] F. Weninger, B. Schuller, M. Wöllmer, and G. Rigoll, “Localization of non-linguistic

events in spontaneous speech by non-negative matrix factorization and long short-term memory," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5840–5843.

[64] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Proc. of International Workshop on Machine Listening in Multisource Environments (CHiME)*, 2011, pp. 36–40.

[65] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*. Springer, Berlin, Germany, 2012, pp. 341–371. [Online]. Available: http://hal.inria.fr/hal-00708805

[66] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. Independent Component Analysis and Blind Signal Separation (ICA)*, 2004, pp. 494–499.

[67] C. Lopes and F. Perdigão, "Speech event detection by non-negative matrix deconvolution," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2007.

[68] C. Cotton and D. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, oct. 2011, pp. 69–72.

[69] D. FitzGerald and M. Cranitch, "Sound source separation using shifted non-negative tensor factorisation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[70] Y. Mitsufuji, M. Liuni, A. Baker, and A. Roebel, "Online non-negative tensor deconvolution for source detection in 3DTV audio," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 3082–3086.

[71] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of Audio Engineering Society*, pp. 1004–1025, 2005.

[72] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda, "Analytical approach to wave field reconstruction filtering in spatio-temporal frequency domain," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 685–696, 2013.

[73] S. Koyama, K. Furuya, Y. Haneda, and H. Saruwatari, "Source-location-informed sound field recording and reproduction," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 881–894, 2015.

[74] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography.* Academic Press, 1999.

[75] J. Ahrens, *Analytic Methods of Sound Field Synthesis.* Springer Berlin Heidelberg, 2012.

[76] B. Kulis, M. A. Sustik, and I. S. Dhillon, "Low-rank kernel learning with Bregman matrix divergences," *Journal of Machine Learning Research*, vol. 10, pp. 341–376, 2009.

[77] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[78] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[79] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[80] T. Higuchi and H. Kameoka, "Unified approach for underdetermined BSS, VAD, dereverberation and DOA estimation with multichannel factorial HMM," in

Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2014, pp. 562–566.

[81] S. Lee, S. H. Park, and K. Sung, “Beamspace-domain multichannel nonnegative matrix factorization for audio source separation,” *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 43–46, 2012.

[82] D. Fitzgerald, A. Liutkus, and R. Badeau, “Projection-based demixing of spatial audio,” *IEEE/ACM Transaction on Audio, Speech Language Processing*, vol. 24, no. 9, pp. 1560–1572, 2016.

[83] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 276–280.

[84] T. Taniguchi and T. Masuda, “Linear demixed domain multichannel nonnegative matrix factorization for speech enhancement,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 476–480.

[85] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small - room acoustics,” *The Journal of the Acoustical Society of America (JASA)*, vol. 65, no. 4, pp. 943–950, 1979.

[86] S. M. Kay, *Fundamentals of Statistical Signal Processing.* Prentice Hall PTR, 1993.

[87] X. Guo, S. Uhlich, and Y. Mitsufuji, “NMF-based blind source separation using a linear predictive coding error clustering criterion,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 261–265.

[88] M. Spiertz and V. Gnann, “Source-filter based clustering for monaural blind source separation,” in *Proc. International Conference on Digital Audio Effects (DAFx)*, 2009.

[89] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[90] “In signal separation evaluation campaign (SiSEC 2008): http://www.sisec.wiki.irisa.fr.”

[91] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[92] R. M. Gray, “Toeplitz and circulant matrices: A review,” Tech. Rep., 2001.

[93] J. Allen and D. Berkley, “Image method for efficiently simulating small room acoustics,” *The Journal of the Acoustical Society of America (JASA)*, vol. 65, no. 4, 1979.

[94] L. H. Thomas, “Elliptic Problems in Linear Differential Equations over a Network,” Columbia University, Tech. Rep., 1949.

[95] J. Nikunen and T. Virtanen, “Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 6677–6681.

[96] Y. Mitsufuji, S. Koyama, and H. Saruwatari, “Multichannel blind source separation based on non-negative tensor factorization in wavenumber domain,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 56–60.

[97] J. Löfberg, “YALMIP: A toolbox for modeling and optimization in MATLAB,” in *Proc. IEEE International Conference on Robotics and Automation*, 2004.

[98] J. F. Sturm, “Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones,” *Optimization Methods and Software*, vol. 11, no. 1-4, pp. 625–653,

1999.

[99] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” in *Proc. International Conference on Language Resources and Evaluation (LREC)*, pp. 965–968.

[100] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, “Blind spatial subtraction array for speech enhancement in noisy environment,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 650–664, 2009.

[101] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, “Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 654–669, 2015.

[102] D. FitzGerald, M. Cranitch, and E. Coyle, “On the use of the beta divergence for musical source separation,” in *Proc. Irish Signals and Systems Conference (ISSC)*, 2008.

[103] S. Doclo and M. Moonen, “Gsvd-based optimal filtering for single and multimicrophone speech enhancement,” *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[104] T. J. Klasen, M. Moonen, T. V. den Bogaert, and J. Wouters, “Preservation of interaural time delay for binaural hearing aids through multi-channel wiener filtering based noise reduction,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[105] N. Bertin, C. Févotte, and R. Badeau, “A tempering approach for Itakura–Saito non-negative matrix factorization. with application to music transcription,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 1545–1548.

List of Publications

Journal papers and articles

[J1] Yuki Mitsufuji, Axel Roebel, ”On the Use of a Spatial Cue as Prior Information for Stereo Sound Source Separation Based on Spatially Weighted Non-Negative Tensor Factorization,” *EURASIP Journal of Advancement of Signal Processing*, issue 1, 2014.

[J2] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, Yuki Mitsufuji, ”Open-Unmix - A Reference Implementation for Music Source Separation,” *Journal of Open Source Software*, vol. 4, no. 41, pp. 1667, 2019.

[J3] Yuki Mitsufuji, Stefan Uhlich, Norihiro Takamune, Daichi Kitamura, Shoichi Koyama, Hiroshi Saruwatari, ”Multichannel Non-Negative Matrix Factorization Using Banded Spatial Covariance Matrices in Wavenumber Domain,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 49–60, 2019.

[J4] Yu Maeno, Yuki Mitsufuji, Prasanga N Samarasinghe, Naoki Murata, Thushara D Abhayapala, ”Spherical-Harmonic-Domain Feedforward Active Noise Control Using Sparse Decomposition of Reference Signals from Distributed Sensor Arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 656–670, 2019.

[J5] Tetsu Magariyachi, Yuki Mitsufuji, ”Analytic Error Control Methods for Efficient Rotation in Dynamic Binaural Rendering of Ambisonics,” *Journal of the Acoustical Society of America*, vol. 147, issue 1, 2020.

[J6] Jihui Aimee Zhang, Naoki Murata, Yu Maeno, Prasanga N. Samarasinghe, Thushara D. Abhayapala, Yuki Mitsufuji, ”Coherence-Based Performance Analysis on Noise Reduction in Multichannel Active Noise Control Systems,” *Journal of the Acoustical Society of America*, vol. 148, issue 3, 2020.

[J7] Yuki Mitsufuji, Norihiro Takamune, Shoichi Koyama, Hiroshi Saruwatari, ”Multichannel Blind Source Separation Based on Evanescent-Region-Aware Non-Negative Tensor Factorization in Spherical Harmonic Domain,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (conditionally accepted), 2020.

International conferences

[C1] Yuki Mitsufuji, Axel Roebel, ”Sound Source Separation Based on Non-Negative Tensor Factorization Incorporating Spatial Cue as Prior Knowledge,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 71–75, 2013.

[C2] Yuki Mitsufuji, Marco Liuni, Alex Baker, Axel Roebel, ”Online Non-Negative Tensor Deconvolution for Source Detection in 3DTV Audio,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3082–3086, 2014.

[C3] Xin Guo, Stefan Uhlich, Yuki Mitsufuji, ”NMF-Based Blind Source Separation Using a Linear Predictive Coding Error Clustering Criterion,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 261–265, 2015.

[C4] Stefan Uhlich, Franck Giron, Yuki Mitsufuji, ”Deep Neural Network Based Instrument Extraction from Music,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2135–2139, 2015.

[C5] Yuki Mitsufuji, Shoichi Koyama, Hiroshi Saruwatari, ”Multichannel Blind Source Separation Based on Non-Negative Tensor Factorization in Wavenumber Domain,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 56–60, 2016.

[C6] Keiichi Osako, Yuki Mitsufuji, Rita Singh, Bhiksha Raj, ”Supervised Monaural Source Separation Based on Autoencoders,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 11–15, 2017.

[C7] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, Yuki Mitsufuji, ”Improving Music Source Separation Based on Deep Neural Networks Through Data Augmentation and Network Blending,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 261–265, 2017.

[C8] Naoya Takahashi, Yuki Mitsufuji, ”Multi-Scale Multi-Band DenseNets for Audio Source Separation,” in *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 21–25, 2017.

[C9] Yu Maeno, Yuki Mitsufuji, Thushara D. Abhayapala, ”Mode Domain Spatial Active Noise Control Using Sparse Signal Representation,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 211–215, 2018.

[C10] Yuki Mitsufuji, Asako Tomura, Kazunobu Ohkuri, ”Creating a Highly-Realistic” Acoustic Vessel Odyssey” Using Sound field Synthesis with 576 Loudspeakers,” in *Proceedings of Audio Engineering Society (AES) Conference on Spatial Reproduction-Aesthetics and Science*, 2018.

[C11] Naoya Takahashi, Nabarun Goswami, Yuki Mitsufuji, ”MMDenseLSTM: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation,” in *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[C12] Yu Maeno, Yuki Mitsufuji, Prasanga N Samarasinghe, Thushara D Abhayapala, ”Mode-domain Spatial Active Noise Control Using Multiple Circular Arrays,” in *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 441–445, 2018.

[C13] Joachim Muth, Stefan Uhlich, Nathanael Perraudin, Thomas Kemp, Fabien Cardinaux, Yuki Mitsufuji, ”Improving DNN-based Music Source Separation Using Phase Features,” in *Proceedings of Joint Workshop on Machine Learning for Music at ICML, IJCAI/ECAI and AAMAS*, 2018.

[C14] Naoya Takahashi, Purvi Agrawal, Nabarun Goswami, Yuki Mitsufuji, ”PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation,” in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2713–2717, 2018.

[C15] Wei-Hsiang Liao, Yuki Mitsufuji, Keiichi Osako, Kazunobu Ohkuri, ”Microphone Array Geometry for Two Dimensional Broadband Sound Field Recording,” in *Proceedings of 145th Audio Engineering Society (AES) Convention*, 2018.

[C16] Naoya Takahashi, Sudarsanam Parthasaarathy, Nabarun Goswami, Yuki Mitsufuji, ”Recursive Speech Separation for Unknown Number of Speakers,” in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1348–1352, 2019.

[C17] Naoki Murata, Jihui Zhang, Yu Maeno, Yuki Mitsufuji, ”Global and Local Mode-Domain Adaptive Algorithms for Spatial Active Noise Control Using Higher-Order Sources,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 526–530, 2019.

[C18] Naoya Takahashi, Mayank Kumar Singh, Sakya Basak, Parthasaarathy Sudarsanam, Sriram Ganapathy, Yuki Mitsufuji, ”Improving Voice Separation by Incorporating End-To-End Speech Recognition,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 41–45, 2020.

[C19] Yu Maeno, Yuhta Takida, Naoki Murata, Yuki Mitsufuji, ”Array-Geometry-Aware Spatial Active Noise Control Based on Direction-of-Arrival Weighting,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 8414–8418, 2020.

Appendix

Incorporation of Spatial Cues and Spatially Weighted NTF

The following text appeared in the journal article [J1].

1 Introduction

In the last few decades, NMF has become one of the most prevalent techniques to tackle the underdetermined source separation problem where the number of sources is greater than or equal to the number of observations. NMF is based on the idea that a mixture is a composite of a number of object basis elements, each of which represents an underlying characteristic of the sources. Estimation is carried out by simple matrix factorization, with all the elements being non-negative. Cost function for NMF estimation has long been investigated by several researchers [17, 27, 28]. In particular, the Itakura–Saito divergence is known to be an appropriate cost function for approximation of audio spectra due to its scale-invariant nature. To complete NMF-based source separation, clustering of the decomposed basis elements follows the factorization to properly classify them to corresponding sources. A large number of related techniques have been developed so far [21, 22, 23].

Taking advantage of prior information for the purpose of enhancing the performance of NMF has been widely investigated. Smaragdis et al. have attempted to make use of a user-guided humming for the extraction of melodies in a mixture [58]. Diknen et al. investigated the Bayesian NMF model assigning different prior distributions for tonal and percussive signals [59]. Ewert et al. presented an extended approach that uses additional score information to guide the NMF process [60]. Although a number of researches related to the prior knowledge of TF features have been introduced, it is not yet common to incorporate a spatial feature (hereafter, spatial cue) in NMF due to its algorithm framework. Since NMF which separates a single-channel signal does not produce any spatial information during the separation process, it has been difficult to associate a spatial cue until the emergence of MNMF.

NTF, known as one of the MNMF techniques, extends the NMF idea to tensors. An n-way tensor is a generalization of the mathematical concepts of scalar, vector, and matrix (e.g., a two-way tensor is a matrix). Specifically, a three-way tensor, which can be regarded as a collection of multichannel spectrograms, is being investigated for use in NTF [49, 51, 52]. Extension to the third dimension provides another matrix that describes the energy distribution of each basis component on every channel, which can also be regarded as spatial information. This technique enables the NMF approach to be adapted to easily accept a spatial cue [61].

This Appendix proposes a promising method to enhance NTF performance, taking advantage of a spatial cue given by users or from accompanying images. The enhancement is mainly achieved by introducing weights on bin-wise NTF cost functions, which differentiates a target component from other components. Since a spatial cue indicates which bins of the tensor spectrogram are important, it is possible to improve the quality of an approximation to the specific bins of the tensor by giving more weights to bins where the target is likely to exist and less weights to the others. Virtanen et al. proposed perceptually weighted NMF that provides perceptually motivated weights for each critical band in each frame in accordance with the loudness perception of the human auditory system [26]. Nevertheless, to our knowledge, no research regarding NMF that incorporates the weighting function for spatially focusing on target component estimation has been proposed. The evaluation results show that this method is advantageous in terms of separation quality over conventional PARAFAC-NTF and other source separation techniques such as DUET [6, 7, 8].

It should be noted that apart from NTF, there exist other approaches to address the source separation problem in multichannel audio. Especially, the algorithms based on local Gaussian models [30] using the SCM for encoding spatial positions of source signals [31] have been shown to outperform the simpler NTF approach that will be used in the following. A rank-1 convolutive model assuming target sources existed in a non-reverberant environment has been proposed in [33]. Full-rank unconstrained model with NMF was further introduced by Arberet et al. to account for reverberant conditions [32, 35]. A modular framework for these algorithms has been presented in [36, 37]. These models present a significant improvement in terms of separation quality as compared to the NTF model. A problem with these models that use the GEM algorithm for optimization is the significant increase in computational complexity when compared to the NTF model. Sawada et al. have addressed this problem by introducing multiplicative update rules in place of the GEM algorithm, but still their optimization requires significantly more computation time compared to single-channel NMF [44]. While we believe that the proposed weighting scheme can improve performance with other existing multichannel factorization algorithms, we selected the NTF algorithm for investigation into the effectiveness of our weighting scheme to limit the computational costs.

2 NTF-based Source Separation

2.1 Non-negative Tensor Factorization

A multichannel audio signal that has been transformed into a set of spectrograms (one for each of the A channels) can be regarded as a three-way tensor $\mathbf{V}$ and approximated by $\widehat{\mathbf{V}}$. $\widehat{\mathbf{V}}$ is created as a superposition of K feature tensors, each produced by means of an outer product of three vectors $\mathbf{q}_k$, $\mathbf{w}_k$, and $\mathbf{h}_k$, respectively representing the channel, frequency, and time factors of the feature tensor. To adapt the NTF representation $\widehat{\mathbf{V}}$ to the target tensor spectrogram $\mathbf{V}$, the following optimization problem is solved:

$$\min_{\mathbf{Q},\mathbf{W},\mathbf{H}} \sum_{a,f,t} \xi_{aft} D_\beta(v_{aft}|\hat{v}_{aft}) + \lambda(\mathbf{H}) \quad \text{s.t. } \mathbf{Q},\mathbf{W},\mathbf{H} \geq 0 \ , \tag{A.1}$$

with

$$\hat{v}_{aft} = \sum_k q_{ak} w_{fk} h_{kt} \ .$$

Here the matrices $\mathbf{Q}$, $\mathbf{W}$, and $\mathbf{H}$ are assembled from the vectors $\mathbf{q}_k$, $\mathbf{w}_k$, and $\mathbf{h}_k$, having elements q_{ak}, w_{fk}, and h_{kt}. The elements of the tensor $\widehat{\mathbf{V}}$ are denoted as v_{aft} where a indicates the channel index, f the bin index of the spectrogram, and t the time index of the spectrogram. $\lambda(\mathbf{H})$ represents additional constraints on matrix $\mathbf{H}$, which are taken into account during minimization of the cost function. The β-divergence, D_β, is suitable for NTF, allowing the separation quality to be changed, subject to the parameter β [102]. When β is equal to 2, 1, or 0, the NTFs are called EUC-NTF, KL-NTF, or IS-NTF, respectively. ξ_{aft} denotes one of the bins of the weighting tensor, $\mathbf{G}$, in bin-wise β-divergence. It allows controlling the impact of the error observed in the different elements of $\mathbf{V}$. For standard PARAFAC-NTF, $\xi_{aft} = 1$ for all the bins.

The update rules for training the three matrices are derived from the derivatives of the cost function:

$$\mathbf{Q} \leftarrow \mathbf{Q} \odot \left(\frac{\langle \mathbf{G} \odot \mathbf{V} \odot \widehat{\mathbf{V}}^{\odot(\beta-2)}, \mathbf{W} \circ \mathbf{H} \rangle_{\{2,3\},\{1,2\}}}{\langle \mathbf{G} \odot \widehat{\mathbf{V}}^{\odot(\beta-1)}, \mathbf{W} \circ \mathbf{H} \rangle_{\{2,3\},\{1,2\}}} \right)^{\odot\gamma(\beta)} , \tag{A.2}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\frac{\langle \mathbf{G} \odot \mathbf{V} \odot \widehat{\mathbf{V}}^{\odot(\beta-2)}, \mathbf{Q} \circ \mathbf{H} \rangle_{\{1,3\},\{1,2\}}}{\langle \mathbf{G} \odot \widehat{\mathbf{V}}^{\odot(\beta-1)}, \mathbf{Q} \circ \mathbf{H} \rangle_{\{1,3\},\{1,2\}}} \right)^{\odot\gamma(\beta)} , \tag{A.3}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\langle \mathbf{G} \odot \mathbf{V} \odot \widehat{\mathbf{V}}^{\odot(\beta-2)}, \mathbf{Q} \circ \mathbf{W} \rangle_{\{1,2\},\{1,2\}} + \nabla^-_{\mathbf{H}}\lambda(\mathbf{H})}{\langle \mathbf{G} \odot \widehat{\mathbf{V}}^{\odot(\beta-1)}, \mathbf{Q} \circ \mathbf{W} \rangle_{\{1,2\},\{1,2\}} + \nabla^+_{\mathbf{H}}\lambda(\mathbf{H})} \right)^{\odot\gamma(\beta)} , \tag{A.4}$$

where $\nabla_{\mathbf{H}}\lambda(\mathbf{H}) = \nabla^+_{\mathbf{H}}\lambda(\mathbf{H}) - \nabla^-_{\mathbf{H}}\lambda(\mathbf{H})$, both $\odot$ and / denote element-wise calculations, $\mathbf{A} \circ \mathbf{B}$ denotes $A \times F \times K$ tensor with elements $a_{ak}b_{fk}$ when $\mathbf{A}$ and $\mathbf{B}$ are $A \times K$ and $F \times K$ [53], and $\langle \mathbf{A}, \mathbf{B} \rangle_{\{C\},\{D\}}$ denotes a contracted product [6]. Setting parameter γ to the proper value guarantees that the cost function decreases monotonically when $\xi_{aft} = 1$ for all the bins and the constraints are zero [28].

2.2 Wiener Filtering

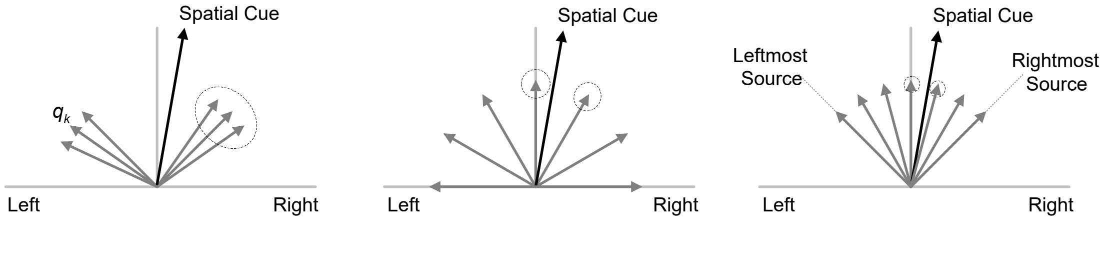
As soon as the approximation of spectrogram $\widehat{\mathbf{V}}$ composed of multiple basis elements $\mathbf{Q}$, $\mathbf{W}$, and $\mathbf{H}$ has been completed by NTF, Wiener filtering is followed to extract the target signal such that

$$y_{aft} = \frac{\sum_{k \in K_{\mathrm{tar}}} q_{ak} w_{fk} h_{kt}}{\hat{v}_{aft}} x_{aft} , \tag{A.5}$$

where K_{tar} denotes the collection of bases considered as the target group. x_{aft} and y_{aft} denote STFT of input audio signal and the separated target signal, respectively. It should be noted that the more sophisticated method called multichannel Wiener filter employing spatial covariance matrices is known to give a better performance in more complex mixing scenario [103, 104].

3 Incorporation of Spatial Cues

We devised two ways of incorporating a spatial cue. Fig. A.1 shows a two-dimensional representation of a channel matrix, $\mathbf{Q}$, for 2ch-stereo signals. The small arrows represent the basis elements of the channel matrix. Their positions depend on the values for each channel: for example, the basis element $q_k = [0.5, 0.5]^T$ means that the source is coming from the center, and the basis element $q_k = [0.9, 0.1]^T$ means that the source is closer to the left channel. It can be represented more intuitively by the following equation,



[idea1] PARAFAC-NTF

[idea2] fixed-Q NTF

Proposed method

Fig. A.1. two-dimensional representation of a channel matrix $\mathbf{Q}$. Three different solutions incorporating NTF.

$$\overleftarrow{Q}_k = 2\tan^{-1}\left(\frac{q_{1,k}}{q_{0,k}}\right)^{\delta}, \tag{A.6}$$

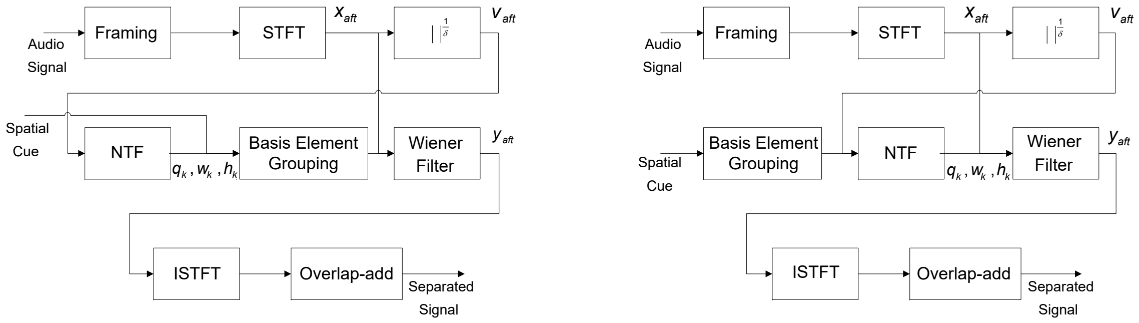
where $\delta = 1.0$ when magnitude spectrums are used as an input of NTF and $\delta = 0.5$ for power spectrums. $\overleftarrow{Q}$ denotes the angles of the arrows in radians clockwise from the horizontal axis in Fig. A.1. The big arrow indicates a spatial cue that is specified independently from outside. It is totally independent of the positions of the small arrows at this moment, and it is given in the same format as the elements of $\overleftarrow{Q}$, specifically called $\overleftarrow{Q}_{\mathrm{sc}}$. However, it should be noted that these arrows serve only for visualization purposes and are different from the azimuth angles in the real world.

Idea 1 (Fig. A.1, left) applies standard PARAFAC-NTF to audio signals. Factorization produces the channel matrix, $\mathbf{Q}$, the elements of which will be linked to the spatial cue at the end of the process. This may, however, pose a problem when the spatial cue is far from the basis element candidates (Fig. A.1, left). Interpolation between the two groups (three arrows on the left and three arrows on the right) in another space might not be helpful in creating sound in the direction of the spatial cue. However, idea 1 yields good performance when the spatial cue and the basis element candidates are sufficiently close to each other. We call idea 1 PARAFAC-NTF (p-NTF).

In idea 2 (Fig. A.1, center), the basis elements of the channel matrix, $\mathbf{Q}$, are evenly spaced before the start of the NTF process. The directions remain fixed throughout the process while matrix $\mathbf{W}$ and matrix $\mathbf{H}$ are trained by means of NTF update rules. The basis elements located at both sides of the spatial cue are selected as target elements. We call idea 2 fixed-Q NTF (f-NTF), and each direction is numbered with a direction index, o, ranging from 0 to $O - 1$.

Fig. A.2 shows block diagrams of two different solutions. The big difference that can be observed in this figure is that in p-NTF, the spatial cue cannot be passed to the NTF process, whereas in f-NTF it is, which allows NTF to take advantage of spatial information. Two things make idea 2 worth focusing on: the computational efficiency, since the channel matrix, $\mathbf{Q}$, does not have to be updated thanks to spatial information provided from a spatial cue; and the potential improvement in quality due to the prior knowledge provided by the spatial cue.

The next section presents more details and a variant of the f-NTF method that is called spatial cue (sc-NTF).



[idea1] PARAFAC-NTF (p-NTF)

[idea2] fixed-Q NTF (f-NTF)

Fig. A.2. Block diagrams of two different solutions. Block diagrams for two different solutions using NTF.

4 Model and Derivation of Update Rules

4.1 Choice of Divergence

As mentioned in Sec. 2.1, three settings of the β-divergence are commonly used for NMF and NTF: Euclidean distance ($\beta = 2$), KL divergence ($\beta = 1$), and Itakura–Saito (IS) divergence ($\beta = 0$). Their differences were investigated by C. Févotte [27]. One important characteristic of the IS divergence that is not shared with the two other types of divergence is that the absolute scale of given audio does not affect the total cost of the divergence. That is, the unnoticeably small spectrogram bins can be approximated as well as the dominant bins. We assume that IS-NTF is thus more appropriate when a relatively small signal might come from a direction close to that of the spatial cue. However, this assumption is probably true only when there is little ambient noise [63]. Thus, we selected IS-NTF for our initial experiments and used noise-free input signals, such as commercial music. Another motivation for employing IS divergence comes from a statistical perspective. It has been shown that the ML estimation of a sum of complex Gaussian components representing for each spectral bin is equivalent to minimizing IS divergence between the ideal and estimated power spectrograms [20]. While there are clear advantages of the IS divergence, IS-NTF suffers from the fact that it is more often caught in local minima.

The simplest solution to this problem might be to perform a number of training runs and then select the best results from among them. Another approach to mitigating this effect is tempering NTF by changing the type of divergence during the iterations [105]. For example, the training could start with EUC-NTF (NTF based on Euclidean distance), which is relatively robust with regard to local minima, and finish up with IS-NTF, which produces better results. This would require that developers carefully control β, and more iterations than usual would probably be needed.

4.2 Initialization of Channel Matrix

The initialization of channel matrix, $\mathbf{Q}$, is based on the understanding of the matrix Q explained in Section 3.

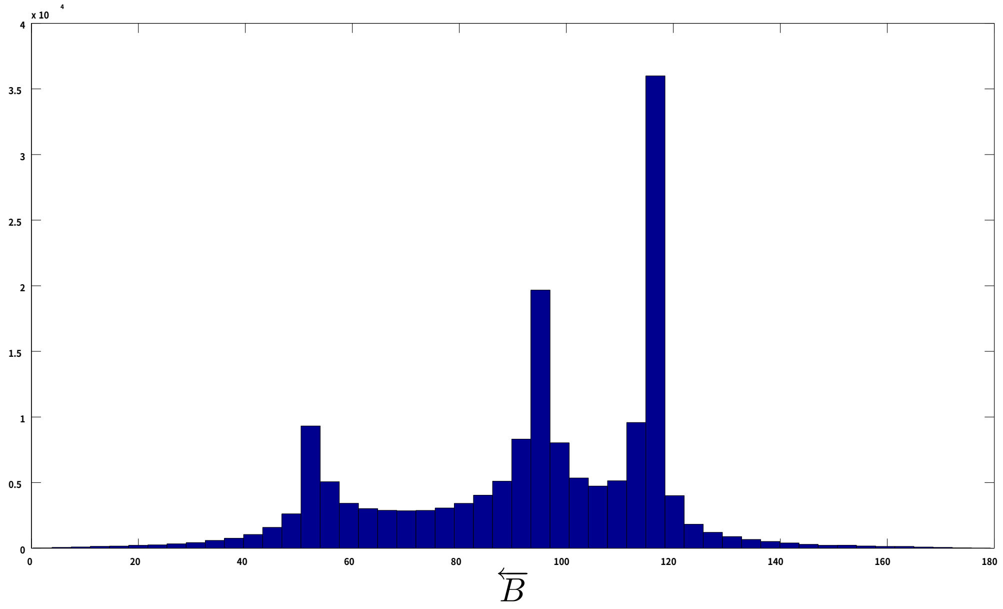


Fig. A.3. Histogram of $\overleftarrow{B}$ of the three-source mixture. Three peaks are observed.

$$\overleftarrow{B}_{ft} = 2\tan^{-1}\left(\frac{v_{1,ft}}{v_{0,ft}}\right)^{\delta}, \tag{A.7}$$

where $v_{0,ft}$ and $v_{1,ft}$ are the spectrogram bins for the left and right channels, respectively. This feature concerns the arrows in Fig. A.1 and their relationships to each spectrogram bin. It is possible to determine the locations of sources with respect to the bins by searching for the peaks in the histogram of $\overleftarrow{B}$ (Fig. A.3), which represents the dominant presence of the sources. The basis elements are preferentially allocated by initializing channel matrix, $\mathbf{Q}$, based on the histogram of $\overleftarrow{B}$: More elements are allocated to directions where sources are likely to exist, although some are allocated to cover all directions. Fig. A.3 shows the histogram of $\overleftarrow{B}$ calculated from a mixture of three audio instruments placed in different positions. The length of arrows corresponds to a frequency of each bin. As we can see the arrows in three directions in the histogram, it is highly likely that sources exist at 50° to 60°, 90° to 100°, and 110° to 120°. However, the allocation of basis elements to the left of the leftmost peak and to the right of the rightmost peak is not required since the superposition of the sources never appears outside of these peaks. It is therefore preferable to allocate basis elements inside the range spanned by the left and rightmost peaks that exist in the measured histogram of $\overleftarrow{B}$, as can be seen in the right image of Fig. A.1.

4.3 Initialization of Frequency Matrix and Time Matrix

Initialization of frequency matrix, $\mathbf{W}$, and time matrix, $\mathbf{H}$, is simply carried out by taking advantage of information of the histogram such that

$$w_{fk} = \frac{1}{N_{Grp(o)}A}\sum_{a}\sum_{t\in Grp(o)} v_{aft} \quad f,t,k \in Grp(o), \tag{A.8}$$

$$h_{kt} = \frac{1}{N_{Grp(o)}A}\sum_{a}\sum_{f\in Grp(o)} v_{aft} \quad f,t,k \in Grp(o), \tag{A.9}$$

where $N_{Grp(o)}$ denotes the frequency per bin of the histogram, and $Grp(o)$ denotes the collection of bases allocated to the direction with the index o. Normalization of the matrices follows for the purpose of concentrating the energy of input tensor $\mathbf{V}$ into $\mathbf{H}$.

$$\mathbf{Q}_k = \mathbf{Q}_k/|\mathbf{Q}_k|_1,\ \mathbf{W}_k = \mathbf{W}_k/|\mathbf{W}_k|_1, \mathbf{H}_k = |\mathbf{Q}_k|_1|\mathbf{W}_k|_1\mathbf{H}_k, \tag{A.10}$$

$$e_o = \sum_{o}\sum_{k\in Grp(o)} |\mathbf{H}_k|_1, \tag{A.11}$$

where $|\cdot|_1$ denotes the l1-norm, and e_o denotes the directional energy associated with the direction index o.

4.4 Weighting Function

Since the spatial cue indicates which direction should be given preference, and since the histogram of $\overleftarrow{B}$ indicates which source is dominant for a given direction, it is possible to approximate the spectrogram bin associated with the spatial cue more precisely than other bins. This is easy to accomplish by using the proper weighting tensor, $\mathbf{G}$, in the cost function:

$$\xi_{aft} = \exp\left(-\frac{\Psi}{O}\left|\frac{\overleftarrow{Q}_{\mathrm{sc}} - \overleftarrow{Q}_k}{\Delta\overleftarrow{Q}}\right|\right) \quad f,t,k \in Grp(o), \tag{A.12}$$

where Ψ determines the shape of the exponential function. Fig. A.4 changes the weighting parameter Ψ and $\overleftarrow{Q}_k$ that creates different shapes while forcing $\overleftarrow{Q}_{\mathrm{sc}}$ to point toward 100°. The weighting values of different Ψ when $\overleftarrow{Q}_k = 72$ are accentuated with markers. When Ψ equals 0, all the weights for bin-wise cost functions become 1, which boils down the update rules of sc-NTF described in Sec. 2.1 to those used for PARAFAC-NTF.

4.5 Constraints

The energy for each direction should be estimated by adding all the basis elements in matrix $\mathbf{H}$ over time, equal to the procedure done by (A.11). The estimated energy is fixed so that it can be used as a reference to constrain the energy distribution of the estimated tensor. This should reduce the likelihood of being trapped in local minima. Here, we again use the IS divergence to measure distance. The constraint on energy in a given direction is

$$\lambda(\hat{v}_{aft}) = \mu\sum_{o=0}^{O-1} D_{\mathrm{IS}}\left(\sum_{aft\in Grp(o)} v_{aft} \;\middle|\; \sum_{aft\in Grp(o)} \hat{v}_{aft}\right). \tag{A.13}$$

By taking into account the normalization procedure in (A.10), the equation can be boiled down to

$$\lambda(\mathbf{H}) = \mu\sum_{o=0}^{O-1} D_{\mathrm{IS}}\left(e_o \;\middle|\; \sum_{k\in Grp(o)} |\mathbf{H}_k|_1\right), \tag{A.14}$$

$$\text{s.t. } \mathbf{Q}_k = \mathbf{Q}_k/|\mathbf{Q}_k|_1,\ \mathbf{W}_k = \mathbf{W}_k/|\mathbf{W}_k|_1, \mathbf{H}_k = |\mathbf{Q}_k|_1|\mathbf{W}_k|_1\mathbf{H}_k.$$

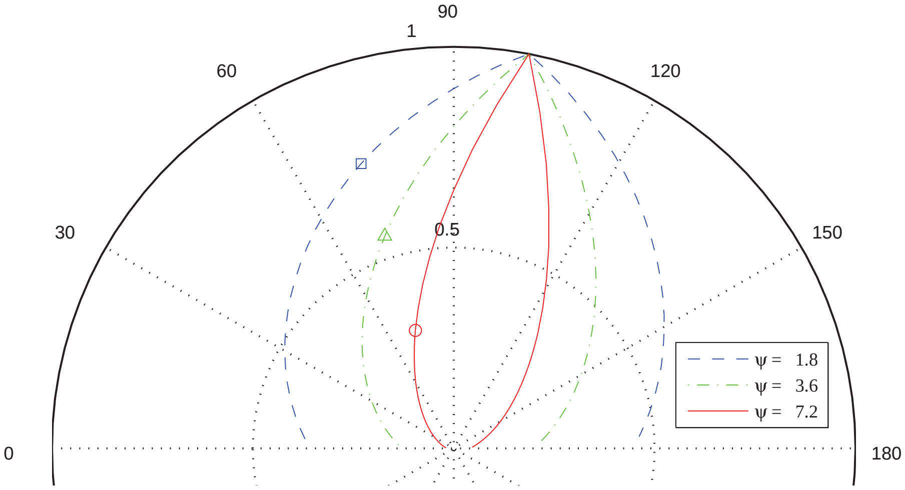


Fig. A.4. Spatial representation of the weighting function, **G**. The weighting function controlled its shape by the weighting parameter Ψ. Three different shapes are shown when $\overleftarrow{Q}_{\mathbf{sc}} = 100$.

For IS-NTF, the following should hold for the derivative of the constraint:

$$\nabla\lambda(h_k) = \frac{\mu}{e_o} - \frac{\mu \sum_{k \in Grp(o)} |\mathbf{H}_k|_1}{(e_o)^2}. \tag{A.15}$$

5 Evaluation

5.1 Separation Quality

BSS Eval of MATLAB was used to evaluate the above method. It gives three standard metrics for source separation: signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) [89]. SDR is a global measure of the quality of source separation that encompasses the two other metrics, SIR indicates how well the target source is separated from interference, and SAR indicates how well the target source retains sound quality after separation. sc-NTF was compared with p-NTF and f-NTF. 2ch-stereo signals were obtained from the 'Signal Separation Evaluation Campaign' Web site (SiSEC 2008 [90]). More specifically, we used development data from the underdetermined speech and music mixture task. These 2ch-stereo sources contain a number of instruments placed independently in a two-dimensional field. The ground truth registered in a similar format as $\overleftarrow{Q}$ was also obtained from SiSEC 2008. It gives the location of each instrument.

For f-NTF, after separation is conducted, the basis elements pointing in a direction close to the ground truth are selected to be separated out. In contrast, p-NTF requires grouping after training. This difference can be seen in Fig. A.2. Our experiments on two grouping algorithms, k-means and k-nearest neighbor, to the ground truth showed that

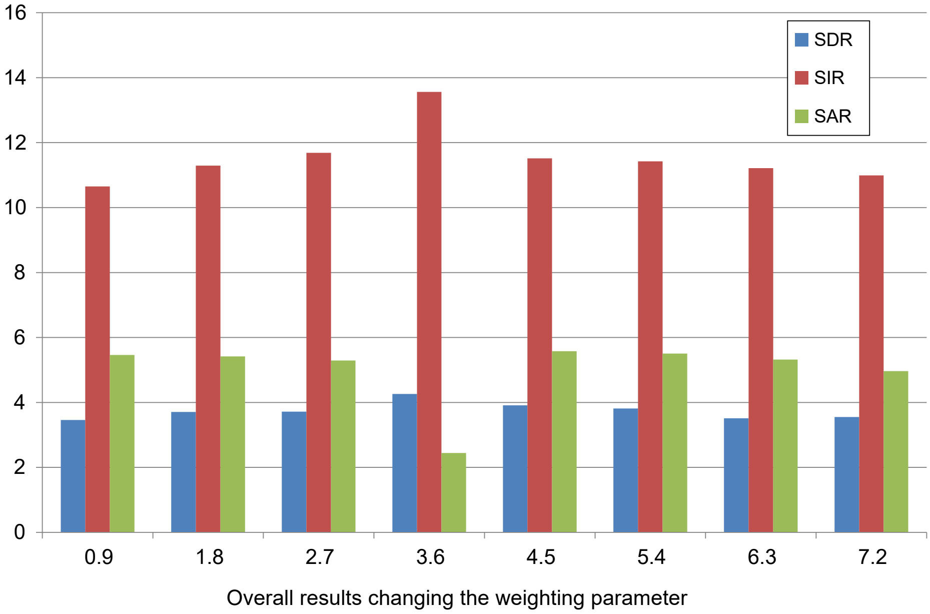


Fig. A.5. Test results of different shapes of the weighting function. Overall results of different settings of Ψ is shown, from 0.9 to 7.2.

the latter yielded better results. The number of centroids is determined according to the number of basis components.

On the other hand, for sc-NTF, the bases are selected beforehand due to the link established between the allocated basis elements and the spatial cue. Resynthesis is followed by Wiener filtering to create the output signals used for evaluation (1,024-point FFT, half overlap, the number of the basis elements $K = 90$, and the number of directions $O = 18$). Tests were run 10 times to obtain an average, indicated by a bar, and 95% confidence, indicated by a line on top of the bar. This test was almost exactly the same as the one described by Févotte et al. in their 2011 paper [53]. The only difference was in the number of bases: They used $K = 9$ and we used $K = 90$. γ in the cost function was set to 1. We obtained better results when $\Psi = 3.6$ for the weighting tensor $\mathbf{G}$ and $\mu = 300$ for the constraint. More details with results using different settings of Ψ can be found in Fig. A.5.

Fig. A.6 shows test results for harmonic sound (nodrums), and Fig. A.7 shows results for percussive sound (wdrums). Most of the results show that sc-NTF outperforms both f-NTF and p-NTF. It is important to note that these results were obtained by sacrificing the accuracy of approximation of sources far from the spatial cue. This can be deduced from the final value of the cost function. Table A.1 shows the IS divergence per bin for four methods including sc-NTF without the weighting function $\mathbf{G}$. sc-NTF using the weighting function produces worse results than p-NTF and sc-NTF without the weighting function in terms of approximation, due to the lesser weights of the broad range and the more weights of the relatively narrow target direction. The 95% confidence for both p-NTF and sc-NTF indicates that local minima were avoided. There is a large variance only in the results for f-NTF, which means that the proposed method helps to avoid being trapped in local minima.

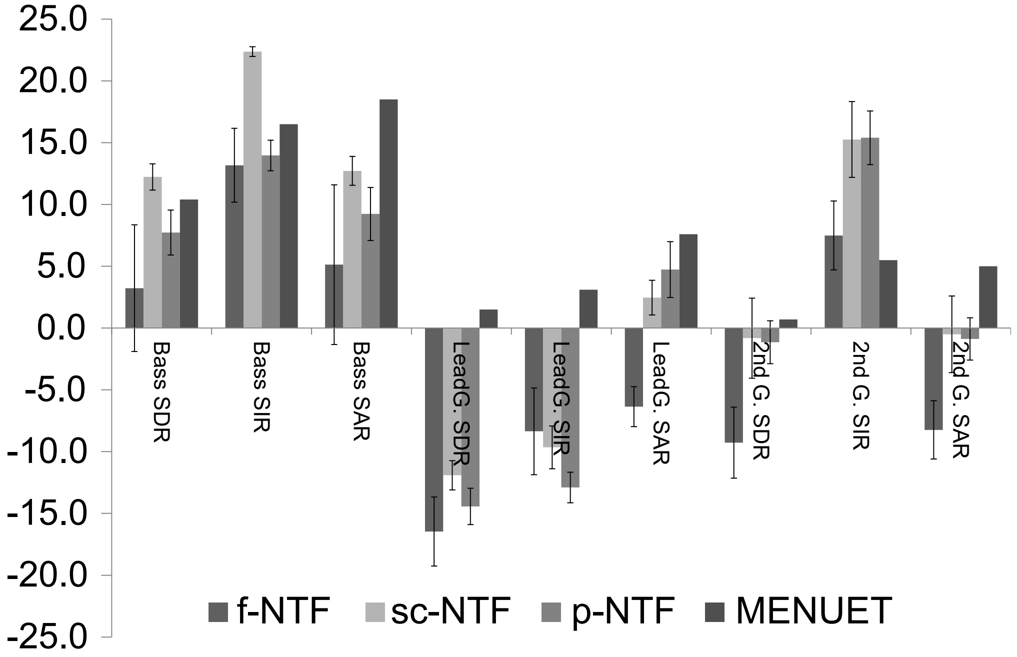


Fig. A.6. Test results for harmonic sound. Dataset nodrums. Mixture of three different instruments: bass, lead guitar, and second lead guitar.

Table A.1. **Comparison of convergence**

	f-NTF	sc-NTF (w/o **G**)	sc-NTF	p-NTF
IS divergence per bin	0.19300	**0.16759**	0.18300	0.16857

Itakura–Saito divergence per bin after 200 iterations.

5.2 Computational Cost

As mentioned earlier, the omission of the calculation of the channel matrix $\mathbf{Q}$ should be a great advantage for sc-NTF. A simple experiment was conducted by measuring the runtime on an Intel Core i7 (2.80 GHz) processor for three NTF methods implemented in MATLAB code. It should be noted that the code was not particularly optimized by incorporating external libraries written in, for instance, C language. Instead, the built-in functions automatically provided by MATLAB, such as division and log functions, were used. The conditions of the experiment were the same as those for the quality evaluation described in Sec. 5.1 (the number of channels $A = 2$, the number of frequency bins $F = 513$, and the number of frames $T = 314$). Fig. A.8 shows that p-NTF takes almost four times as much computational power as the other NTF methods. Although sc-NTF requires initialization involving division, inverse tangents, and calculation of the histogram, the runtime is only slightly longer than that of f-NTF. Since the size of each matrix is different due to the required resolution, the computational power needed to update each matrix is not the same. Each matrix needs at least two contracted products

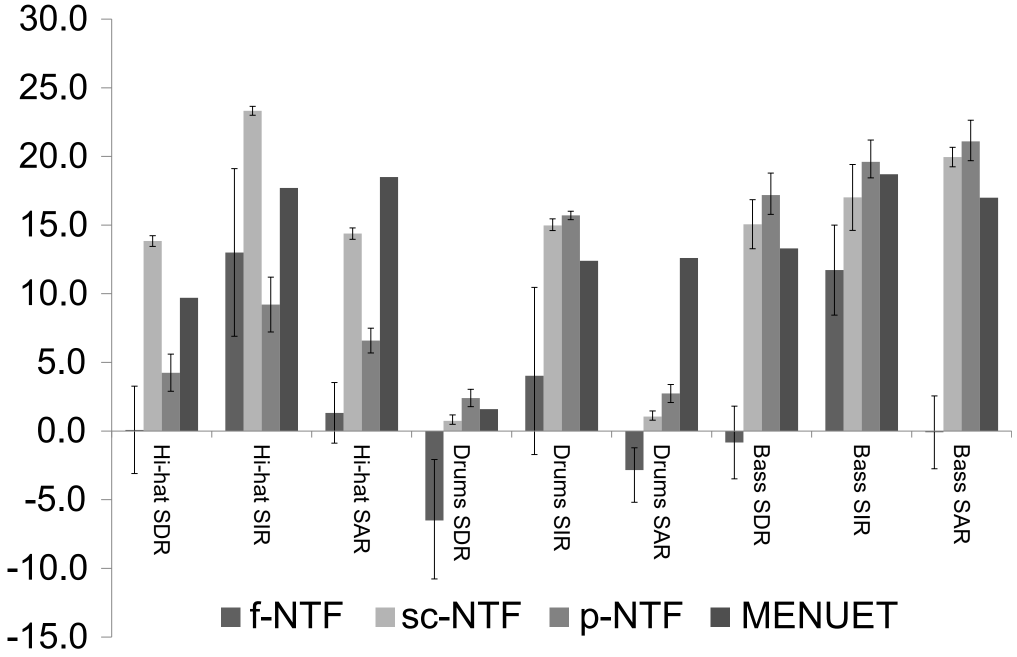


Fig. A.7. Test results for percussive sound. Dataset wdrums. Mixture of three different instruments: hi-hat, drums, and bass.

and single division: $2 \times F \times T + A \times K$ for the channel matrix, $2 \times A \times T + F \times K$ for the frequency matrix, and $2 \times A \times F + T \times K$ for the time matrix. Thus, the results should depend on the numbers of occurrences of A, F, T, and K. In most cases, $A << F$ or T, which means that updating will probably take longer for the channel matrix, $\mathbf{Q}$, than for the other two matrices. Another important point is that the burden of the built-in functions depends on the features of the LSI used to implement the NTF: Division generally needs a couple of instructions, but multiplication and addition usually need only one.

5.3 Comparison with DUET

An evaluation that compares sc-NTF with DUET has been carried out. Since a number of evolved versions of DUET have been proposed, we employed one of the most straightforward extensions of DUET called Multiple Sensor DUET (MENUET) [8]. Although the framework of MENUET requires the number of sources for the clustering at the end of the process, it is usually unknown in real world, particularly in applications such as Sound Zoom. Instead of providing the number of sources, which can be regarded as providing another piece of prior knowledge, we incorporated a spatial cue in the DUET system so that the closest centroid to the spatial cue can be extracted as a target centroid. The number of centroids is set to 18 to fully cover all the azimuths. Fig. A.6 and A.7 show the comparison results between two different source separation techniques. In particular, the results of the two guitars give us a deep insight such that MENUET performs better than sc-NTF when separating such signals that have similar frequency characteristics, on

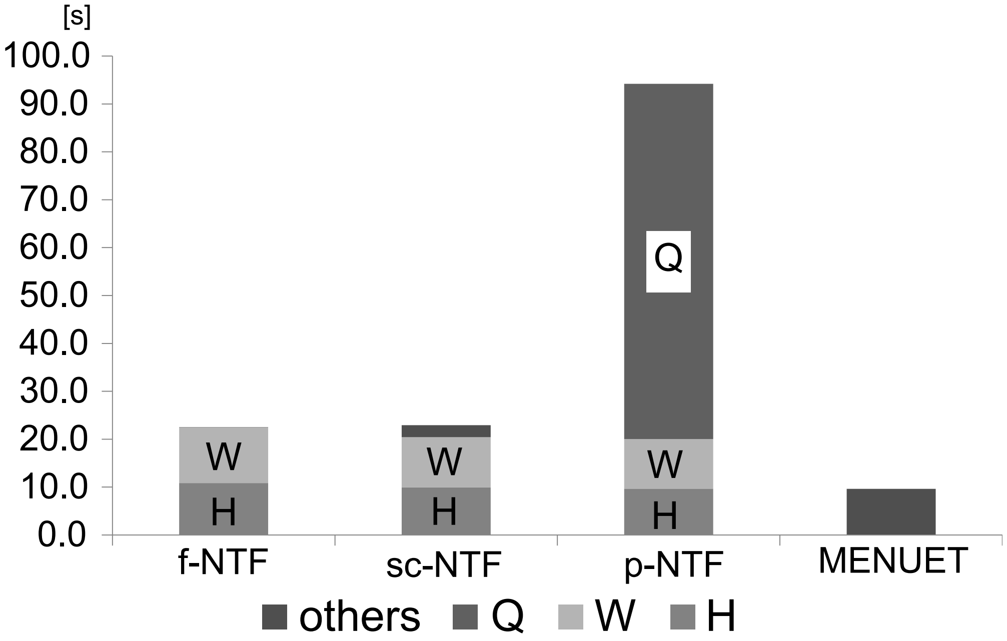


Fig. A.8. Comparison of computational cost. Runtime test for the three different NTFs. The runtime of sc-NTF includes the calculation of initialization, weighting function, and constraints.

the condition that the two sources are not coming from the same directions. The worse results of sc-NTF may be attributed to the nature of NTF that greedily exploits not only spatial information, but also TF information for the approximation of the cost function. On the other hand, sc-NTF performs better in the case of separating the two percussions in spite of the close positions of the two instruments. This is due to the capability of NMF to capture repetitive structures of signals. Computational costs of MENUET can be seen in Fig. A.8. The number of iterations for clustering in MENUET is 30.

6 Conclusion

We developed a new method of enhancing NTF performance by introducing weights on the NTF cost function, which is achieved by incorporating a spatial cue into the system. Two ways of incorporating a spatial cue into an NTF framework were devised. The one that employs a fixed channel matrix, $\mathbf{Q}$, was further developed to improve the separation quality. The association of the spatial cue with the histogram of $\overleftarrow{B}$ clarifies which spectrogram bins should be given preference to obtain a better approximation. An evaluation of separation quality, which was carried out as in previous studies, demonstrated the effectiveness of the weighting tensor, $\mathbf{G}$, and the energy constraints. In addition, the omission of the calculation of $\mathbf{Q}$ was a great advantage in the runtime test. In short, our algorithm combines the computational cost of f-NTF with the separation quality of p-NTF. sc-NTF also showed competitive results against the variant of the DUET algorithm.