Effective Optimization Algorithms for Blind and Supervised Music Source Separation with Nonnegative Matrix Factorization

KITAMURA DAICHI

Doctor of Philosophy

Department of Informatics
School of Multidisciplinary Sciences
SOKENDAI (The Graduate University for Advanced Studies)

Effective Optimization Algorithms for Blind and Supervised Music Source Separation with Nonnegative Matrix Factorization

by

Daichi Kitamura

Dissertation

submitted to the Department of Informatics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

SOKENDAI (The Graduate University for Advanced Studies)

March 2017

Committee

Advisor	Dr. Nobutaka Ono Associate Professor of National Institute of Informatics/SOKENDAI
Subadvisor	Dr. Junichi Yamagishi Associate Professor of National Institute of Informatics/SOKENDAI
Subadvisor	Dr. Hiroshi Saruwatari Professor of The University of Tokyo
Examiner	Dr. Ken Hayami Professor of National Institute of Informatics/SOKENDAI
Examiner	Dr. Imari Sato Professor of National Institute of Informatics/SOKENDAI

Acknowledgments

This dissertation is a summary of three years of study carried out at School of Multidisciplinary Sciences, SOKENDAI (The Graduate University for Advanced Studies), Japan.

I would like to express the deepest appreciation to Dr. Nobutaka Ono, Associate Professor of National Institute of Informatics and SOKENDAI, my dissertation adviser, for his continuous teaching and essential advice on both technical and non technical issues. This work could not have been accomplished without his well-directed advice, helpful suggestions, and fruitful discussions with him. I have learned many valuable aspects of being a researcher from his attitude toward study and have always enjoyed conducting research with him.

I would also like to express my deep gratitude to Dr. Hiroshi Saruwatari, Professor of The University of Tokyo, my dissertation subadviser, for his valuable guidance and constant encouragement. Without his encouragement help and guidance, this dissertation would not have materialized.

I would like to express my sincere thanks to Dr. Junichi Yamagishi, Associate Professor of National Institute of Informatics and SOKENDAI, my dissertation subadviser, for his instructive comments for my research and on the dissertation.

I would like to offer my special thanks to Dr. Ken Hayami and Dr. Imari Sato, Professors of National Institute of Informatics and SOKENDAI, members of the dissertation committee, for their valuable comments on the dissertation.

This work could not have been achieved without the collaboration of many researchers. I especially thank Dr. Hiroshi Sawada and Dr. Hirokazu Kameoka, researchers at Nippon Telegraph and Telephone Corporation, and Dr. Yu Takahashi and Dr. Kazunobu Kondo, researchers at Yamaha Corporation, for

their beneficial and valuable comments and advice.

Many staff and members of my research group have supported me in carrying out experiments and writing this dissertation at National Institute of Informatics, SOKENDAI, and The University of Tokyo; I would especially like to express my appreciation of the clerical supports for me, valuable discussions with them on technical issues, and their provision of a comfortable research environment. I also wish to express my deep gratitude to Ms. Shigeko Morimoto, a secretary of our laboratory, for her kind help and support in all aspects of my research.

I received satisfactory financial supports from Japan Society for the Promotion of Science as a Research Fellow DC1 and Research Organization of Information and Systems as a Research Assistant with Superordinate Wages during my Ph.D. pursuit. I deeply appreciate their supports of a research fund and salaries.

I appreciate the opportunity to have studied with all the members in our research group at National Institute of Informatics, SOKENDAI, and The University of Tokyo. I thank Dr. Trung-Kien Le, Dr. Keiko Ochi, and Dr. Shinji Takaki who are Postdoctoral Fellows at National Institute of Informatics, for useful discussions on this work. I would also like to thank Mr. Keisuke Imoto who is a Ph.D. student and a colleague of mine from the same laboratory for working hard together. I also thank many foreign internship students in our laboratory.

Finally, I wish to show my great appreciation to my family, Masao Kitamura, Chizu Kitamura, and Izumi Masutani, for their support over many years, and I would also like to offer my special thanks to Ms. Manami Kitamura who has always supported my academic life since I enrolled in Master Course of Nara Institute of Science and Technology.

Abstract

In this dissertation, to address a music source separation problem, several optimization algorithms are proposed. Music source separation is a technique to extract or separate specific music sources from an observed mixture signal that contains multiple music instrumental and vocal sounds. There are many feasible applications for this technique, for example, audio remixing by users, automatic music transcription, and musical instrument education. A general audio source separation problem has been investigated for a long time, particularly in the speech signal processing field to reduce background noise and enhance only the speech signal in the observation. Many techniques have been proposed for various recording conditions in the past, and they can roughly be divided into two situations: determined (or overdetermined) and underdetermined cases. In the determined situation, sufficient number of observations (microphones used in the recording) can be utilized for solving the separation problem, whereas the underdetermined situation, which includes monaural observation, basically lacks such multi-dimensional information. Also, presence of external prior information (supervision) such as music scores, source locations, or sound examples of each source in the mixture is another important issue. The source separation techniques without any prior information is often called blind source separation, which is the most difficult but a practical technique.

The objective of this dissertation is to develop an effective optimization algorithm for the music source separation and to achieve satisfactory separation performance. Two main topics are here addressed: determined (and overdetermined) blind source separation and single-channel (underdetermined) semi-supervised source separation. The semi-supervised source separation

exploits sound examples of only the target source for the separation, namely, only the target source is extracted from the mixture. In both the topics, an important property of music signals is focused to effectively capture their structures. Since typical music signals consist of limited number of components such as discrete pitches and musical notes and include many reiteration of similar or the same spectral patterns (timbers), the power spectrogram of music signals tends to have a low-rank structure. On the basis of this nature in music signals, for both the topics discussed in this dissertation, a popular algorithm of matrix decomposition called nonnegative matrix factorization (NMF) is exploited for modeling the structure of music signals. By applying NMF to the spectrogram of audio signal, the frequently appearing spectral patterns and their time-varying gains can be extracted as bases and activations. These components are useful for modeling the audio signals and achieving the source separation. For the problem of determined blind source separation, independent component analysis (ICA) and its multivariate extension, independent vector analysis (IVA), are traditional and reliable approaches and can provide good separation results particularly for a mixture signal of speech. These approaches estimate spatial demixing filters by assuming that the sources are mutually independent. This assumption is valid in a practical mixture signal and make the separation problem solvable in a fully blind fashion. However, the separation accuracy of ICA and IVA for music signals is not satisfactory. This is because the general music signals frequently contain spectral overlaps and co-occurrences between sources, which result in a harmony of music, and these properties weaken the inherent independence between the sources. Also, the both methods assume only the non-Gaussian source distribution as an unspecific source model and do not utilize any information about the structure in the spectrogram of each source. To solve this problem, in this dissertation, the unified method of NMF and IVA called independent low-rank matrix analysis (ILRMA) is proposed, which performs simultaneous estimation of the spectrogram structure of each source and their spatial demixing filters. The optimization algorithm in ILRMA ensures faster convergence, more stable performance, and better computational efficiency compared with conventional methods including multichannel extension of NMF (MNMF), which is a state-of-the-art method for source separation. Also,

theoretical relationships between IVA, MNMF, and ILRMA are revealed, namely, ILRMA is essentially equivalent to MNMF with a constraint for the mixing system, and IVA is also a special case of ILRMA.

For the single-channel semi-supervised source separation task, semi-supervised NMF, which aims to extract only the target source from the mixture, is the most popular approach. In this method, sound examples of the target source are utilized for preparing the supervised bases (spectral dictionary) of the target source. However, when the target source and the other sources in the mixture signal share similar or the same spectral patterns (bases), the separation performance of semi-supervised NMF is degraded because such shared components cannot be separated. This fact means that the supervised bases must be discriminative from the other bases of non-target sources. On the basis of this fact, in this dissertation, a new training algorithm that provides discriminative supervised bases is proposed for semi-supervised NMF. In this method, other sound examples, which are candidates of the non-target signals in the observed mixture, are utilized only for learning which spectral components will be frequently shared between the target and non-target sources. Furthermore, a new efficient initialization scheme for NMF is proposed. Since an optimization in NMF requires initial values for bases and activations, all the results of applications based on NMF always depend on the initialization. The proposed initialization is based on a maximization of mutual independence between the activations using nonnegative ICA algorithm. The efficacy of the proposed method for several source separation tasks including ILRMA and semi-supervised NMF with discriminative basis training is experimentally confirmed.

List of Abbreviations

BF	Beamforming
BSILRMA	Basis-shared independent low-rank matrix analysis
BSS	Blind source separation
DAE	Denoising autoencoder
DC	Directional clustering
DNN	Deep neural network
DOA	Direction of arrival
EM	Expectation-maximization
EU distance	Squared Euclidean distance
EUNMF	Nonnegative matrix factorization based on squared Euclidean distance
FDICA	Frequency-domain independent component analysis
FSNMF	Full-supervised nonnegative matrix factorization
ICA	Independent component analysis
i.i.d.	Independent and identically distributed
ILRMA	Independent low-rank matrix analysis
IP	Iterative projection
IS divergence	Itakura–Saito divergence
ISNMF	Nonnegative matrix factorization based on Itakura–Saito divergence
IVA	Independent vector analysis
KAM	Kernel additive model
KL divergence	Generalized Kullback–Leibler divergence

KLNMF	Nonnegative matrix factorization based on generalized Kullback–Leibler divergence
MFCC	Mel-frequency cepstrum coefficients
MIDI	Musical instrument digital interface
ML	Maximum likelihood
MNMF	Multichannel nonnegative matrix factorization
MSM	Multichannel sparse modeling
MU	Multiplicative update
MWF	Multichannel Wiener filtering
NICA	Nonnegative independent component analysis
NMF	Nonnegative matrix factorization
NNDSVD	Nonnegative double singular value decomposition
PCA	Principal component analysis
PLCA	Probabilistic latent component analysis
REPET	Repeating pattern extraction technique
SDR	Signal-to-distortion ratio
SSNMF	Semi-supervised nonnegative matrix factorization
STFT	Short-time Fourier transform
SVD	Singular value decomposition
TDOA	Time difference of arrival
TFM	Time-frequency masking

Contents

List of Figures xvii

List of Tables xxv

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Prior Work . . . 2
- 1.3 Contributions . . . 4
- 1.4 Outline . . . 4

2 Preliminaries 7
- 2.1 Introduction . . . 7
- 2.2 Mathematical Formulation . . . 8
  - 2.2.1 Multichannel Mixing and Demixing Systems . . . 8
  - 2.2.2 Single-Channel Mixing System . . . 12
- 2.3 Existing Conventional Techniques and Their Categorization . . . 12
- 2.4 Motivations for Developing New Algorithms . . . 17
- 2.5 Basic Principle of NMF . . . 22
- 2.6 Summary . . . 26

3 Determined and Overdetermined Blind Source Separation Based on Independent Low-Rank Matrix Analysis 27
- 3.1 Introduction . . . 27
- 3.2 Basic Principles of ICA, FDICA, IVA, and ISNMF . . . 28

3.2.1 ICA and FDICA . . . 28
3.2.2 IVA . . . 31
3.2.3 ISNMF . . . 36
3.2.4 Time-Varying Gaussian IVA . . . 40
3.3 Independent Low-Rank Matrix Analysis . . . 41
3.3.1 Motivation and Strategy . . . 41
3.3.2 Derivation of Cost Function . . . 42
3.3.3 Update Rules . . . 44
3.3.4 Summary of Algorithm . . . 48
3.4 Relationship between IVA, MNMF, and ILRMA . . . 51
3.4.1 Generative Model in MNMF and Spatial Covariance . . . 51
3.4.2 Existing MNMF Models . . . 52
3.4.3 Equivalence between ILRMA and MNMF with Rank-1 Spatial Model . . . 53
3.5 Experimental Analysis of ILRMA using Artificial Observation . . . 56
3.5.1 Difference between Assumption in Source Model . . . 56
3.5.2 Difference between Assumption in Spatial Model . . . 57
3.5.3 Experimental Validation . . . 58
3.6 Comparison of Speech and Music Separation Performance . . . 64
3.6.1 Datasets . . . 64
3.6.2 Experimental Analysis of Optimal Number of Bases . . . 65
3.6.3 Comparison of Separation Performance . . . 68
3.6.4 Experiments on Three-Source Case with Music Signals . . . 78
3.6.5 Experimental Analysis of Optimal Window Length . . . 80
3.7 Extension of ILRMA for Overdetermined and Reverberant Recording 89
3.7.1 PCA for Overdetermined BSS . . . 89
3.7.2 Relaxation of Rank-1 Spatial Model in ILRMA . . . 90
3.7.3 Clustering with Spectral Correlations . . . 91
3.7.4 Auto-Clustering with Basis-Shared ILRMA . . . 92
3.7.5 Experiments and Results . . . 94
3.8 Summary . . . 99

4 Single-Channel Semi-Supervised Source Separation Based on Discriminative Nonnegative Matrix Factorization 101
4.1 Introduction . . . 101
4.2 Existing NMF-Based Single-Channel Source Separation . . . 102
4.3 Conventional Supervised NMF and Discriminative Training of Supervised Bases . . . 103
4.3.1 Conventional Supervised NMF . . . 103
4.3.2 Drawback in Supervised NMF and Motivation for Discriminative Basis Training . . . 107
4.3.3 Algorithm of Discriminative Basis Training for FSNMF . . . 108
4.4 New Algorithm for Discriminative Basis Training . . . 109
4.4.1 Strategy . . . 109
4.4.2 Discriminative and Reconstructive Basis Training in SSNMF 111
4.5 Experiments . . . 112
4.5.1 Simple Experiment Using Piano and Flute Tones . . . 112
4.5.2 Music Source Separation . . . 112
4.6 Summary . . . 116

5 Effective Initialization for Nonnegative Matrix Factorization Based on Statistical Independence 119
5.1 Introduction . . . 119
5.2 Conventional NMF Initializations . . . 120
5.2.1 Initialization with Random Values . . . 121
5.2.2 Initialization without Random Values . . . 121
5.3 Efficient NMF Initialization Based on ICA . . . 122
5.3.1 Motivation and Strategy . . . 122
5.3.2 Combination of PCA and ICA . . . 124
5.3.3 Proposed Initialization using NICA . . . 125
5.3.4 Proposed Initialization using ICA and Differential of Data Matrix . . . 126
5.3.5 Nonnegativization . . . 127
5.4 Experimental Comparisons . . . 128
5.4.1 Performance as Initial Value for NMF . . . 128

5.4.2 Full-Supervised Audio Source Separation 133
5.5 Application to ILRMA and Discriminative SSNMF 136
5.5.1 BSS Based on ILRMA with Various Initialization for NMF 136
5.5.2 Discriminative SSNMF with Various Initialization for NMF 137
5.6 Summary 138

6 Conclusion **143**
6.1 Summary of Dissertation 143
6.2 Future Works 145

Bibliography **147**

A Derivation of Shape Parameter for Artificial Random Spectrogram with Constant Kurtosis **177**

List of Figures

2.1 Mixing system when $N = 3$ and $M = 2$, where only direct paths are depicted as arrows. 9
2.2 Discreteness in music signals, where score is beginning of Prelude (op. 28, no. 7) by Frédéric Chopin. Typical music consists of limited number of discrete parts. 18
2.3 Example of power spectrogram of drums sound obtained from "another dreamer-the ones we love" in SiSEC2011 dataset, where grayscale indicates spectral power and white is stronger than black. 19
2.4 Example of power spectrogram of guitar sound obtained from "another dreamer-the ones we love" in SiSEC2011 dataset, where grayscale indicates spectral power and white is stronger than black. 19
2.5 Example of power spectrogram of vocals sound obtained from "another dreamer-the ones we love" in SiSEC2011 dataset, where grayscale indicates spectral power and white is stronger than black. 20
2.6 Example of power spectrogram of male speech sound obtained from "dev1_male3_src" in SiSEC2011 dataset, where grayscale indicates spectral power and white is stronger than black. 20
2.7 Example of cumulative singular values of music and speech spectrograms, where all signals are truncated to be the same signals length. 21
2.8 Decomposition model of simple NMF, where $K = 2$. Basis matrix involves representative spectral patterns, and activation matrix represents time-varying gains for each basis. 24

3.1 Permutation problem in FDICA and its solver, where $N = M = 2$. 32
3.2 Mixing and demixing model in IVA, where $N = M = 2$. 33
3.3 Spherically symmetric multivariate Laplace distribution, where $\bar{s}_{ij,n}$ can be considered as either real or imaginary part of $s_{ij,n}$ and $I = 2$ (bivariate case). Two frequency components $s_{1j,n}$ and $s_{2j,n}$ are uncorrelated but have mutual dependences, which is called higher-order correlation. 34
3.4 Principles of source estimation in (a) FDICA and (b) Laplace IVA. Separation filter (demixing matrix) is optimized so that estimated signals obey non-Gaussian source model. Whereas FDICA assumes non-Gaussian source distribution $p(s_{ij,n})$ for each frequency component, IVA assumes non-Gaussian multivariate source distribution $p(\boldsymbol{s}_{j,n})$ that has spherically symmetric property. 36
3.5 Circularly symmetric complex Gaussian distribution. Probability does not depend on phase $\arg(q_{ijl})$ but only depends on amplitude $|q_{ijl}|$ or power $|q_{ijl}|^2$ because of circularly symmetric property. . . 37
3.6 Comparison of source models (variance structures) in (a) IVA and (b) ISNMF, where grayscale in each time-frequency slot indicates scale of variance. IVA has uniform variance over frequency bins, and all the frequency bins have the same activations (time-varying gains), whereas ISNMF employs limited number of bases to capture low-rank structure resulting in more flexible source model. 41
3.7 Relationship between IVA, ILRMA, and MNMF from viewpoint of flexibility of spatial and source models. 56
3.8 Artificial source that consists of R bases. 58
3.9 Artificial DOA with Gaussian distributions. 60
3.10 Separation results of (a) first source and (b) second source for various numbers of bases. 61
3.11 Separation results of (a) first source and (b) second source for various angles. 63
3.12 Separation results of (a) first source and (b) second source for various variances. 63

3.13 Recording conditions of impulse responses (a) E2A and (b) JR2 for two-source case. 65
3.14 Average SDR improvements for female speech (dev1) with 1 m microphone spacing and 130 ms reverberation time: (a) first speaker and (b) second speaker. 67
3.15 Average SDR improvements for song ID4 with impulse response E2A: (a) guitar and (b) synth. 67
3.16 Cumulative singular values of each source spectrogram in dev1 female speech and song ID4 music, where all sources are truncated to be the same signal length. 68
3.17 Average SDR improvements for female speech (dev1) with 1 m microphone spacing, where reverberation time is (a) 130 ms and (b) 250 ms. 71
3.18 Average SDR improvements for male speech (dev1) with 1 m microphone spacing, where reverberation time is (a) 130 ms and (b) 250 ms. 72
3.19 Average SDR improvements for music signal song ID3 with impulse response (a) E2A and (b) JR2. 73
3.20 Average SDR improvements for music signal song ID4 with impulse response (a) E2A and (b) JR2. 74
3.21 Convergence of z_{1k} from $k = 1$ to $k = K$ in music signal case. . . . 75
3.22 SDR convergence for music signal song ID4 with impulse response E2A: (a) guitar and (b) synth. 77
3.23 Results of subjective scores obtained by Thurstone pairwise comparison method, where 48 pairs of separated speech and 48 pairs of separated music signals are presented in random order to 14 examinees, who selected which signal they preferred from the viewpoint of total quality of separated sound. Scores show relative tendency of selection. 78
3.24 Probability of selection regarding difference between two subjective scores. 79
3.25 Recording conditions of impulse responses (a) E2A and (b) JR2 for three-source case. 80

3.26 Average SDR improvements for music signal song ID4 in three-source case with impulse response (a) E2A and (b) JR2. 81

3.27 Averaged SDR improvements averaged for music signal song ID1 with impulse response JR2 (reverberation time is 470 ms): (a) $L = 5$ and $K = 10$, (b) $L = 10$ and $K = 20$, (c) $L = 20$ and $K = 40$, (d) $L = 30$ and $K = 60$, (e) $L = 40$ and $K = 80$, and (f) $L = 50$ and $K = 100$. Scores are averaged for two sources and 10 trials with different pseudorandom seeds. 84

3.28 Averaged SDR improvements averaged for music signal song ID2 with impulse response JR2 (reverberation time is 470 ms): (a) $L = 5$ and $K = 10$, (b) $L = 10$ and $K = 20$, (c) $L = 20$ and $K = 40$, (d) $L = 30$ and $K = 60$, (e) $L = 40$ and $K = 80$, and (f) $L = 50$ and $K = 100$. Scores are averaged for two sources and 10 trials with different pseudorandom seeds. 85

3.29 Averaged SDR improvements averaged for music signal song ID3 with impulse response JR2 (reverberation time is 470 ms): (a) $L = 5$ and $K = 10$, (b) $L = 10$ and $K = 20$, (c) $L = 20$ and $K = 40$, (d) $L = 30$ and $K = 60$, (e) $L = 40$ and $K = 80$, and (f) $L = 50$ and $K = 100$. Scores are averaged for two sources and 10 trials with different pseudorandom seeds. 86

3.30 Averaged SDR improvements averaged for music signal song ID4 with impulse response JR2 (reverberation time is 470 ms): (a) $L = 5$ and $K = 10$, (b) $L = 10$ and $K = 20$, (c) $L = 20$ and $K = 40$, (d) $L = 30$ and $K = 60$, (e) $L = 40$ and $K = 80$, and (f) $L = 50$ and $K = 100$. Scores are averaged for two sources and 10 trials with different pseudorandom seeds. 87

3.31 Number of bases in power spectrogram of each source when its cumulative singular value reaches 80% or 90%: (a) song ID1, guitar (source 1) and vocals (source 2), (b) song ID2, guitar (source 1) and vocals (source 2), (c) song ID3, violins synth. (source 1) and vocals (source 2), and (d) song ID4, guitar (source 1) and synth. (source 2). 88

3.32 Mixing system of each spectrogram slot when $N = M = 2$; (a) has a linear time-invariant mixing system and there is no reverberation; (b) has some leaked components from the previous frame because of reverberation. 90
3.33 Algorithms of (a) conventional and (b) proposed methods ($N = 2$, $M = 4$, and $Q = 2$), where subscripts for i and j are omitted. 92
3.34 Hierarchical clustering using correlation cor ($N = 2$, $M = 4$, and $Q = 2$), where all sets must have the same number of signals and subscripts for i and j are omitted.. 93
3.35 Recording condition of impulse response used in experiment of reverberant signals. 94
3.36 Average SDR improvements for song ID1 used in experiment of reverberant signals. 97
3.37 Average SDR improvements for song ID2 used in experiment of reverberant signals. 97
3.38 Average SDR improvements for song ID3 used in experiment of reverberant signals. 98
3.39 Average SDR improvements for song ID4 used in experiment of reverberant signals. 98

4.1 Training and separation stages in (a) FSNMF and (b) SSNMF. . . . 104
4.2 Difference between (a) conventional and (b) proposed algorithms of separation, where black components correspond to target source and gray components correspond to interfering source. Proposed method utilizes discriminative bases ($\boldsymbol{F}'$) that has unique component of target source for separation, and target source is synthesized using reconstructive bases ($\boldsymbol{F}$) that has complete spectral components. 110
4.3 Spectral bases obtained from simple NMF: (a) C5 piano tone and (b) C6 flute tone. 113
4.4 Spectral bases obtained from (4.16), where (a) is discriminative basis $\boldsymbol{F}'$ of piano tone and (b) and (c) are the other bases in $\boldsymbol{T}$. . . 114

4.5 Average SDR improvement for each mixture and each number of iterations in (4.16). 117

5.1 Example of SDR improvements of music source separation, where FSNMF initialized by random values with various pseudorandom seeds is performed. 121

5.2 Geometry of (a) optimal, (b), orthogonal, and (c) close bases, where black dots indicate observed data points in positive orthant, gray area indicates cone defined by data points, broken lines indicate edges of cone, $\boldsymbol{f}_k$ denotes kth NMF basis, $\Phi = K = 2$, and $\Psi = 10$. 123

5.3 Assumption of proposed method, where nonnegative activations are assumed to be independent of each other. 125

5.4 Convergence of cost function in (a) NICA and (b) ICA. 129

5.5 Convergences of cost function in EUNMF, where NICA1–NICA3 are depicted in (a), ICA1–ICA3 are depicted in (b), and conventional methods are depicted in both. 130

5.6 Convergences of cost function in KLNMF, where NICA1–NICA3 are depicted in (a), ICA1–ICA3 are depicted in (b), and conventional methods are depicted in both. 131

5.7 Convergences of cost function in ISNMF, where NICA1–NICA3 are depicted in (a), ICA1–ICA3 are depicted in (b), and conventional methods are depicted in both. 132

5.8 SDR improvement of supervised NMF for (a) vocals and (b) other. 135

5.9 Process flow of BSS based on ILRMA with NMF initialization method. 136

5.10 Average SDR improvement of ID1 for each number of iterations in (4.16) with various NMF initializations. 139

5.11 Average SDR improvement of ID2 for each number of iterations in (4.16) with various NMF initializations. 139

5.12 Average SDR improvement of ID3 for each number of iterations in (4.16) with various NMF initializations. 140

5.13 Average SDR improvement of ID4 for each number of iterations in (4.16) with various NMF initializations. 140
5.14 Average SDR improvement of ID5 for each number of iterations in (4.16) with various NMF initializations. 141
5.15 Average SDR improvement of ID6 for each number of iterations in (4.16) with various NMF initializations. 141

List of Tables

2.1 Categorization of typical existing techniques for audio source separation 13

3.1 Models of mixing system, spatial covariance, power spectrogram, and their optimization in each method 53
3.2 Estimated values of shape parameter κ so that kurtosis of $\boldsymbol{FG}$ is adjusted to 50 for each $\mathcal{R}$ 59
3.3 Music sources for two-source case 65
3.4 Experimental conditions 66
3.5 Experimental conditions used in Ozerov's MNMF 69
3.6 Averaged SDR improvements over various speech signals and sources with same recording conditions in two-source case 75
3.7 Averaged SDR improvements over various music signals and sources with same impulse response in two-source case 75
3.8 Music sources for three-source case 79
3.9 Averaged SDR improvements over various music signals and sources with same impulse response in three-source case 80
3.10 Computational times (s) for separation of song ID1 with impulse response E2A in three-source case 82
3.11 Experimental conditions used in analysis of optimal window length 82
3.12 Music sources used in experiment of reverberant signals 95
3.13 Characteristics of each method used in experiment of reverberant signals 95
3.14 Experimental conditions used in experiment of reverberant signals 96

3.15 Computational times for separation of song ID3 used in experiment of reverberant signals (s) 99

4.1 Mixture $\mathbf{\Lambda}$ with target and non-target sources 115
4.2 Sample sounds of non-target source $\boldsymbol{N}^{(\mathrm{train})}$ for preparing $\boldsymbol{S}_{\mathrm{mix}}^{(\mathrm{train})}$. 115

5.1 Examples of computational time in each process (s) 133
5.2 Averaged SDR improvements over various speech signals and sources with same recording conditions, where these scores are obtained by ILRMA w/o partitioning function with various NMF initializations 137
5.3 Averaged SDR improvements over various music signals and sources with same impulse response, where these scores are obtained by ILRMA w/o partitioning function with various NMF initializations 138

1 Introduction

1.1 Background

Audio source separation is a technique for separating specific source signals from a mixture signal and has been intensively studied for several decades. This technique mainly focuses on mixed signals of speech, which can be used for many applications including speech enhancement, automatic speech recognition, reduction of undesired background noise, and hearing aid systems. Many applications assume the use of microphone array, which consists of several synchronized microphones, for recording sound sources in a multichannel format. In particular, when the number of microphones (channels) is equal to or grater than the number of sources, which is called determined or overdetermined situation, blind source separation (BSS) technique is often applied to solve the audio source separation problem. The benefit of using BSS is accessibility for many systems because BSS techniques do not require any information about the recording environment, mixing system, or source locations. Audio source

separation for underdetermined or single-channel signals is also a potential feature because the variety of its applications is wider than the source separation techniques using microphone array. Since there is no satisfactory information in the observations, this problem is tougher than the determined problem, and some assumptions or training data for sources in a mixture are required to solve the separation effectively.

The audio source separation for music signals has also attracted considerable interest in recent yeas. This is also used for many applications, for example, remixing of existing music, automatic music transcription, music search and recommendation systems, sound field reproduction of a live concert, and education for musical instruments. When the music signal is recorded by the microphone array (in a determined or an overdetermined situation), BSS techniques can be applied similar to the speech source separation. However, almost all existing music signals are mixed down and provided in a stereo format even though it includes more than two sources. For such signals, the underdetermined or single-channel source separation techniques must be applied to achieve the source separation. As the advantage in music source separation, some prior information of the instrumental sources are often available, for example, a music score, solo-played instrumental signals, and synthetic or sample-recorded sound dataset of specific musical instruments. For this reason, supervised approach of music source separation has also been an active area of research.

1.2 Prior Work

The source separation problem includes several situations depending on its assumptions and conditions, e.g., the number of recoding channels, recording environment, presence of prior information and training data, characteristics of mixed sources (stability and temporal or harmonic structure), and availability of other sensing data (multimodality). BSS for the determined observation is one of the most basic theory in these problems. In particular, *independent component analysis (ICA)* [1] has been well studied since the mid 1990s not only in the filed of acoustic signal processing but also in fields of brain science, radio

engineering, financial engineering, and image signal processing. Since a mixing system of acoustic signals becomes a convolutive mixture due to effect of room reverberation, and it is a more difficult problem than the instantaneous mixing system, ICA has been developed mainly in the field of acoustic signal processing. In the late 1990s, *frequency-domain ICA (FDICA)* [2] was established to deal with such convolutive mixture using Fourier transform. By the advent of *independent vector analysis (IVA)* [3, 4] in 2006, which is an extension of FDICA, a high-quality blind speech separation was achieved.

On the other hand, a new theory of matrix decomposition called *nonnegative matrix factorization (NMF)* [5] was invented in 1999. NMF can extract some meaningful features in an observed data matrix as non-orthogonal basis vectors and is applied to a spectrogram of an acoustic signal, which enables one to extract specific spectral patterns of sources. On the basis of this theory, many techniques for solving the underdetermined and single-channel audio source separation have been proposed, and they have substantially been progressed during the 2000s and 2010s. In particular, *multichannel NMF (MNMF)* [6], which deals with the multichannel signal and simultaneously models the spectral patterns and spatial information (differences between channels) of the recording environment, results in a major contribution to the underdetermined source separation problem. Also, a supervised approach based on NMF [7] was introduced for the single-channel source separation, where training data for the sources in mixture signal are directly exploited for preparing spectral dictionaries of each source. In the very recent past, a method of discriminative training for supervised NMF has actively been addressed [8, 9, 10] to improve separation performance of supervised NMF. As another issue, the effective initialization method for NMF has been investigated since NMF appeared, particularly in the field of machine learning. Since an optimization algorithm in NMF requires initial values for the variables and the result of decomposition strongly depends on their initialization, the effective initialization is one of the attractive research topics.

1.3 Contributions

The objective of this dissertation is to develop an effective optimization algorithm for the music source separation and to achieve satisfactory separation performance. On the basis of the nature in music signals, such as low-rank and discrete structure of the music spectrogram, NMF is suitably exploited for modeling these structures throughout this dissertation. The following two major problems in music source separation are mainly addressed here:

- Determined and overdetermined BSS
- Single-channel semi-supervised source separation

For the former issue, an effective unified method of NMF and IVA is proposed as an extension of traditional ICA algorithms. In addition, intriguing relationships between IVA, MNMF, and the proposed method are theoretically revealed. For the latter research topic, a new optimization algorithm for discriminative training of supervised bases is developed, which can be applied even in semi-supervised situations. The algorithm approximately solves a conventional bilevel optimization problem for obtaining discriminative bases. Furthermore, an effective initialization scheme based on statistical independence of NMF components is considered and applied to the two proposed methods described above.

1.4 Outline

Chapter 2 presents some basic preliminaries of audio source separation, which include mathematical formulation, overview of conventional techniques, motivations for developing new algorithms, and principle of NMF. In Chap. 3, I propose a new effective algorithm for determined BSS task and discuss the relationship between conventional and the proposed methods. Also, the extended algorithm for overdetermined and reverberant signals is proposed. The efficacies of these methods are validated via experiments. Chapter 4 deals with a single-channel semi-supervised source separation problem. After explaining the

conventional approaches for training discriminative features in full-supervised situation, a new discriminative training algorithm in semi-supervised situation is developed. In Chap. 5 an initialization problem for general NMF optimization problem is discussed, and an efficient initialization scheme based on statistical independence is proposed. Also, the efficacy of the proposed initialization for NMF-based source separation task is experimentally confirmed. Finally, Chap. 6 concludes the whole contents and contributions in this dissertation.

2

Preliminaries

2.1 Introduction

In this chapter, I provide some preliminaries, which are necessary for later discussion. After giving a mathematical formulation of general source separation, I explain an overview of existing conventional techniques and categorize them in terms of prior conditions (assumptions) for solving the source separation problems. Next, motivations for developing new algorithms of music source separation are clarified, then I introduce a key ingredient of this dissertation, which is a matrix decomposition algorithm called NMF [11, 5, 12, 13, 14]. Finally, I summarize the contents in this chapter.

2.2 Mathematical Formulation

2.2.1 Multichannel Mixing and Demixing Systems

We consider a mixing system of N sound sources and M channels (microphones) as the following expression [15]:

$$\tilde{x}_m(\tau) = \sum_{n=1}^{N} \tilde{c}_{nm}(\tau) \tag{2.1}$$

$$= \sum_{n=1}^{N} \sum_{\tau'=0}^{L_{\text{filter}}-1} \tilde{a}_{nm}(\tau)\tilde{s}_n(\tau - \tau'), \tag{2.2}$$

where $\tau = 1, 2, \cdots, \tau_{\text{end}}$ is the integral index of discrete-time, $n = 1, 2, \cdots, N$ and $m = 1, 2, \cdots, M$ are the integral indices of sources and microphones, respectively, $\tilde{x}_m(\tau)$ is the observed (recorded) time-domain signal of the mth microphone, $\tilde{c}_{nm}(\tau)$ is the observed nth source signal obtained by the mth microphone, $\tilde{a}_{nm}(\tau)$ is a filter coefficient of impulse response that models the acoustic path from the nth source to the mth microphone, L_{filter} is the filter length of $\tilde{a}_{nm}$, and $\tilde{s}_n(\tau)$ is the time-domain signal of the nth original source. The mixing model can also be expressed by vector form as follows:

$$\tilde{\boldsymbol{x}}(\tau) = \sum_{n=1}^{N} \tilde{\boldsymbol{c}}_n(\tau) \tag{2.3}$$

$$= \sum_{n=1}^{N} \sum_{\tau'=0}^{L_{\text{filter}}-1} \tilde{\boldsymbol{a}}_n(\tau)\tilde{s}_n(\tau - \tau'), \tag{2.4}$$

where $\tilde{\boldsymbol{x}}(\tau) = (\tilde{x}_1(\tau) \ \cdots \ \tilde{x}_M(\tau))^{\mathrm{T}}$ is the observed multichannel vector, $\tilde{\boldsymbol{c}}_n(\tau) = (\tilde{c}_{n1}(\tau) \ \cdots \ \tilde{c}_{nM}(\tau))^{\mathrm{T}}$ is the observed multichannel vector of nth source, which is often called spatial source image [16], $\tilde{\boldsymbol{a}}_n(\tau) = (\tilde{a}_{n1}(\tau) \ \cdots \ \tilde{a}_{nM}(\tau))^{\mathrm{T}}$ includes the filter coefficients of acoustic paths from nth source to all the microphones, and $\cdot^{\mathrm{T}}$ denotes the vector or matrix transpose. The equations (2.2) and (2.4) show the convolutive mixture of original N sources, which simulates the reverberant mixture in a recording environment, and the filter length L_{filter} corresponds to the length of the reverberation time. In an anechoic case ($L_{\text{filter}} = 1$), the mixing

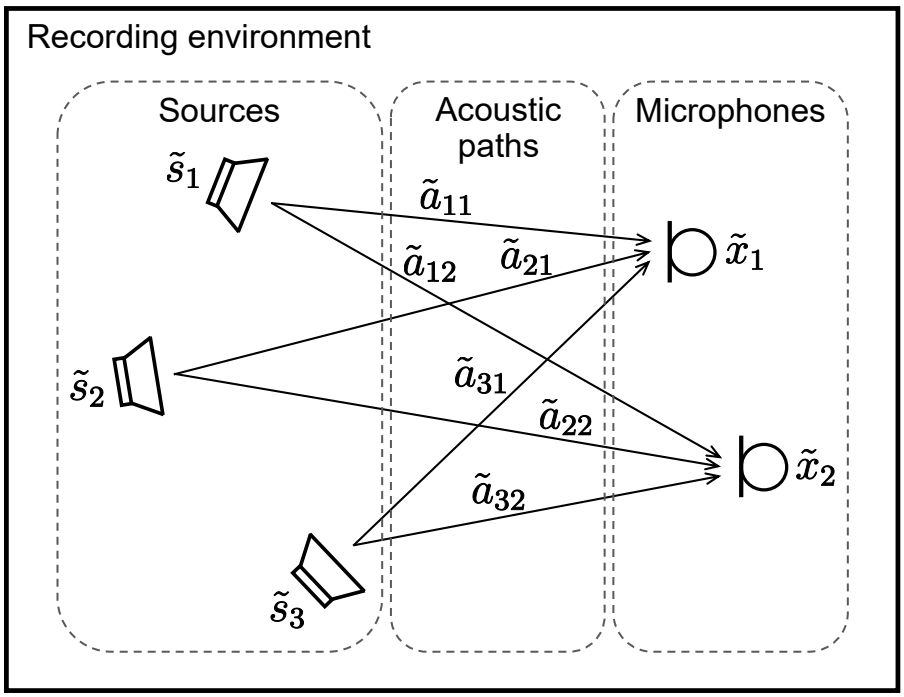

Figure 2.1: Mixing system when $N = 3$ and $M = 2$, where only direct paths are depicted as arrows.

system represented above becomes an instantaneous mixture in a time domain. Figure 2.1 shows the mixing system, where only the direct paths between sources and microphones are depicted as the arrows. Note that since $\tilde{a}_{nm}$ is a filter with length L_{filter}, there are many other paths with delays due to the reflection in the recording environment. The objective of source separation is not the estimation of original dry source $\tilde{s}_n(\tau)$ but the estimation of the separated observation $\tilde{c}_{nm}(\tau)$ or $\tilde{\boldsymbol{c}}_n(\tau)$. The estimation problem of $\tilde{s}_n(\tau)$ from $\tilde{c}_{nm}(\tau)$ or $\tilde{\boldsymbol{c}}_n(\tau)$ is generally called dereverberation [17, 18, 19], which is out of the scope and is not treated in this dissertation.

The mixing system can be transformed via short-time Fourier transform (STFT) [20, 21] with an analysis window. In particular, when the length of the analysis window is sufficiently longer than L_{filter}, the convolutive mixture model can be transformed into the instantaneous mixture model in the time-frequency

domain as

$$x_{ij,m} = \sum_{n=1}^{N} c_{ij,nm} = \sum_{n=1}^{N} a_{i,nm} s_{ij,n}, \tag{2.5}$$

or in the vector form

$$\boldsymbol{x}_{ij} = \sum_{n=1}^{N} \boldsymbol{c}_{ij,n} = \sum_{n=1}^{N} \boldsymbol{a}_{i,n} s_{ij,n}, \tag{2.6}$$

where $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$ are integral the indices of frequency bins and time frames obtained via STFT, respectively, and $x_{ij,m}$, $c_{ij,nm}$, $a_{i,nm}$, and $s_{ij,n}$ are the complex-valued STFT coefficients of the time domain signals $\tilde{x}_m$, $\tilde{c}_{nm}$, $\tilde{a}_{nm}$, and $\tilde{s}_{ij,n}$, respectively. The vectors in (2.6) are defined as $\boldsymbol{x}_{ij} = (x_{ij,1} \ \cdots \ x_{ij,M})^{\mathrm{T}}$, $\boldsymbol{c}_{ij,n} = (c_{ij,n1} \ \cdots \ c_{ij,nM})^{\mathrm{T}}$, $\boldsymbol{a}_{i,n} = (a_{i,n1} \ \cdots \ a_{i,nM})^{\mathrm{T}}$, and $\boldsymbol{s}_{ij} = (s_{ij,1} \ \cdots \ s_{ij,N})^{\mathrm{T}}$, where note that $\boldsymbol{x}_{ij}$, $\boldsymbol{c}_{ij,n}$, and $\boldsymbol{a}_{ij,n}$ are the multichannel $(M \times 1)$ vectors and $\boldsymbol{s}_{ij}$ is the multisource $(N \times 1)$ vector. The vector $\boldsymbol{a}_{i,n}$ is called array manifold vector [22] or steering vector [23, 24], which models the acoustic paths for nth source in frequency domain. We here consider that the steering vector $\boldsymbol{a}_{i,n}$ is time-invariant, namely, all the spatial locations of sources and microphones do not change along the time frame. The time-invariant instantaneous mixture in the time-frequency domain leads to the following simple mixing representation:

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \tag{2.7}$$

where

$$\boldsymbol{A}_i = \begin{pmatrix} \boldsymbol{a}_{i,1} & \cdots & \boldsymbol{a}_{i,N} \end{pmatrix} = \begin{pmatrix} a_{i,11} & \cdots & a_{i,N1} \\ \vdots & \ddots & \vdots \\ a_{i,1M} & \cdots & a_{i,NM} \end{pmatrix} \tag{2.8}$$

is called *mixing matrix* whose size is $M \times N$. If the mixing matrix is a full-rank square matrix $(M = N)$, we can define an inverse matrix of $\boldsymbol{A}_i$ that separates the

sources in $\boldsymbol{x}_{ij}$ as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}, \tag{2.9}$$

$$y_{ij,n} = \boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}, \tag{2.10}$$

where $\boldsymbol{y}_{ij} = (y_{ij,1} \ \cdots \ y_{ij,N})^{\mathrm{T}}$ is the estimated (separated) multisource vector,

$$\boldsymbol{W}_i = \begin{pmatrix} \boldsymbol{w}_{i,1}^{\mathrm{H}} \\ \vdots \\ \boldsymbol{w}_{i,N}^{\mathrm{H}} \end{pmatrix} = \begin{pmatrix} w_{i,11}^{*} & \cdots & w_{i,1M}^{*} \\ \vdots & \ddots & \vdots \\ w_{i,N1}^{*} & \cdots & w_{i,NM}^{*} \end{pmatrix} \tag{2.11}$$

is called *demixing matrix*, $\boldsymbol{w}_{i,n} = (w_{i,n1} \ \cdots \ w_{i,nM})^{\mathrm{T}}$ is called *demixing filter* for nth source, $\cdot^{\mathrm{H}}$ denotes the Hermitian transpose, and $\cdot^{*}$ denotes the conjugate of complex value. The estimated time-domain signal $\tilde{y}_n(\tau)$ can be calculated by applying inverse STFT with overlap-save [25] and the appropriate synthesis window [26, 27] to $y_{ij,n}$. The estimation problem of the separated signals $\tilde{y}_n(\tau)$ without knowing any information about the mixing system $\tilde{\boldsymbol{a}}_n(\tau)$ is often called BSS. In this problem, calibration of the microphones is not required because we cannot distinguish the frequency responses of the microphones and the effects caused by the filter coefficients $\tilde{\boldsymbol{a}}_n(\tau)$, namely, the difference of characteristics between the microphones is absorbed by $\tilde{\boldsymbol{a}}_n(\tau)$.

Hereafter, I denote the complex-valued spectrograms ($I \times J$ time-frequency matrix) of, nth source signal, mth observed signal, and nth separated signal as $\boldsymbol{S}_n \in \mathbb{C}^{I \times J}$, $\boldsymbol{X}_m \in \mathbb{C}^{I \times J}$, and $\boldsymbol{Y}_n \in \mathbb{C}^{I \times J}$, respectively. Also, the third-order tensor of source, observed, and separated signals are denoted using sans-serif upright font as $\mathsf{S} \in \mathbb{C}^{I \times J \times N}$, $\mathsf{X} \in \mathbb{C}^{I \times J \times M}$, and $\mathsf{Y} \in \mathbb{C}^{I \times J \times N}$, respectively. Moreover, the third-order tensor with a subscript denotes the sliced matrix or the fiber vector [28] in the original tensor. For example, $\mathsf{X}_{i::}$, $\mathsf{X}_{:j:}$, and $\mathsf{X}_{::m}$ denote the $J \times M$, $I \times M$, and $I \times J$ sliced matrices in X, respectively. Also, $\mathsf{X}_{ij:}$, $\mathsf{X}_{i:m}$, and $\mathsf{X}_{:jm}$ denote the $M \times 1$, $J \times 1$, and $I \times 1$ fiber (column) vectors in X, respectively.

2.2.2 Single-Channel Mixing System

For the case of $M = 1$, the mixing system can be defined as

$$\tilde{x}(\tau) = \sum_{n=1}^{N} \tilde{c}_n(\tau) \tag{2.12}$$

$$= \sum_{n=1}^{N} \sum_{\tau'=0}^{L_{\text{filter}}-1} \tilde{a}_n(\tau)\tilde{s}_n(\tau - \tau'). \tag{2.13}$$

In the time-frequency domain, (2.12) and (2.13) are transformed as

$$x_{ij} = \sum_{n=1}^{N} c_{ij,n} \tag{2.14}$$

$$= \sum_{n=1}^{N} a_{i,n} s_{ij,n}. \tag{2.15}$$

The single-channel source separation is more difficult problem than that for multichannel signals because differences of amplitudes and phases between channels cannot be utilized. Thus, typical single-channel source separation techniques employ some strong constraints or a powerful a priori knowledge for achieving the objective.

2.3 Existing Conventional Techniques and Their Categorization

Audio source separation techniques can roughly be divided in terms of two aspects; determinacy of mixing system and presence of external supervised information. The former issue is related to the numbers of sources and channels (N and M). Since the source separation is an inverse problem, its difficulty directly depends on these conditions. When the number of sources is equal or less than the number of channels ($N \leq M$), the source separation becomes determined or overdetermined problem. On the other hand, when the number of sources is grater than the number of channels ($N > M$), it is called underdetermined

Table 2.1: Categorization of typical existing techniques for audio source separation

Situations	Blind	Supervised			
		Spectral supervision		Spatial supervision	
		Sound examples	Source activities	Source locations	Steering vectors
Determined or overdetermined	FDICA IVA	Multichannel DNN	User-guided IVA	Fixed BF Adaptive BF	Robust adaptive BF
		Spectral supervision		Spatial supervision	
		Sound examples	Source activities	Source locations	Steering vectors
Underdetermined	Sparse coding TFM TDOA clustering MNMF	Multichannel DNN Hybrid method	User-guided MNMF	TFM	Dictionary-based MSM
		Spectral supervision		Spatial supervision	
		Sound examples	Source activities	Source locations	Steering vectors
Single-channel	TFM REPET KAM	Supervised NMF DAE	Informed NMF	–	–

problem, which is tougher than the previous situation. The latter issue is related to a presence of a priori knowledge for sources. For example, source locations (spatial positions) can be used for a multichannel source separation techniques. For the music separation, scores are powerful prior information for estimating source activities. Also, some instrumental sequences may be available in advance to train the specific source spectra. Table 2.1 summarizes a categorization of typical existing techniques for audio source separation techniques, where they are categorized from the viewpoints of the problem determinacy and the presence of external supervised information. I explain the details of these typical techniques below.

In the determined and overdetermined situation, ICA [1, 29, 30, 31, 32, 33, 34, 35] has been the most successful algorithm for the source separation problem. ICA utilizes the assumption of statistical independence between sources and estimates the demixing filters from the mixture observations in a fully blind fashion, which is called BSS. For BSS in audio signals, the sources are convolved by the room reverberation as (2.4). Many ICA-based separation techniques for delayed and convolved sources were proposed [36, 37, 38, 39, 40]. Also, FDICA [2, 41, 42, 34, 43, 35] was developed as another approach for solving the signal deconvolution using STFT. In FDICA, the demixing matrix in the frequency domain, $\boldsymbol{W}_i$, is estimated for the separation. This method is more stable and more efficient compared with ICA deconvolution in the time domain because we can easily treat the convolutive mixing system as the simple instantaneous

mixture (2.7) by applying STFT to the signals.

For supervised source separation in determined and overdetermined situations, both fixed and adaptive beamforming (BF) techniques have been widely used. The source separation based on fixed BF assumes that the locations of all sources and microphones are known, where the number of microphones should be large for the accurate separation. Adaptive BF [44, 45, 46, 24] exploits additional criteria, such as minimum variance and distortionless constraint, to adaptively reduce the background diffuse noise and extract the target sources. It is revealed that the estimation in adaptive BF are essentially equivalent to that in BSS based on FDICA [47] whereas FDICA is a blind (unsupervised) technique. However, even if the locations of sources and microphones are known, BF-based methods fail to accurately separate the sources when the source signals are convolved by room reverberations, which particularly arise in audio recording. This is because the directions of steering vectors spatially spread around the true source direction, and the distortionless constraint in adaptive BF cannot ensure the quality of estimated sources. In some limited cases, the pretrained steering vectors can be used for BFs as "strict" spatial supervision including spatial spreads of each source, where the mismatch between the trained steering vectors and an observation becomes further important problem. Robust adaptive BF [48, 49, 50, 51, 52] was developed to improve the robustness against such mismatch of steering vectors.

On the other hand, in the underdetermined situation, ICA has been used to estimate not the demixing filters but the ICA bases, which is known as an estimation of overcomplete bases [30, 53, 32]. This approach develops to new methods such as a sparse coding [54, 55, 56, 57] and a time-frequency masking (TFM) [58, 59, 60], which are related to the methods in the machine learning or pattern recognition field. These methods are based on the sparseness of signals, which is a strong and practical assumption, and can solve the BSS problem even in the underdetermined situation. However, since the sparseness assumption does not always hold completely, the sound quality of separated signals is markedly degraded owing to the generation of artificial noise. Reticently, TFM is utilized for the estimation of steering vectors [61, 62], and it is reported that this hybrid approach gives better suppression of diffuse noise in the overdetermined

situation. Clustering-based underdetermined source separation is developed for multichannel observations [58, 63, 64, 65, 60, 66], where these techniques rely on time-difference-of-arrival (TDOA) estimations for each source at multiple microphones. However, under severe reverberant conditions, TDOA estimations become unreliable and these clustering-based techniques do not work well.

As another approach, NMF [11, 5, 12, 13, 14] has been introduced for single-channel BSS [67, 68, 69, 70, 71, 72, 73, 74, 75, 76]. NMF is a low-rank approximation of an observed nonnegative matrix under nonnegative constraint, and a small number of meaningful bases can be extracted from the observed matrix. For acoustic signals, an amplitude or a power spectrogram is used as the observed matrix in NMF, and the source separation is achieved by clustering the decomposed bases into each source. In order to utilize spatial information of each source as a criterion of the clustering, NMF is extended to multichannel model [77, 78, 79, 80, 6, 81, 82, 83, 84], which has a potential to solve BSS even in the underdetermined situation. In addition, in recent years, several new approaches based on a repetitive structure of music signals were proposed for single-channel blind situation, which are called repeating pattern extraction technique (REPET) [85, 86, 87, 88, 89, 90] and kernel additive model (KAM) [91, 92, 93, 94].

As supervised approaches for single-channel signals, supervised NMF [95, 7, 96, 97, 98] is the most reliable method, which directly utilizes training sequences of each source for clustering the NMF bases. For the multichannel signals, a hybrid method of binary TFM and supervised NMF was proposed [99], where the sources are first separated based on spatial cues, then supervised NMF is effectively applied. A score-informed or user-guided approach with NMF or MNMF [100, 82, 101, 102, 103] has also been a popular technique for providing better separation exploiting an external information about the source activities as supervision. This supervised approach is also applied to the determined BSS, e.g., user-guided IVA [104]. In addition, multichannel sparse modeling (MSM) with large-scale spatial dictionary [105, 106] is another approach for spatially informed underdetermined source separation. In this method, acoustic paths between many spatial locations and microphones are measured or calculated in advance to prepare a large-scale spatial dictionary. The mixing system is

adaptively estimated using sparse modeling and the dictionary. The mismatch between the dictionary and actual observation is also estimated for the robust separation [106].

In recent years, underdetermined or single-channel source separation based on deep neural network (DNN) or denoising autoencoder (DAE) has been a very active research topic [107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117]. Many literatures investigate the DNN-based source separation for speech signals [107, 108, 109, 110, 112, 113, 114, 116, 117, 118] and music signals [111, 113, 115] so far. In the method [118], the NMF decomposition for source modeling in MNMF is replaced to DNN, resulting in developing a new multichannel DNN-based audio source separation with spectral supervision. For spatial supervision, multichannel features are exploited as an input data of DAE to train the spatial information [119, 120, 114]. However, for the convolutive BSS in time domain, thousands of coefficients in the separation filters must be trained even with 300 ms reverberation and 8-kHz sampling rate. Thus, to train the spatial features using DNN with practical dataset may be almost impossible in a real situation. In addition, DAE-based methods require many pairs of clean and contaminated source signals. This is a crucial problem in music source separation because we cannot prepare sufficient number of such pairs in a practical situation.

The main objective of this dissertation is to advance the audio source separation techniques and develop more practical algorithms that give us better separation performance. In addition, this dissertation mainly focuses on the separation of music signals. For this reason, I will aim to only the following situations:

- Determined and overdetermined BSS
- Single-channel semi-supervised source separation

The first issue, which will be addressed in Chap. 3, is a classical BSS problem, where we do not assume the spatial supervision. As a motivation to treat BSS, in a realistic situation, it is almost impossible to accurately train the whole spatial information including source locations and spatial spreads caused by reverberations. The almost all audio recordings are always the only once,

and completely the same situation (mixing system) is never reproduced. The mismatch-robust approach such as robust adaptive BF is one of the possible solution, but the blind estimation is most preferable if it can solve the problem. For the determined and overdetermined observations, ICA-based approaches have a potential to blindly solve the source separation and have been successful. However, it usually aims to the separation of speech signals, and the music signals have not often been treated so far. This might be because the conventional ICA-based BSS does not provide satisfactory performance for music signals. If we achieve a better music separation in this situation, various applications using microphone array can be realized, e.g., remixing or editing of live-recorded music.

The second issue, which will be addressed in Chap. 4, is also important for the music source separation because almost all music signals are provided as a stereo format including more than two sources, namely, the underdetermined situation. In particular, source separation for single-channel signals is the most basic framework and can easily be applied for the other situations. Moreover, for music signals, the users can easily produce synthetic instrumental sounds using musical instrument digital interface (MIDI) synthesizer or easily record the actual instrumental sounds of their interesting source part. Therefore, the spectral supervised approach can also be applied for music source separation. This dissertation is mainly focused on a semi-supervised method, which utilizes a training sequence for only the target instruments. This approach is more practical for many applications than preparing the training sequences for all the sources, namely, a full-supervised method.

2.4 Motivations for Developing New Algorithms

For determined BSS of speech signals, the conventional ICA-based methods are well investigated and can achieve better separation performance. However, for music signals, the separation accuracy tends to be degraded. The main reason is that typical music signal frequently contains co-occurring sources with many overlapped spectra, which results in harmony with multiple instrumental and vocal sources. In such signals, the statistical independence between the

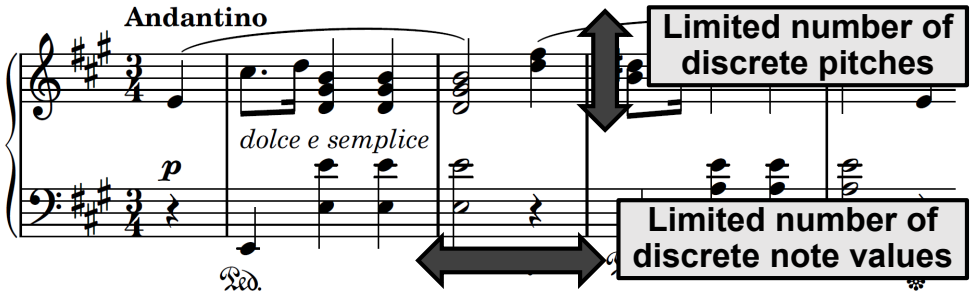

Figure 2.2: Discreteness in music signals, where score is beginning of Prelude (op. 28, no. 7) by Frédéric Chopin. Typical music consists of limited number of discrete parts.

sources is weakened, and ICA-based separation often causes errors of capturing each source. Also, the conventional ICA-based methods utilize only the source distribution $p(\boldsymbol{Y}_n)$ for the source model, and it has been empirically determined (e.g., Laplace distribution as a source distribution of speech signals [121]) without any information about the specific time-frequency structure of each source. Since BSS is the inverse problem, the separation performance directly depends on the accuracy of the source distribution $p(\boldsymbol{Y}_n)$ or the source model. We can expect that if we employ stricter source model in ICA algorithm, the separation performance for music signals would be improved.

In this dissertation, I focus on a property of music signals to effectively model their spectrograms (source models). Typical music signals include many reiteration of similar or the same instrumental timbres, melody patterns, chords, harmonies, and refrains with a stable rhythm. Also, the music signals typically consist of limited number of components, for example, steady musical tones, discrete pitches, and discrete notes as shown in Fig. 2.2. This property means that the spectrogram of a music signal tends to be a low-rank matrix compared with a speech spectrogram. Here, I explain this property with some example music and speech signals.

Figures 2.3–2.6 show the power spectrograms of drums, guitar, vocals, and male speech signals, respectively. These audio signals are obtained from SiSEC2011 dataset [122]. The power spectrograms are calculated by STFT using

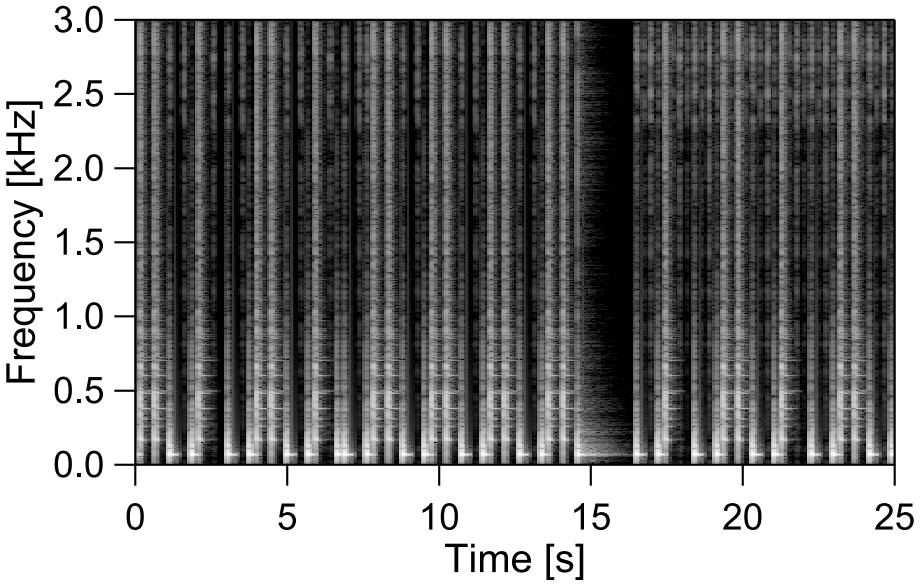

Figure 2.3: Example of power spectrogram of drums sound obtained from "another dreamer-the ones we love" in SiSEC2011 dataset, where grayscale indicates spectral power and white is stronger than black.

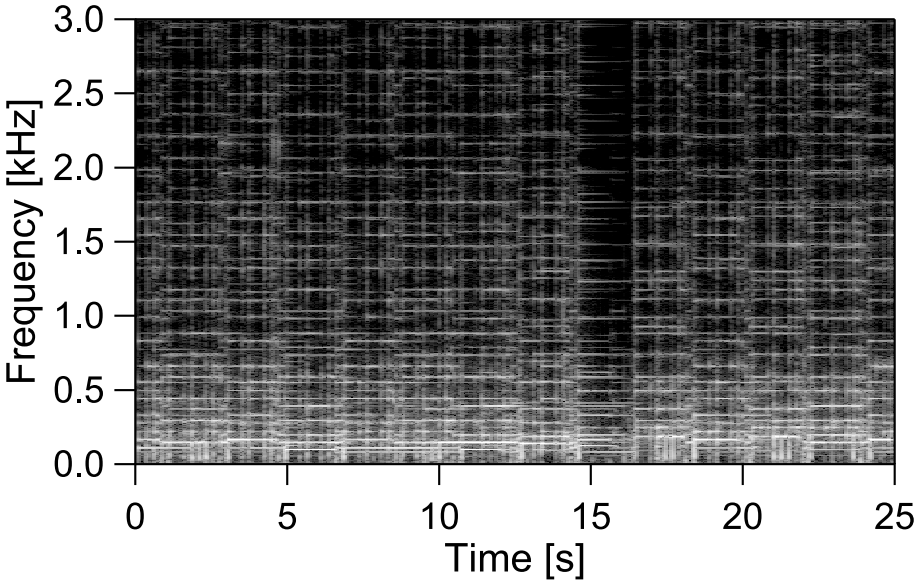

Figure 2.4: Example of power spectrogram of guitar sound obtained from "another dreamer-the ones we love" in SiSEC2011 dataset, where grayscale indicates spectral power and white is stronger than black.

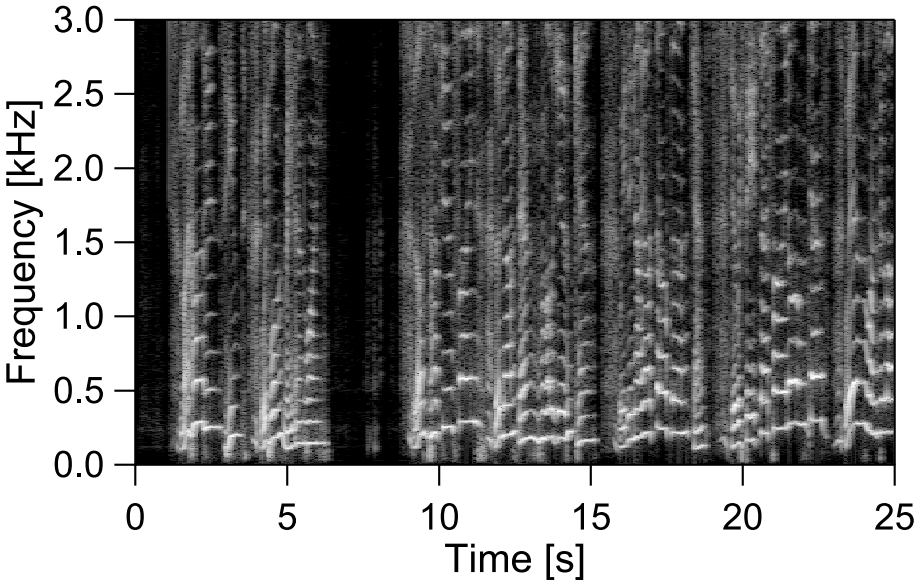

Figure 2.5: Example of power spectrogram of vocals sound obtained from "another dreamer-the ones we love" in SiSEC2011 dataset, where grayscale indicates spectral power and white is stronger than black.

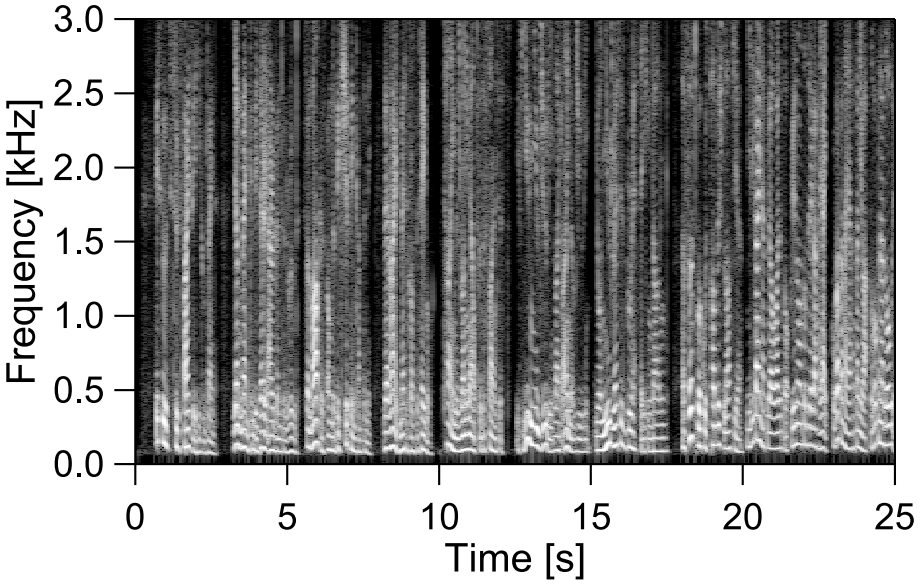

Figure 2.6: Example of power spectrogram of male speech sound obtained from "dev1_male3_src" in SiSEC2011 dataset, where grayscale indicates spectral power and white is stronger than black.

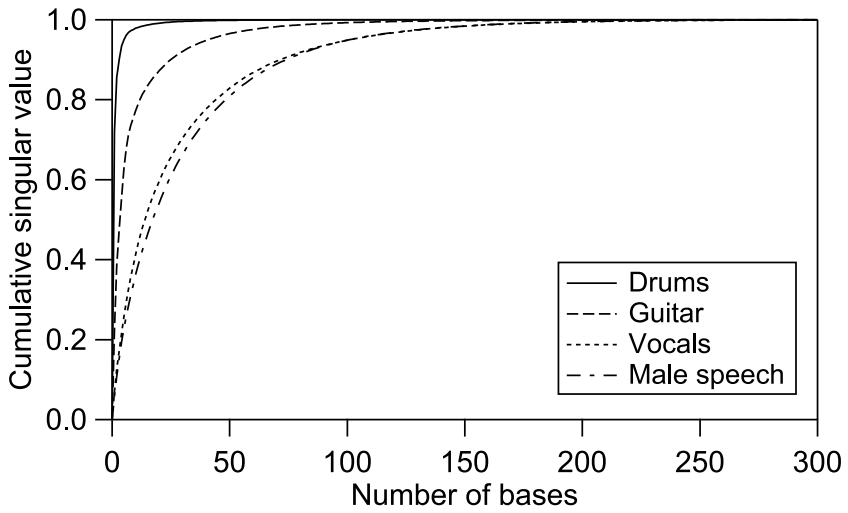

Figure 2.7: Example of cumulative singular values of music and speech spectrograms, where all signals are truncated to be the same signals length.

256-ms-long Hamming window with 192-ms-long overlap. Also, Fig. 2.7 shows their cumulative singular values. It is obviously confirmed that the drums and guitar signals contain many similar or the same spectral patterns compared with the vocals and speech signals, which results in reducing the rank of the power spectrogram as shown in Fig. 2.7. Therefore, if we can extract a limited number of significant spectral patterns from the low-rank music spectrogram as the bases, the signal can effectively be modeled using such bases as the "low-rank approximation." The approximated model spectrogram of each source can directly be used as the strict source model in conventional ICA algorithm.

For this reason, I consider that NMF decomposition is certainly suitable for modeling the spectrogram of music signals because we can blindly extract significant nonnegative bases and their coefficients, which correspond to the frequently appearing spectral patterns and their time-varying gains. Indeed, many tasks related to the music signals have been addressed using NMF, e.g., [123, 124, 79, 125, 71, 126, 127, 100, 128, 129, 130, 131, 132, 98, 99, 133]. Therefore, the unification of ICA-based source separation and the NMF-based source modeling will provide better performance for the multichannel music source separation task.

As another issue, the conventional single-channel source separation based on semi-supervised or full-supervised NMF (SSNMF or FSNMF) does not always provide satisfactory separation although it utilizes the training sequence for target source or all the sources. This is because the pre-trained NMF bases (supervised bases) may represent not only the relevant source but also different sources when the sources share similar or the same spectral patterns. Even if the supervised bases are prepared for all the sources in FSNMF, this problem may occur, resulting in the degradation of separation performance. As a means of avoiding this problem, several techniques for training discriminative supervised bases were proposed [134, 135, 8, 9, 136, 137, 10]. However, these methods are only applicable to the full-supervised situation, and no one investigates the discriminative training for SSNMF.

In this dissertation, I contribute the above-mentioned issues, namely, a unified algorithm of ICA-based BSS and NMF source model for the determined and overdetermined BSS problems in Chap. 3 and a discriminative NMF basis training for single-channel semi-supervised source separation problem in Chap. 4. Since both contributions are based on NMF algorithm, in the next section, I provide a basic principle of NMF as a preliminary for the later chapters.

2.5 Basic Principle of NMF

NMF is an unsupervised data decomposition technique and a latent variable analysis that can be used for, e.g., topics recovery, feature learning, clustering, temporal segmentation, filtering and source separation, and coding. There has been a variety of successful applications: text mining [138, 139], image signal processing (e.g., object discovery [140], face recognition [141], tagging [142], denoising and inpainting [143], texture classification [144], hashing [145], and watermarking [146]), feature extraction [147, 148] and artifact rejection [149, 150] for electroencephalography signals, and even in the bioinformatics field [151, 152, 153]. In audio signal processing, source separation based on NMF for both speech and music signals is well-studied and still a growing research topic as described in Sect. 2.3. Moreover, audio denoising [154, 155], audio inpainting [156, 157], compression [158], and music transcription [123, 159, 160, 161, 125, 162, 163] are

also investigated with NMF.

NMF approximately decomposes an observed nonnegative data matrix $\boldsymbol{\Delta} \in \mathbb{R}_{\geq 0}^{\Phi \times \Psi}$ into a product of two nonnegative matrices $\boldsymbol{F} \in \mathbb{R}_{\geq 0}^{\Phi \times K}$ and $\boldsymbol{G} \in \mathbb{R}_{\geq 0}^{K \times \Psi}$ as

$$\boldsymbol{\Delta} \approx \hat{\boldsymbol{\Delta}} = \boldsymbol{F}\boldsymbol{G}, \tag{2.16}$$

$$\delta_{\phi\psi} \approx \hat{\delta}_{\phi\psi} = \sum_{k=1}^{K} f_{\phi k} g_{k\psi}, \tag{2.17}$$

where $\boldsymbol{F} = (\boldsymbol{f}_1 \ \cdots \ \boldsymbol{f}_K)$ is called *basis matrix* that involves NMF bases $\boldsymbol{f}_k$ as the column vectors, $\boldsymbol{G} = (\boldsymbol{g}_1 \ \cdots \ \boldsymbol{g}_K)^{\mathrm{T}}$ is called *activation matrix* that involves coefficient vectors $\boldsymbol{g}_k$ for each basis as the row vectors, $\delta_{\phi\psi}$, $\hat{\delta}_{\phi\psi}$, $f_{\phi k}$, and $g_{k\psi}$ are the nonnegative entries of the matrices $\boldsymbol{\Delta}$, $\hat{\boldsymbol{\Delta}}$, $\boldsymbol{F}$, and $\boldsymbol{G}$, respectively, Φ and Ψ are the numbers of rows and columns of the matrix $\boldsymbol{\Delta}$, respectively, K is the number of bases, and $\phi = 1, 2, \cdots, \Phi$, $\psi = 1, 2, \cdots, \Psi$, and $k = 1, 2, \cdots, K$ are the integral indices for the rows, columns, and bases. Since the main objective of NMF is to reduce the dimensionality and find a low-rank representation with nonnegative parts, the number of bases K should be set to a small value as $K \ll \min(\Phi, \Psi)$. The basis vectors can be considered as nonnegative parts representing the observed data, where only linear combinations with nonnegative coefficients are allowed. As a result, the obtained bases and activations can often be interpreted intuitively, and they should be the latent and meaningful features in the observed data matrix $\boldsymbol{\Delta}$. For audio signals, a magnitude or power spectrogram obtained via STFT is often used as the input nonnegative data matrix $\boldsymbol{\Delta}$. In this case, the bases vectors correspond to the frequently-appearing spectral patterns in the spectrogram, and the activation vectors represent time-varying gains of each spectral pattern. Figure 2.8 shows an example of NMF decomposition for audio spectrogram data, where $\boldsymbol{\Delta}$ includes two harmonic tones with some overlaps. The obtained basis vectors in $\boldsymbol{F}$ correctly capture the spectra of each tone, and their gains are represented as the activation vectors in $\boldsymbol{G}$. In this decomposition, the two rank-1 matrices, $\boldsymbol{f}_1\boldsymbol{g}_1^{\mathrm{T}}$ and $\boldsymbol{f}_2\boldsymbol{g}_2^{\mathrm{T}}$, represent each tonal spectrogram, and they are superposed for approximately representing the observed spectrogram as $\boldsymbol{\Delta} \approx \boldsymbol{f}_1\boldsymbol{g}_1^{\mathrm{T}} + \boldsymbol{f}_2\boldsymbol{g}_2^{\mathrm{T}}$. Therefore, the source separation based on NMF can be considered as a clustering problem of K bases into N sources.

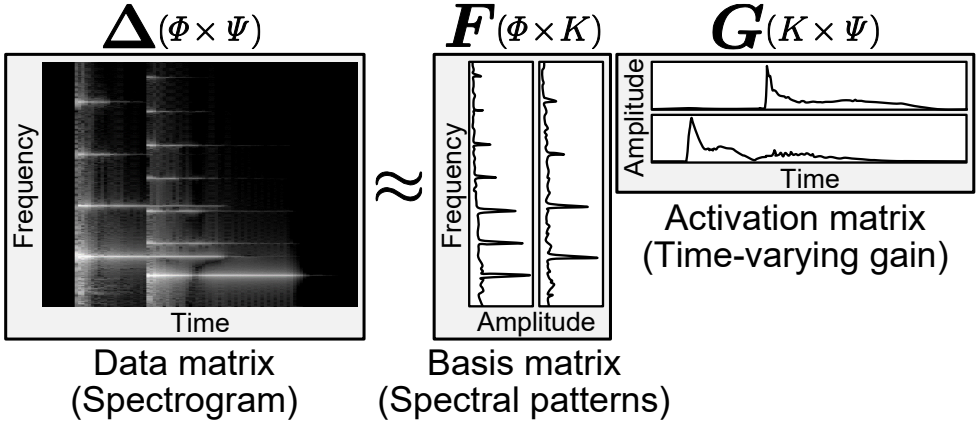

Figure 2.8: Decomposition model of simple NMF, where $K = 2$. Basis matrix involves representative spectral patterns, and activation matrix represents time-varying gains for each basis.

In NMF, we find the decomposed bases and their activations by minimization of the distance or divergence between the observed data $\boldsymbol{\Delta}$ and an approximated model $\hat{\boldsymbol{\Delta}} = \boldsymbol{F}\boldsymbol{G}$ as follows:

$$\min_{\boldsymbol{F},\boldsymbol{G}} \mathcal{D}(\boldsymbol{\Delta} \| \hat{\boldsymbol{\Delta}}) \quad \text{s.t. } f_{\phi k}, g_{k\psi} \geq 0 \quad \forall \phi, \psi, k, \tag{2.18}$$

where $\mathcal{D}(\cdot\|\cdot)$ denotes an arbitrary divergence (dissimilarity). Various divergences have been utilized, e.g., the squared Euclidean distance (EU distance), generalized Kullback–Leibler divergence (KL divergence) [164], and Itakura–Saito divergence (IS divergence) [165]. As a generalized criterion, β divergence [166] $\mathcal{D}_\beta(\cdot\|\cdot)$ is also introduced into NMF:

$$\mathcal{D}_\beta(\boldsymbol{\Delta} \| \hat{\boldsymbol{\Delta}}) = \begin{cases} \displaystyle\sum_{\phi,\psi} \left(\frac{\delta_{\phi\psi}^\beta}{\beta(\beta-1)} + \frac{\hat{\delta}_{\phi\psi}^\beta}{\beta} - \frac{\delta_{\phi\psi}\hat{\delta}_{\phi\psi}^{\beta-1}}{\beta-1} \right) & (\beta \in \mathbb{R}_{\setminus\{0,1\}}) \\ \displaystyle\sum_{\phi,\psi} \left(\delta_{\phi\psi} \log \frac{\delta_{\phi\psi}}{\hat{\delta}_{\phi\psi}} + \hat{\delta}_{\phi\psi} - \delta_{\phi\psi} \right) & (\beta = 1) \\ \displaystyle\sum_{\phi,\psi} \left(\frac{\delta_{\phi\psi}}{\hat{\delta}_{\phi\psi}} - \log \frac{\delta_{\phi\psi}}{\hat{\delta}_{\phi\psi}} - 1 \right) & (\beta = 0) \end{cases} . \tag{2.19}$$

When $\beta = 2$, $\beta = 1$, and $\beta = 0$, the β divergence becomes identical to EU distance, KL divergence, and IS divergence, respectively. In the context of audio source separation, KL divergence and IS divergence are often used because they usually give us a better separation performance than EU distance.

Since the simultaneous minimization of $\boldsymbol{F}$ and $\boldsymbol{G}$ based on (2.18) is not convex, regardless of the type of divergence, the closed-form solution for (2.18) have yet to be found. However, for the alternating optimization of $\boldsymbol{F}$ and $\boldsymbol{G}$, efficient iterative update rules have been derived [5, 12, 167, 168] using the auxiliary function technique [12], which is the generalized optimization approach of expectation-maximization (EM) algorithm [169] and is also known as the majorization-minimization algorithm [170, 171, 172]. For β-divergence-based NMF, the following multiplicative update (MU) rules efficiently minimize the value of cost function (2.18):

$$f_{\phi k} \leftarrow f_{\phi k} \left[\frac{\sum_{\psi} \delta_{\phi\psi} g_{k\psi} \left(\sum_{k'} f_{\phi k'} g_{k'\psi} \right)^{\beta-2}}{\sum_{\psi} g_{k\psi} \left(\sum_{k'} f_{\phi k'} g_{k'\psi} \right)^{\beta-1}} \right]^{\varphi(\beta)}, \tag{2.20}$$

$$g_{k\psi} \leftarrow g_{k\psi} \left[\frac{\sum_{\phi} f_{\phi k} \delta_{\phi\psi} \left(\sum_{k'} f_{\phi k'} g_{k'\psi} \right)^{\beta-2}}{\sum_{\phi} f_{\phi k} \left(\sum_{k'} f_{\phi k'} g_{k'\psi} \right)^{\beta-1}} \right]^{\varphi(\beta)}, \tag{2.21}$$

where $\varphi(\beta)$ is given by

$$\varphi(\beta) = \begin{cases} \dfrac{1}{2-\beta} & (\beta < 1) \\ 1 & (1 \leq \beta \leq 2) \\ \dfrac{1}{\beta - 1} & (\beta > 2) \end{cases}. \tag{2.22}$$

The update rules (2.20)–(2.21) can also be rewritten in a matrix form as follows:

$$\boldsymbol{F} \leftarrow \boldsymbol{F} \circ \left\{ \frac{\left[\boldsymbol{\Delta} \circ (\boldsymbol{F}\boldsymbol{G})^{\cdot(\beta-2)}\right]\boldsymbol{G}^{\mathrm{T}}}{\left[(\boldsymbol{F}\boldsymbol{G})^{\cdot(\beta-1)}\right]\boldsymbol{G}^{\mathrm{T}}} \right\}^{\cdot\varphi(\beta)}, \tag{2.23}$$

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \circ \left\{ \frac{\boldsymbol{F}^{\mathrm{T}}\left[\boldsymbol{\Delta} \circ (\boldsymbol{F}\boldsymbol{G})^{\cdot(\beta-2)}\right]}{\boldsymbol{F}^{\mathrm{T}}\left[(\boldsymbol{F}\boldsymbol{G})^{\cdot(\beta-1)}\right]} \right\}^{\cdot\varphi(\beta)}, \tag{2.24}$$

where $\circ$ and the quotient symbol for matrices denote the Hadamard product (entrywise multiplication) and entrywise division, respectively, and the dotted exponent for matrices denotes entrywise exponent. The iteration of these update rules ensures the monotonic decrease of the cost function. In (2.20) and (2.21), the initial values for $f_{\phi k}$ and $g_{k\psi}$ must be given for all ϕ, ψ, and k. This fact means that the initial values influence the decomposed solution because there exist many local minimum solutions. Since all the results in NMF-based applications including the methods treated in Chaps. 3 and 4 directly depend on these initial values, the effective initialization method for NMF is one of the big problems. This issue will be treated in Chap. 5.

2.6 Summary

In this chapter, a mathematical formulation for general source separation problems was introduced. Next, typical and popular source separation methods were reviewed. In addition, motivations for developing new source separation algorithms were explained. Finally, a key ingredient of this dissertation, NMF, was introduced with theoretical and mathematical principles. In the following chapters, the main contribution of this dissertation will be discussed; determined and overdetermined BSS will be addressed in Chap. 3, single-channel semi-supervised source separation will be treated in Chap. 4, and a better initialization for NMF will be addressed in Chap. 5.

3

Determined and Overdetermined Blind Source Separation Based on Independent Low-Rank Matrix Analysis

3.1 Introduction

In this chapter, I address the determined BSS problem and propose a new efficient algorithm that unifies conventional ICA-based BSS and NMF-based source model. First, I introduce some basic principles of ICA, FDICA, IVA [3, 4, 173], and NMF based on IS divergence (hereafter referred to as ISNMF) [125], which are necessary for the main contribution in this chapter. Next, a new efficient BSS technique is described with its motivations and optimization algorithms.

After giving an explanation of the relationship between IVA, MNMF, and the proposed method, the efficacy of the proposed method for BSS task is validated via experimental analysis and comparison. In addition, further extension of the proposed method for overdetermined BSS problem is addressed. Finally, the whole contents in this chapter are summarized.

3.2 Basic Principles of ICA, FDICA, IVA, and ISNMF

While many text books for ICA have been published [30, 31, 32, 33] so far, I briefly introduce some basic principles of ICA and FDICA. Also, IVA, and ISNMF are explained with their motivations. These algorithms and their statistical backgrounds are important and necessary for introducing the motivations of the proposed method.

3.2.1 ICA and FDICA

ICA, which is sometimes regarded as synonymous with BSS, relies on non-Gaussianity, namely, the independent sources $\tilde{s}_n(\tau)$ in the observed mixture $\tilde{x}_m(\tau)$ are inherently generated from non-Gaussian distributions. This assumption makes the BSS problem solvable because the mixture of sources tends towards a Gaussian distribution even if the original sources themselves obey non-Gaussian distributions, which is known as the central limit theorem. Therefore, if we determine the statistical model of separated signals $\tilde{y}_n(\tau)$ as $p(\tilde{y}_n)$ and if $M = N$ (determined) and $L_{\text{filter}} = 1$ (instantaneous mixture), we can basically estimate the original $\tilde{y}_n(\tau)$ from the mixture $\tilde{\boldsymbol{x}}(\tau)$ by ICA, while the sources must truly obey the non-Gaussian distributions.

Let us assume that $M = N$, $L_{\text{filter}} = 1$, and the mixing system is described as

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} \tag{3.1}$$

$$\equiv \boldsymbol{A}\boldsymbol{y}, \tag{3.2}$$

where tildes and the time index τ are omitted for the simplicity, and all the

variables are real values. The probability $p(\boldsymbol{x})$ can be written using Jacobian

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = |\det \boldsymbol{A}^{-1}|, \tag{3.3}$$

as

$$p(\boldsymbol{x}) = |\det \boldsymbol{A}^{-1}| \cdot p(\boldsymbol{y}) \tag{3.4}$$

$$= |\det \boldsymbol{W}| \cdot \prod_n p(y_n), \tag{3.5}$$

where $\boldsymbol{W} = \boldsymbol{A}^{-1}$, and $p(\boldsymbol{s}) = \prod_n p(s_n)$ are obtained by assuming mutual independence between y_n for all the sources. Since the signals $\boldsymbol{x}$, $\boldsymbol{s}$, and $\boldsymbol{y}$ have τ_{end} samples, the likelihood of $\boldsymbol{W}$ is given by

$$\mathcal{L}(\boldsymbol{W}) = \prod_\tau \prod_n p(\boldsymbol{w}_n^{\mathrm{T}} \boldsymbol{x}(\tau))|\det \boldsymbol{W}|, \tag{3.6}$$

where $\boldsymbol{w}_n$ is a row vector in $\boldsymbol{W}$. Also, the log-likelihood function is obtained as

$$\log \mathcal{L}(\boldsymbol{W}) = \sum_\tau \sum_n p(\boldsymbol{w}_n^{\mathrm{T}} \boldsymbol{x}(\tau)) + \tau_{\text{end}} \log |\det \boldsymbol{W}|. \tag{3.7}$$

By replacing the sum of τ to the expectation operator $\mathrm{E}[\cdot]$, we can rewrite (3.7) as

$$\frac{1}{\tau_{\text{end}}} \log \mathcal{L}(\boldsymbol{W}) = \mathrm{E}\left[\sum_n p(\boldsymbol{w}_n^{\mathrm{T}} \boldsymbol{x})\right] + \log |\det \boldsymbol{W}|. \tag{3.8}$$

The maximum likelihood (ML) estimation in ICA can be obtained by differentiating (3.8) with respect to $\boldsymbol{W}$. The gradient of (3.8) can easily be calculated as

$$\frac{1}{\tau_{\text{end}}} \frac{\partial \log \mathcal{L}(\boldsymbol{W})}{\partial \boldsymbol{W}} = \mathrm{E}\left[\mathcal{S}(\boldsymbol{W}\boldsymbol{x})\boldsymbol{x}^{\mathrm{T}}\right] + \left(\boldsymbol{W}^{\mathrm{T}}\right)^{-1}, \tag{3.9}$$

where

$$\mathcal{S}(\boldsymbol{y}) = \frac{\partial \log p(\boldsymbol{y})}{\partial \boldsymbol{y}} \tag{3.10}$$

is called the score function or nonlinear function in the context of ICA optimization. From (3.9), the steepest gradient descent for ICA can be derived as

$$\boldsymbol{W} \leftarrow \boldsymbol{W} + \eta \left\{ \mathrm{E}\left[\mathcal{S}(\boldsymbol{W}\boldsymbol{x})\boldsymbol{x}^{\mathrm{T}} \right] + \left(\boldsymbol{W}^{\mathrm{T}} \right)^{-1} \right\}, \tag{3.11}$$

where η is a stepsize parameter. This algorithm is often called Bell–Sejnowski algorithm, which is firstly derived from another ICA principle called Infomax approach [174]. The algorithm (3.11) is extended to a natural gradient [175, 33] method that is based on a geometric structure in the parameter space, as

$$\boldsymbol{W} \leftarrow \boldsymbol{W} + \eta \left\{ \mathrm{E}\left[\mathcal{S}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{y}^{\mathrm{T}} \right] + \boldsymbol{I} \right\} \boldsymbol{W}. \tag{3.12}$$

This algorithm is more efficient than (3.11) because the gradient for the parameters are defined in Riemannian metric, and the inverse calculation of $\boldsymbol{W}$ is omitted. In addition, for the minimization of (3.9), fast and stable update rules called iterative projection (IP) based on the auxiliary function technique have been proposed [176]. Note that there exists signal permutation and scaling ambiguities in ICA solution, namely,

$$\boldsymbol{y} \leftarrow \boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{y} \tag{3.13}$$

is also a solution for any permutation matrix $\boldsymbol{\Upsilon}$ and diagonal matrix $\boldsymbol{\Lambda}$. This is because the solution in ICA is only based on a statistic source model $p(\tilde{y}_n)$ and its independence between the sources. In many ICA applications, the source model $p(\tilde{y}_n)$ or its score function is empirically set to the appropriate ones, e.g., Laplace distribution for super-Gaussian sources.

For the BSS problem of acoustic signals, the sources are convolved with reverberation in a recording environment as represented in (2.4), and it becomes a deconvolution problem. Many ICA-based deconvolution techniques for solving BSS problem in time-domain were proposed [36, 37, 38, 39, 40]. However, the estimation of inverse filters in the time domain is still a tough problem because the number of parameters drastically increases when the filter length L_{filter} becomes large. Instead of solving the time-domain deconvolution,

FDICA [2, 41, 42, 34, 43, 35] was proposed. In this approach, the instantaneous mixture in the frequency domain is assumed as (2.6). When the mixing filter length L_{filter} is much shorter than the length of analysis window in STFT, this assumption becomes valid. In FDICA, the simple ICA is independently applied in each of frequency bins, and the frequency-wise demixing matrix $\boldsymbol{W}_i$ is estimated for the separation. The optimization algorithm for FDICA is identical to that for simple ICA with complex-valued signals [121, 177]. However, in FDICA, the permutation ambiguity of separated signals becomes a serious problem because the separated components in each frequency bin must be correctly aligned as shown in Fig. 3.1. This alignment problem is often called permutation problem, and many criteria for solving this ambiguity have been proposed [178, 179, 180, 181]. The popular permutation solver is using direction of arrival (DOA) of each source [178], where DOA can be calculated from the estimated demixing filter $\boldsymbol{w}_n$. In [180], the correlations between frequency bins are simultaneously exploited with DOA information for solving the permutation problem. The scaling ambiguity should also be solved after the estimation of $\boldsymbol{W}_i$ in FDICA. The simplest way for recovering signal scales is projecting them to the observed signals, as

$$\hat{\boldsymbol{c}}_{ij,n} \leftarrow \boldsymbol{W}_i^{-1} (\boldsymbol{e}_n \circ \boldsymbol{y}), \tag{3.14}$$

where $\hat{\boldsymbol{c}}_{ij,n} = (\hat{c}_{ij,n1} \ \cdots \ \hat{c}_{ij,nM})^{\mathrm{T}}$ is an estimated source image whose scale is fitted to the observed signals at each microphone, and $\boldsymbol{e}_n$ denotes the $M \times 1$ unit vector with the nth element equal to unity. This calculation is called the back projection technique [179].

3.2.2 IVA

IVA [3, 4, 173] is a multivariate extension of FDICA and can solve the BSS problem while avoiding the permutation problem. In IVA, we assume the multivariate source vector $\boldsymbol{s}_{j,n}$, observed vector $\boldsymbol{x}_{j,m}$, and separated vector $\boldsymbol{y}_{j,n}$,

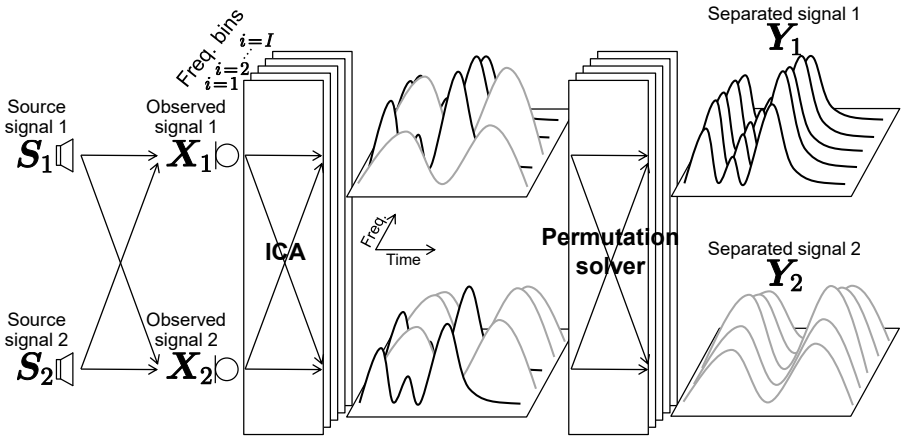

Figure 3.1: Permutation problem in FDICA and its solver, where $N = M = 2$.

which consist of all the frequency bins, as

$$\boldsymbol{s}_{j,n} = (s_{1j,n} \ \cdots \ s_{Ij,n})^{\mathrm{T}}, \tag{3.15}$$

$$\boldsymbol{x}_{j,m} = (x_{1j,m} \ \cdots \ x_{Ij,m})^{\mathrm{T}}, \tag{3.16}$$

$$\boldsymbol{y}_{j,n} = (y_{1j,n} \ \cdots \ y_{Ij,n})^{\mathrm{T}}. \tag{3.17}$$

Figure 3.2 shows the mixing and demixing model in IVA, where $N = M = 2$. In IVA, all the source, observed, and separated signals are represented as frequency vector variables, whereas FDICA independently models each of the frequency components resulting in the permutation problem. In addition, higher-order correlations between the frequency components in each source (or separated) vector are introduced by assuming spherically symmetric multivariate source distributions $p(\boldsymbol{s}_{j,n}) \approx p(\boldsymbol{y}_{j,n}) = p(y_{1j,n}, \cdots, y_{Ij,n})$, where the spherically symmetric property means that the distribution is a function of only the norm of multivariate vector variable, i.e., $p(\boldsymbol{y}_{j,n}) = f(\|\boldsymbol{y}_{j,n}\|)$.

In the literature [3, 4, 173], a spherically symmetric multivariate Laplace distribution [182, 183] was exploited as a super-Gaussian source distribution for

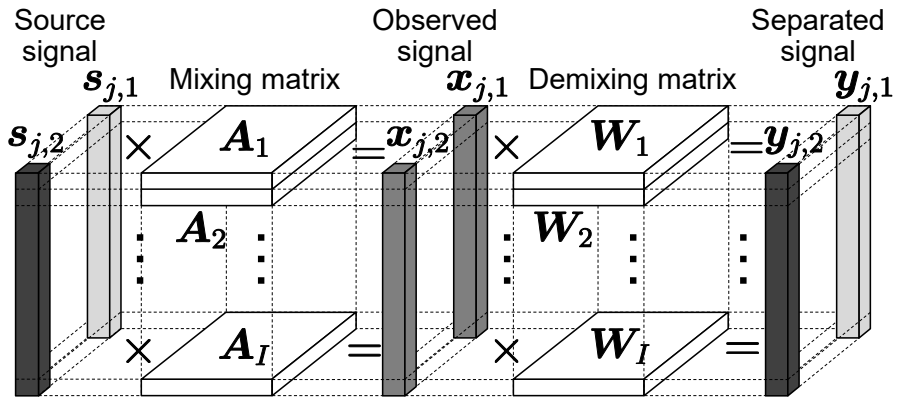

Figure 3.2: Mixing and demixing model in IVA, where $N = M = 2$.

modeling speech sources. This distribution is shown in Fig. 3.3 and is defined as

$$p(\boldsymbol{s}_{j,n}) \approx p(\boldsymbol{y}_{j,n}) = \rho \exp\left(-\sqrt{\sum_i \left|\frac{y_{ij,n}}{r_{i,n}}\right|^2}\right), \tag{3.18}$$

where ρ is a normalization term and $r_{i,n}$ is the variance, which determines the signal scale of $y_{ij,n}$. Since the source distribution has a spherically symmetric property, higher-order correlations between the frequency components in each source are assumed, which results in avoiding the permutation problem. Hereafter, IVA based on the source distribution (3.18) is referred to as *Laplace IVA*.

From the generative source model $p(\boldsymbol{s}_{j,1}, \cdots, \boldsymbol{s}_{j,N}) \approx p(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N})$ and the demixing system (2.9), $p(\boldsymbol{x}_{j,1}, \cdots, \boldsymbol{x}_{j,M})$ can be obtained by multiplying $p(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N})$ by the Jacobian

$$\frac{\partial(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N})}{\partial(\boldsymbol{x}_{j,1}, \cdots, \boldsymbol{x}_{j,M})} = \prod_i |\det \boldsymbol{W}_i|^2; \tag{3.19}$$

note that the Jacobian for a complex-valued variable is the square of the Jacobian for a real-valued variable [177]. Therefore, the likelihood function $\mathcal{L}(\boldsymbol{W})$ of the

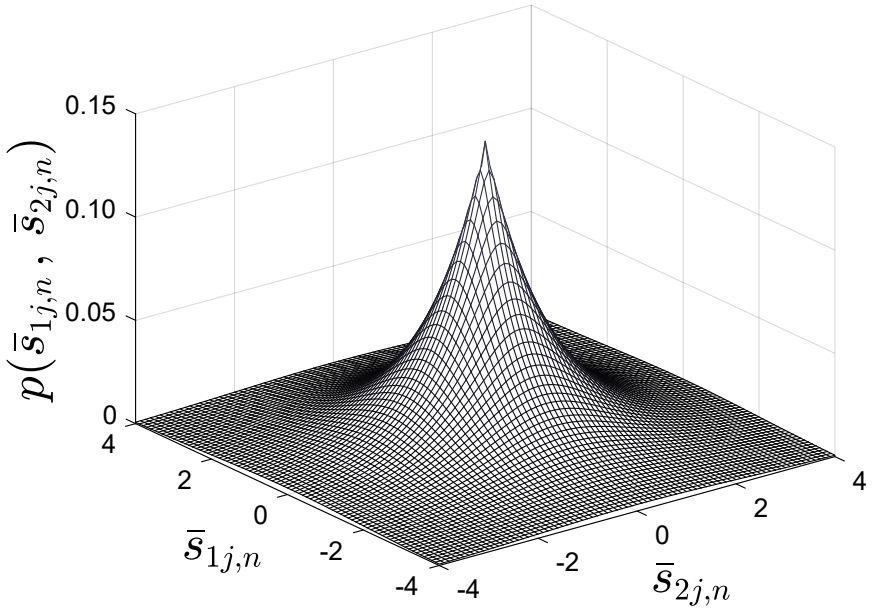

Figure 3.3: Spherically symmetric multivariate Laplace distribution, where $\bar{s}_{ij,n}$ can be considered as either real or imaginary part of $s_{ij,n}$ and $I = 2$ (bivariate case). Two frequency components $s_{1j,n}$ and $s_{2j,n}$ are uncorrelated but have mutual dependences, which is called higher-order correlation.

parameter set $\boldsymbol{W} = \{\boldsymbol{W}_i | i = 1, \cdots, I\}$ is given as

$$\mathcal{L}(\boldsymbol{W}) = \prod_j p(\boldsymbol{x}_{j,1}, \cdots, \boldsymbol{x}_{j,M} | \boldsymbol{W}) \tag{3.20}$$

$$= \prod_j \left[p(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N}) \cdot \prod_i |\det \boldsymbol{W}_i|^2 \right] \tag{3.21}$$

$$= \prod_j \left\{ \left[\prod_n p(\boldsymbol{y}_{j,n}) \right] \cdot \prod_i |\det \boldsymbol{W}_i|^2 \right\}, \tag{3.22}$$

where $p(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N}) = \prod_n p(\boldsymbol{y}_{j,n})$ is obtained by assuming mutual independence between $\boldsymbol{y}_{j,n}$ for all the sources. The negative log-likelihood function can be

calculated as

$$-\log \mathcal{L}(\boldsymbol{W}) = -\sum_{i,j} \log |\det \boldsymbol{W}_i|^2 - \sum_{j,n} \log p(\boldsymbol{y}_{j,n}) \tag{3.23}$$

$$= -2J \sum_{i} \log |\det \boldsymbol{W}_i| + \sum_{j,n} \mathcal{G}(\boldsymbol{y}_{j,n}), \tag{3.24}$$

where $\mathcal{G}(\boldsymbol{y}_{j,n}) = -\log p(\boldsymbol{y}_{j,n})$ is called the contrast function, which depends on the source distribution $p(\boldsymbol{y}_{j,n})$. Note that since $y_{ij,n} = \boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}$, the separated signal $\boldsymbol{y}_{j,n}$ includes the optimization variable $\boldsymbol{W}_i$. The ML estimation based on (3.24) is equivalent to the well-known estimation [29, 32] that maximizes the independence between all the sources with the KL divergence $\mathcal{D}_{\mathrm{KL}}$ as follows:

$$\sum_{j} \mathcal{D}_{\mathrm{KL}} \left(p(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N}) \| \prod_{n} p(\boldsymbol{y}_{j,n}) \right)$$

$$= \sum_{j} \int p(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N}) \log \frac{p(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N})}{\prod_n p(\boldsymbol{y}_{j,n})} d\boldsymbol{y}_{j,1} \cdots d\boldsymbol{y}_{j,N} \tag{3.25}$$

$$= \text{const.} - 2J \sum_{i} \log |\det \boldsymbol{W}_i| + \sum_{j,n} \mathcal{G}(\boldsymbol{y}_{j,n}). \tag{3.26}$$

On the basis of the source distribution (3.18), the contrast function $\mathcal{G}(\boldsymbol{y}_{j,n})$ and the cost function in Laplace IVA can be obtained as follows:

$$\mathcal{G}(\boldsymbol{y}_{j,n}) = -\log \rho + \|\boldsymbol{y}_{j,n}\|_2, \tag{3.27}$$

$$-\log \mathcal{L}(\boldsymbol{W}) = \text{const.} - 2J \sum_{i} \log |\det \boldsymbol{W}_i| + \sum_{j,n} \|\boldsymbol{y}_{j,n}\|_2, \tag{3.28}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. Also, the variance is set to $r_{i,n} = 1$ for all i and n because the scales of separated signals cannot be determined by ICA or IVA, and they can be recovered by the back-projection technique (3.14) after the separation. Similar to the simple ICA, IP-based efficient update rules have been proposed [184, 185, 186].

Figure 3.4 shows the principles of source estimation in FDICA and Laplace IVA. The demixing matrix $\boldsymbol{W}_i$ is optimized so that the estimated signals $y_{ij,n}$ obay the assumed non-Gaussian source model. Whereas FDICA assumes the

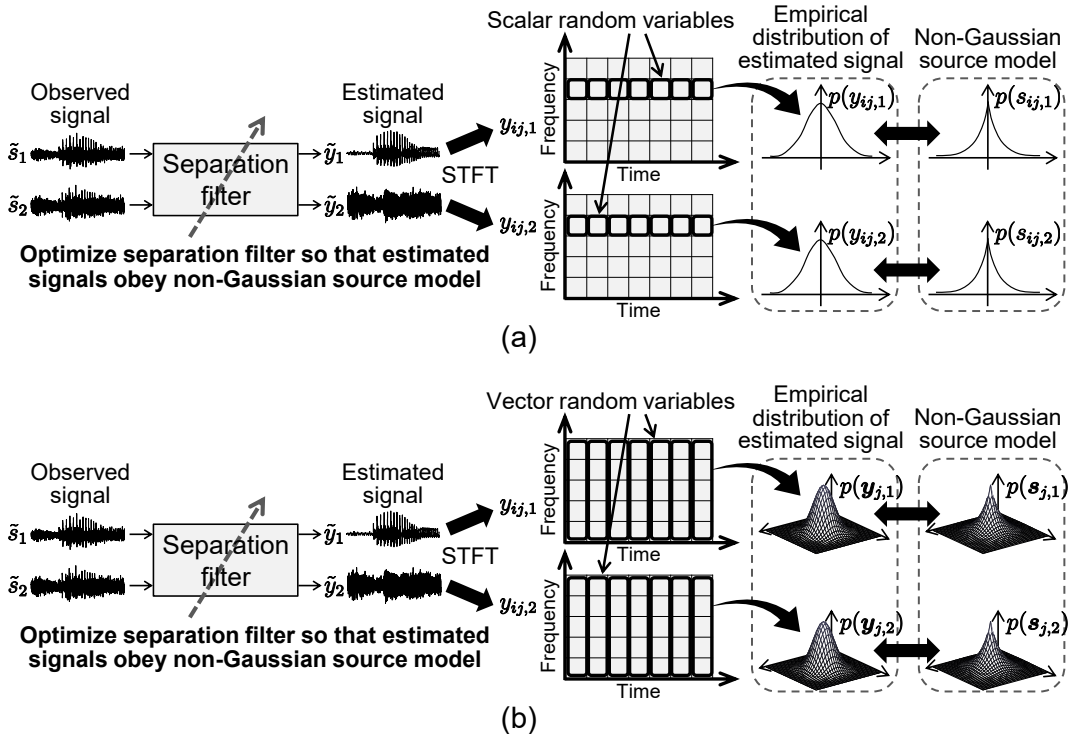

Figure 3.4: Principles of source estimation in (a) FDICA and (b) Laplace IVA. Separation filter (demixing matrix) is optimized so that estimated signals obey non-Gaussian source model. Whereas FDICA assumes non-Gaussian source distribution $p(s_{ij,n})$ for each frequency component, IVA assumes non-Gaussian multivariate source distribution $p(\boldsymbol{s}_{j,n})$ that has spherically symmetric property.

non-Gaussian source distribution $p(s_{ij,n})$ for each frequency component, IVA assumes the non-Gaussian spherically symmetric source distribution $p(\boldsymbol{s}_{j,n})$ for the frequency vector variables.

3.2.3 ISNMF

When we apply NMF to an acoustic signal, the power spectrogram obtained via STFT is considered as an observed nonnegative matrix and can be decomposed into two nonnegative matrices as

$$|\boldsymbol{D}|^{\cdot 2} \approx \boldsymbol{T}\boldsymbol{V}, \tag{3.29}$$

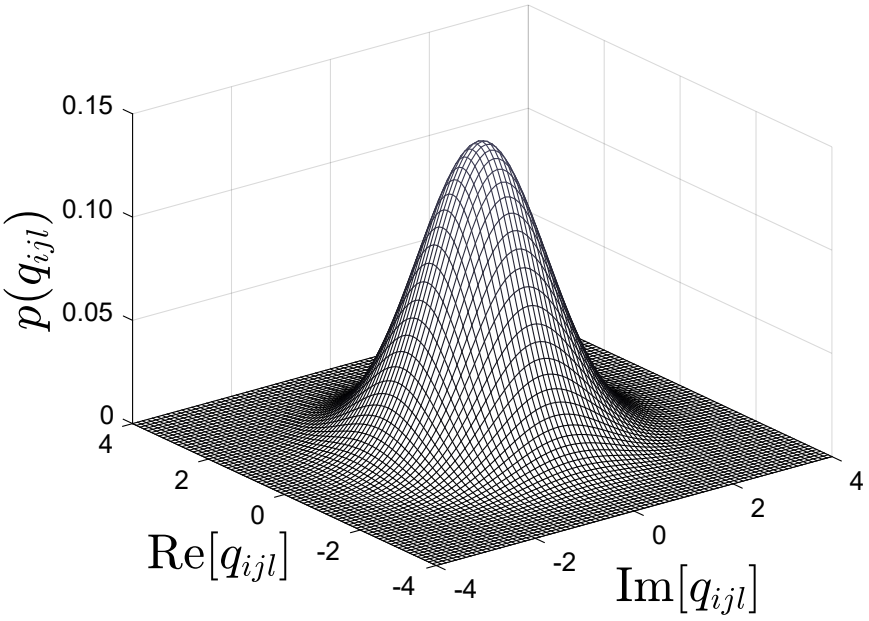

Figure 3.5: Circularly symmetric complex Gaussian distribution. Probability does not depend on phase $\arg(q_{ijl})$ but only depends on amplitude $|q_{ijl}|$ or power $|q_{ijl}|^2$ because of circularly symmetric property.

where $\boldsymbol{D} \in \mathbb{C}^{I \times J}$ is a complex-valued spectrogram, the absolute $|\cdot|$ for matrices denote the entrywise absolute value, $\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{I \times L}$ is the basis matrix, $\boldsymbol{V} \in \mathbb{R}_{\geq 0}^{L \times J}$ is the activation matrix, and L is the number of bases.

Since the transformation of complex spectrograms into power (or amplitude) spectrograms is nonlinear, power spectrograms are non-additive, namely, the power spectrum of the sum of two waveforms is not equal to the sum of the power spectra of the two waveforms. This implies that decomposing a power spectrogram into the sum of additive components does not necessarily lead to an appropriate decomposition of the audio signal. However, the decomposition based on ISNMF applied to the power spectrogram ensues such spectral additivity in the expectation sense, which has been given by Févotte, et al. [125].

Let us assume that L complex-valued spectrograms $q_{ij1}, \cdots, q_{ijL}$ are generated from circularly symmetric complex Gaussian distribution [187], which are

independently defined in each time-frequency slot as follows:

$$p(q_{ijl}) = \frac{1}{\pi r_{ijl}} \exp\left(-\frac{|q_{ijl}|^2}{r_{ijl}}\right), \tag{3.30}$$

where $l = 1, \cdots, L$ is the integral index of L components and r_{ijl} is the nonnegative variance of each distribution. Figure 3.5 shows the circularly symmetric complex Gaussian distribution. Since the distribution has a circularly symmetric property in the complex plane, the probability does not depend on the phase $\arg(q_{ijl})$ and only depends on the amplitude $|q_{ijl}|$ or power $|q_{ijl}|^2$. Note that the variance r_{ijl} corresponds to the expectation value of the power spectrum $|q_{ijl}|^2$, namely, $r_{ijl} = \mathrm{E}[|q_{ijl}|^2]$. When the variance r_{ijl} is large, the distribution becomes wider, and the complex-valued spectrum q_{ijl} with a large power can easily be generated, while the phase of q_{ijl} is always uniformly distributed. In addition, if we assume that the observation d_{ij}, which is the complex-valued entry of $\boldsymbol{D}$, is the sum of the components q_{ijl}, namely, $d_{ij} = \sum_l q_{ijl}$, the following generative model can also be assumed because of the reproductive property in complex Gaussian distributions:

$$p(\boldsymbol{D}) = \prod_{i,j} p(d_{ij}) \tag{3.31}$$

$$= \prod_{i,j} \frac{1}{\pi r_{ij}} \exp\left(-\frac{|d_{ij}|^2}{r_{ij}}\right), \tag{3.32}$$

where $r_{ij} = \sum_l r_{ijl}$. This fact means that the additivity of power spectra $|q_{ijl}|^2$ is held only in the expectation sense, and the superposed component $d_{ij} = \sum_l q_{ijl}$ is also assumed to obey the circularly symmetric complex Gaussian distribution with the superposed variance $r_{ij} = \sum_l r_{ijl}$. In [188], the model of power spectrogram (3.32) is extended to a maximum a posteriori framework using inverse Gamma prior for the variances.

The likelihood function of $\boldsymbol{T}$ and $\boldsymbol{V}$ can be obtained as follows by putting

$r_{ijl} = t_{il}v_{lj}$;

$$\mathcal{L}(\boldsymbol{T}, \boldsymbol{V}) = p(\boldsymbol{D}|\boldsymbol{T}, \boldsymbol{V}) \tag{3.33}$$

$$= \prod_{i,j} \frac{1}{\pi \sum_l t_{il}v_{lj}} \exp\left(-\frac{|d_{ij}|^2}{\sum_l t_{il}v_{lj}}\right), \tag{3.34}$$

where t_{il} and v_{lj} are the nonnegative entries of $\boldsymbol{T}$ and $\boldsymbol{V}$, respectively. The negative log-likelihood function is

$$-\log \mathcal{L}(\boldsymbol{T}, \boldsymbol{V}) = \sum_{i,j} \left(\log \pi + \log \sum_l t_{il}v_{lj} + \frac{|d_{ij}|^2}{\sum_l t_{il}v_{lj}}\right). \tag{3.35}$$

It is clear that the ML estimation based on (3.35) is equivalent to the minimization of IS divergence $\mathcal{D}_{\mathrm{IS}}$ [165] between $|\boldsymbol{D}|^{\cdot 2}$ and $\boldsymbol{TV}$:

$$\mathcal{D}_{\mathrm{IS}}\left(|\boldsymbol{D}|^{\cdot 2}\|\boldsymbol{TV}\right) = \sum_{i,j} \left(\frac{|d_{ij}|^2}{\sum_l t_{il}v_{lj}} - \log \frac{|d_{ij}|^2}{\sum_l t_{il}v_{lj}} - 1\right) \tag{3.36}$$

$$= \text{const.} + \sum_{i,j} \left(\frac{|d_{ij}|^2}{\sum_l t_{il}v_{lj}} + \log \sum_l t_{il}v_{lj}\right). \tag{3.37}$$

Thus, when ISNMF is applied to the observed power spectrogram $|\boldsymbol{D}|^{\cdot 2}$, it is assumed that d_{ij} follows the generative model (3.32) and the components q_{ijl} are mutually independent. The multiplicative update rules for $\boldsymbol{T}$ and $\boldsymbol{V}$ that minimize (3.35) or (3.37) are given by [167]

$$t_{il} \leftarrow t_{il} \sqrt{\frac{\sum_j |d_{ij}|^2 v_{lj} \left(\sum_{l'} t_{il'}v_{l'j}\right)^{-2}}{\sum_j v_{lj} \left(\sum_{l'} t_{il'}v_{l'j}\right)^{-1}}}, \tag{3.38}$$

$$v_{lj} \leftarrow v_{lj} \sqrt{\frac{\sum_i |d_{ij}|^2 t_{il} \left(\sum_{l'} t_{il'}v_{l'j}\right)^{-2}}{\sum_i t_{il} \left(\sum_{l'} t_{il'}v_{l'j}\right)^{-1}}}. \tag{3.39}$$

These MU rules are identical to (2.20) and (2.21) when we set $\beta = 0$.

3.2.4 Time-Varying Gaussian IVA

Laplace IVA employs the spherically symmetric Laplace distribution as a super-Gaussian source distribution. The model ensures that all the frequency components in the same source have higher-order correlation. As another super-Gaussian source model with the higher-order correlation, in [104], the circularly symmetric complex Gaussian distribution with time-varying variance $r_{j,n}$ is introduced to conventional IVA instead of the stationary distribution:

$$\begin{aligned} p(\boldsymbol{y}_{1,n}, \cdots, \boldsymbol{y}_{J,n}) &= \prod_j p(\boldsymbol{y}_{j,n}) \\ &= \prod_j \frac{1}{\pi r_{j,n}} \exp\left(-\frac{\|\boldsymbol{y}_{j,n}\|_2^2}{r_{j,n}}\right), \end{aligned} \tag{3.40}$$

where the time-varying variance $r_{j,n}$ is shared over the frequency bins in each time frame. Similar to (3.18), the distribution (3.40) has the spherically symmetric property for the multivariate vector $\boldsymbol{y}_{j,n}$ because $p(\boldsymbol{y}_{j,n})$ only depends on the vector norm $\|\boldsymbol{y}_{j,n}\|_2$. Also, the distribution is assumed to be mutually independent for time frames and sources. Whereas the temporal source model $p(\boldsymbol{y}_{j,n})$ is based on the Gaussian distribution, the global source model $p(\boldsymbol{Y}_n) = p(\boldsymbol{y}_{1,n}, \cdots, \boldsymbol{y}_{J,n})$ becomes the super-Gaussian distribution because of the time-varying variance $r_{j,n}$ [186]. This time-varying Gaussian source model has been adopted for many techniques, e.g., BSS [189, 190] and dereverberation of speech signals [191]. Hereafter, IVA based on the source distribution (3.40) is referred to as *time-varying Gaussian IVA*.

Figure 3.6 (a) shows the source model (variance structure in a time-frequency region) assumed in time-varying Gaussian IVA. Since the variance $r_{j,n}$ is shared over the frequency bins, it can be interpreted as an uniform (flat) spectral basis. On the other hand, ISNMF has a more flexible source model because the variance r_{ij} is independently defined in each time-frequency slot as shown in Fig. 3.6 (b). It allows us to model the specific time-frequency structure with limited numbers of bases and activations. Similar to (3.40) and Fig. 3.6 (a), the time-frequency-varying source model (3.32) and Fig. 3.6 (b) are the super-Gaussian distribution because of the time-frequency-varying variance $r_{ij,n}$.

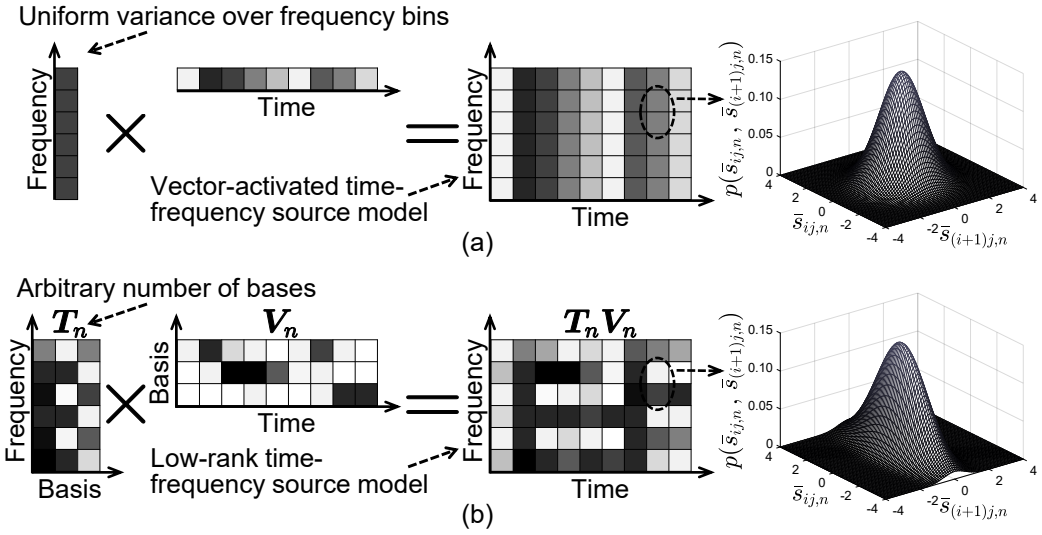

Figure 3.6: Comparison of source models (variance structures) in (a) IVA and (b) ISNMF, where grayscale in each time-frequency slot indicates scale of variance. IVA has uniform variance over frequency bins, and all the frequency bins have the same activations (time-varying gains), whereas ISNMF employs limited number of bases to capture low-rank structure resulting in more flexible source model.

Thus, it can also be used as a source distribution in ICA-based methods.

3.3 Independent Low-Rank Matrix Analysis

3.3.1 Motivation and Strategy

For speech signal separation, Laplace IVA or time-varying Gaussian IVA can achieve better performance than FDICA. However, since only the higher-order correlation defined in (3.18) or (3.40) is utilized as a spectral structure in the source model, IVA cannot treat the specific harmonic structures of each source and lacks flexibility, as shown in Fig. 3.6. For this reason, IVA is not suitable for sources that have characteristic (specific) spectral structures, such as instrumental sounds or music signals. NMF decomposition is suitable for modeling the spectrogram of music or instrumental signals because such signals typically

consist of a limited number of components, for example, steady musical tones, discrete pitches, and discrete notes, as described in Sect. 2.4. This property means that the spectrogram of a music signal tends to be a low-rank matrix compared with a speech spectrogram.

In [104], the temporal power variation of sources provided by a user is exploited as the prior distribution of the time-varying gain $r_{j,n}$, which is defined as an inverse gamma distribution. In [192], a new multichannel source separation method with external model information has been proposed, which is called model-based IVA. In this approach, we consider that the time-frequency variance $r_{ij,n}$ for each source is given by another technique (e.g., single-channel spectral subtraction, voice activity detection, or time-frequency binary masking) applied in advance. The demixing matrix $\boldsymbol{W}_i$ is estimated on the basis of the independence between sources taking the given variance $r_{ij,n}$ into account. These approaches show that the estimation of $\boldsymbol{W}_i$ based on a correct and precise variance will provide better separation performance.

On the basis of these ideas, in this chapter, I introduce ISNMF to IVA for decomposing the sourcewise variance $r_{ij,n}$ using a limited number of NMF bases, where the demixing matrix $\boldsymbol{W}_i$ and the source model $p(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N})$ with the NMF variables are simultaneously estimated in a fully blind fashion. This approach is a natural extension of time-varying Gaussian IVA because we extend the vector source model (frequency-uniform variance) to the low-rank matrix source model (NMF decomposition) as shown in Fig. 3.6. For this reason, hereafter, I call the proposed method *independent low-rank matrix analysis (ILRMA)* [193, 194]. Similarly to standard FDICA or IVA, ILRMA is applicable to the determined case ($M = N$). In the overdetermined case ($M > N$), dimensionality reduction using principal component analysis (PCA) should be applied so that $M = N$.

3.3.2 Derivation of Cost Function

In ILRMA, similarly to ISNMF, the circularly symmetric complex Gaussian distribution is independently assumed to be as follows in each time-frequency

slot as the source model of the separated signal;

$$p(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n) = \prod_j p(\boldsymbol{y}_{j,1}, \cdots, \boldsymbol{y}_{j,N}) \tag{3.41}$$

$$= \prod_{n,j} p(\boldsymbol{y}_{j,n}) \tag{3.42}$$

$$= \prod_{n,i,j} \frac{1}{\pi r_{ij,n}} \exp\left(-\frac{|y_{ij,n}|^2}{r_{ij,n}}\right), \tag{3.43}$$

where $r_{ij,n}$ is the sourcewise variance that corresponds to the expectation of the power spectrogram, namely, $r_{ij,n} = \mathrm{E}[|y_{ij,n}|^2]$. The contrast function and the negative log-likelihood function of the parameter set $\boldsymbol{W}$ and $\boldsymbol{R} = \{r_{ij,n} | i = 1, \cdots, I; j = 1, \cdots J; n = 1, \cdots N\}$ are given as

$$\mathcal{G}(\boldsymbol{y}_{j,n}) = \sum_i \left(\log \pi r_{ij,n} + \frac{|y_{ij,n}|^2}{r_{ij,n}}\right) \tag{3.44}$$

$$= I \log \pi + \sum_i \left(\log r_{ij,n} + \frac{|y_{ij,n}|^2}{r_{ij,n}}\right), \tag{3.45}$$

$$-\log \mathcal{L}(\boldsymbol{W}, \boldsymbol{R}) = \text{const.} - 2J \sum_i \log |\det \boldsymbol{W}_i| + \sum_{i,j,n} \left(\log r_{ij,n} + \frac{|y_{ij,n}|^2}{r_{ij,n}}\right) \tag{3.46}$$

$$\equiv \mathcal{L}_{\mathrm{ILRMA}}. \tag{3.47}$$

Here, I consider two types of $r_{ij,n}$ decomposition depending on the presence of a partitioning function:

$$r_{ij,n} = \sum_l t_{il,n} v_{lj,n}, \tag{3.48}$$

$$r_{ij,n} = \sum_k z_{nk} t_{ik} v_{kj}, \tag{3.49}$$

where $t_{il,n}$ and $v_{lj,n}$ are the nonnegative entries of $\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I \times L}$ and $\boldsymbol{V}_n \in \mathbb{R}_{\geq 0}^{I \times L}$ that are the sourcewise basis and activation matrices, and t_{ik} and v_{kj} are the nonnegative entries of $\boldsymbol{T}$ and $\boldsymbol{V}$ that include K bases and activations, respectively. Moreover, $z_{nk} \in [0, 1]$ is the entry of $\boldsymbol{Z} = (\boldsymbol{z}_1 \ \cdots \ \boldsymbol{z}_N)^{\mathrm{T}} \in \mathbb{R}_{\{0,1\}}^{N \times K}$, which is a partitioning function

that clusters K bases into N sources and satisfies $\sum_n z_{nk} = 1$, and $k = 1, \cdots, K$ is the new basis index. In (3.48), a fixed number of bases, L, is utilized to decompose each separated source spectrogram $|y_{ij,n}|^2$. On the other hand, we can adaptively determine the number of bases for each separated source spectrogram by employing the partitioning function z_{nk} as (3.49). In this model, we only set the total number of bases to K. This approach is reasonable because the optimal number of bases will depend on the time-frequency structure of each source. For a source that consists of a low-rank power spectrogram, such as an instrumental signal, the number of bases should be small, whereas a speech or vocal spectrogram may require more bases for its precise representation. The cost function in ILRMA can be obtained by substituting (3.48) or (3.49) into (3.46).

In Laplace IVA, the variance $r_{i,n}$ is uniformly set to unity over the frequency bins, and is not estimated. This is because the variance only determines the signal scale of $y_{ij,n}$, and it can be restored by the back-projection technique. In time-varying Gaussian IVA, only the activation for the uniform variance is estimated based on the prior information given by users. On the other hand, the variance in ILRMA, $r_{ij,n}$, is blindly estimated by low-rank decomposition using NMF (3.48) or (3.49) to capture the time-frequency structure as shown in Fig. 3.6 (b). It is clear that when the number of bases is set to one for every source and all bases have a flat spectrum, the source models in time-varying Gaussian IVA and ILRMA become identical. This fact shows that ILRMA includes time-varying Gaussian IVA as a special case, which will be discussed in Sect. 3.5.

3.3.3 Update Rules

For the optimization of ICA or IVA, IP-based update rules, which can be derived using the auxiliary function technique, have been proposed [176, 184, 185, 186, 104, 192], and it has been reported that these update rules are faster and more stable than those for a conventional update scheme (e.g., natural gradient method [175, 33]) and that the step size parameter can be omitted in each iteration. Regarding the estimation of $\boldsymbol{W}_i$, the differential of (3.46) w.r.t. $\boldsymbol{W}_i$ becomes equivalent to that of the auxiliary bounding function in Laplace IVA [184]. For this reason, the update rules of $\boldsymbol{W}_i$ based on IP can easily be

derived as follows:

$$\boldsymbol{\mathcal{V}}_{i,n} = \frac{1}{J}\sum_{j}\frac{1}{r_{ij,n}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}}, \tag{3.50}$$

$$\boldsymbol{w}_{i,n} \leftarrow \left(\boldsymbol{W}_i\boldsymbol{\mathcal{V}}_{i,n}\right)^{-1}\boldsymbol{e}_n, \tag{3.51}$$

$$\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n}\left(\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{\mathcal{V}}_{i,n}\boldsymbol{w}_{i,n}\right)^{-\frac{1}{2}}. \tag{3.52}$$

After the update of $\boldsymbol{W}_i$, the separated signal $\boldsymbol{y}_{ij}$ should be updated as

$$y_{ij,n} \leftarrow \boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{ij}. \tag{3.53}$$

If we eliminate the partitioning function z_{nk}, which is ILRMA with (3.48), the differential of (3.46) w.r.t. $t_{il,n}$ or $v_{lj,n}$ becomes identical to the differential of the cost function in ISNMF (3.37). Therefore, the update rules of $t_{il,n}$ and $v_{lj,n}$ are given as

$$t_{il,n} \leftarrow t_{il,n}\sqrt{\frac{\sum_j |y_{ij,n}|^2 v_{lj,n} r_{ij,n}^{-2}}{\sum_j v_{lj,n} r_{ij,n}^{-1}}}, \tag{3.54}$$

$$v_{lj,n} \leftarrow v_{lj,n}\sqrt{\frac{\sum_i |y_{ij,n}|^2 t_{il,n} r_{ij,n}^{-2}}{\sum_i t_{il,n} r_{ij,n}^{-1}}}. \tag{3.55}$$

The estimated source model $r_{ij,n}$ should be updated by (3.48) after each update of $t_{il,n}$ and $v_{lj,n}$.

Alternatively, if we employ the partitioning function z_{nk} to cluster K bases into N specific sources, which is ILRMA with (3.49), we can derive the auxiliary-function-based update rules of z_{nk}, t_{ik}, and v_{kj} by minimizing (3.46) in a similar way to in [12, 167].

Here, I design an upper bound function of (3.46) as the auxiliary function. The first term in (3.46) is a convex function for the variables. Applying Jensen's inequality to this term with an auxiliary variable $\alpha_{ijk,n} \geq 0$ that satisfies

$\sum_k \alpha_{ijk,n} = 1$, we have

$$\frac{1}{\sum_k z_{nk} t_{ik} v_{kj}} \leq \sum_k \frac{\alpha_{ijk,n}^2}{z_{nk} t_{ik} v_{kj}}. \tag{3.56}$$

Also, the third term in (3.46) is a concave function, and we can apply the tangent line inequality to this term with an auxiliary variable $\beta_{ij,n} \geq 0$ as

$$\log \sum_k z_{nk} t_{ik} v_{kj} \leq \frac{1}{\beta_{ij,n}} \left(\sum_k z_{nk} t_{ik} v_{kj} - \beta_{ij,n} \right) + \log \beta_{ij,n}. \tag{3.57}$$

The equality of (3.56) and (3.57) holds if and only if the auxiliary variables are set as follows:

$$\alpha_{ijk,n} = \frac{z_{nk} t_{ik} v_{kj}}{\sum_{k'} z_{nk'} t_{ik'} v_{k'j}}, \tag{3.58}$$

$$\beta_{ij,n} = \sum_k z_{nk} t_{ik} v_{kj}. \tag{3.59}$$

Using these upper bounds, we can design the auxiliary function of (3.46) as

$$\begin{aligned} \mathcal{L}_{\mathrm{ILRMA}} \leq \mathcal{L}^+_{\mathrm{ILRMA}} = \sum_{i,j} \Bigg[& \sum_{n,k} \frac{|y_{ij,n}|^2 \alpha_{ijk,n}^2}{z_{nk} t_{ik} v_{kj}} - 2 \log |\det \boldsymbol{W}_i| \\ & + \frac{1}{\beta_{ij,n}} \left(\sum_k z_{nk} t_{ik} v_{kj} - \beta_{ij,n} \right) + \log \beta_{ij,n} \Bigg]. \end{aligned} \tag{3.60}$$

The update rules for $\mathcal{L}^+_{\mathrm{ILRMA}}$ with respect to each variable are determined by setting the gradient to zero. From $\partial \mathcal{L}^+_{\mathrm{ILRMA}} / \partial z_{nk} = 0$, we obtain

$$\sum_{i,j} \left[-\frac{|y_{ij,n}|^2 \alpha_{ijk,n}^2}{z_{nk}^2 t_{ik} v_{kj}} + \frac{1}{\beta_{ij,n}} t_{ik} v_{kj} \right] = 0. \tag{3.61}$$

By transposing the first term in (3.61) to the right-hand side and multiplying

both sides by z_{nk}^2, we have

$$z_{nk}^2 \sum_{i,j} \frac{1}{\beta_{ij,n}} t_{ik} v_{kj} = \sum_{i,j} \frac{|y_{ij,n}|^2 \alpha_{ijk,n}^2}{t_{ik} v_{kj}}. \tag{3.62}$$

Finally, the MU rule of z_{nk} can be derived by substituting (3.58) and (3.59) into (3.62) as follows:

$$z_{nk} \leftarrow z_{nk} \sqrt{\frac{\sum_{i,j} |y_{ij,n}|^2 t_{ik} v_{kj} r_{ij,n}^{-2}}{\sum_{i,j} t_{ik} v_{kj} r_{ij,n}^{-1}}}, \tag{3.63}$$

$$z_{nk} \leftarrow \frac{z_{nk}}{\sum_{n'} z_{n'k}}, \tag{3.64}$$

where (3.64) is calculated to ensure $\sum_n z_{nk} = 1$. Similarly to (3.63), the update rules of t_{ik} and v_{kj} are obtained as

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_{j,n} |y_{ij,n}|^2 z_{nk} v_{kj} r_{ij,n}^{-2}}{\sum_{j,n} z_{nk} v_{kj} r_{ij,n}^{-1}}}, \tag{3.65}$$

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_{i,n} |y_{ij,n}|^2 z_{nk} t_{ik} r_{ij,n}^{-2}}{\sum_{i,n} z_{nk} t_{ik} r_{ij,n}^{-1}}}, \tag{3.66}$$

The estimated source model $r_{ij,n}$ should be updated by (3.49) after each update of z_{nk}, t_{ik}, and v_{kj}.

Thus, we can estimate all the variables that minimize (3.46) by iterating these update rules. Note that a scale ambiguity exists between $\boldsymbol{W}_i$ and $r_{ij,n}$ because both of them can determine the scale of the separated signal $y_{ij,n}$. Therefore, $\boldsymbol{W}_i$ or $r_{ij,n}$ has a risk of diverging during the optimization. To avoid this problem, the following normalization should be applied at each iteration:

$$\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n} \lambda_n^{-1}, \tag{3.67}$$

$$y_{ij,n} \leftarrow y_{ij,n} \lambda_n^{-1}, \tag{3.68}$$

$$r_{ij,n} \leftarrow r_{ij,n} \lambda_n^{-2}, \tag{3.69}$$

and

$$t_{il,n} \leftarrow t_{il,n}\lambda_n^{-2}, \tag{3.70}$$

should be applied for ILRMA without a partitioning function, or

$$t_{ik} \leftarrow t_{ik}\sum_n z_{nk}\lambda_n^{-2}, \tag{3.71}$$

$$z_{nk} \leftarrow \frac{z_{nk}\lambda_n^{-2}}{\sum_{n'} z_{n'k}\lambda_{n'}^{-2}}, \tag{3.72}$$

should be applied for ILRMA with a partitioning function, where λ_n is an arbitrary sourcewise normalization coefficient, such as the sourcewise average power

$$\lambda_n = \sqrt{\frac{1}{IJ}\sum_{i,j}|y_{ij,n}|^2}. \tag{3.73}$$

These normalizations do not change the value of the cost function (3.46). The scale of the separated signal $y_{ij,n}$ can be restored by applying the back-projection technique (3.14) after the optimization.

3.3.4 Summary of Algorithm

The detailed algorithm of ILRMA is summarized in Algorithms 1 and 2, where $\boldsymbol{P}_n \in \mathbb{R}_{\geq 0}^{I\times J}$ is the power spectrogram whose entry is $p_{ij,n}$, $\max(\cdot,\cdot)$ returns a matrix with the larger elements taken from two inputs in each entry, ε denotes the machine epsilon, $\mathbf{1}^{(\text{size})}$ denotes matrix of ones whose size is denoted as the superscript, and $\mathsf{P} \in \mathbb{R}_{\geq 0}^{I\times J\times N}$ and $\mathsf{R} \in \mathbb{R}_{\geq 0}^{I\times J\times N}$ are third-order tensors whose entries are $p_{ij,n}$ and $r_{ij,n}$, respectively. To avoid division by zero, flooring with the machine epsilon is performed in the update of the NMF variables.

General NMF-based source separation techniques directly use the decomposed components to reconstruct the estimated signals or to calculate a Wiener filter. Since the NMF decomposition is a nonlinear approximation and the additivity of power spectrograms are generally invalid, this separation mecha-

nism often causes artificial distortion in the estimated signals and deteriorates the sound quality. In ILRMA, unlike such NMF-based source separation, the spectrogram decomposition is utilized only for the estimation of a latent source model $r_{ij.n}$, and the source separation is carried out by the linear spatial demixing filter $\boldsymbol{w}_{i,n}$ resulting in less distorted estimated signals. This issue will be confirmed by a subjective comparison of ILRMA and the other methods in Sect. 3.6.3.

Algorithm 1: Algorithm for ILRMA without partitioning function

Initialize $\boldsymbol{W}_i$ with identity matrix and $\boldsymbol{T}_n$ and $\boldsymbol{V}_n$ with nonnegative random values;
Calculate $\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}$ for all i and j;
Calculate $\mathsf{P}_{::n} = |\mathsf{Y}_{::n}|^{\cdot 2}$ and $\mathsf{R}_{::n} = \boldsymbol{T}_n \boldsymbol{V}_n$ for all n, respectively;
repeat
  for $n = 1$ **to** N **do**
    $\boldsymbol{T}_n \leftarrow \max\left(\boldsymbol{T}_n \circ \left[\frac{(\mathsf{P}_{::n} \circ \mathsf{R}_{::n}^{\cdot -2})\boldsymbol{V}_n^{\mathrm{T}}}{\mathsf{R}_{::n}^{-1}\boldsymbol{V}_n^{\mathrm{T}}}\right]^{\cdot \frac{1}{2}}, \varepsilon\right)$;
    $\mathsf{R}_{::n} = \boldsymbol{T}_n \boldsymbol{V}_n$;
    $\boldsymbol{V}_n \leftarrow \max\left(\boldsymbol{V}_n \circ \left[\frac{\boldsymbol{T}_n^{\mathrm{T}}(\mathsf{P}_{::n} \circ \mathsf{R}_{::n}^{\cdot -2})}{\boldsymbol{T}_n^{\mathrm{T}}\mathsf{R}_{::n}^{-1}}\right]^{\cdot \frac{1}{2}}, \varepsilon\right)$;
    $\mathsf{R}_{::n} = \boldsymbol{T}_n \boldsymbol{V}_n$;
    for $i = 1$ **to** I **do**
      $\boldsymbol{\mathcal{U}}_{i,n} = \frac{1}{J}\left\{\mathsf{X}_{i::}^{\mathrm{H}}\left[\mathsf{X}_{i::} \circ \left(\mathsf{R}_{i:n}^{-1}\mathbf{1}^{(1\times M)}\right)\right]\right\}$;
      $\boldsymbol{w}_{i,n} \leftarrow (\boldsymbol{W}_i \boldsymbol{\mathcal{U}}_{i,n})^{-1}\boldsymbol{e}_n$;
      $\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n}(\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{\mathcal{U}}_{i,n}\boldsymbol{w}_{i,n})^{-\frac{1}{2}}$;
    end
  end
  Calculate $\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}$ for all i and j;
  Calculate $\mathsf{P}_{::n} = |\mathsf{Y}_{::n}|^{\cdot 2}$ for all n;
  for $n = 1$ **to** N **do**
    $\lambda_n = \sqrt{\frac{1}{IJ}\sum_{i,j} p_{ij,n}}$;
    for $i = 1$ **to** I **do**
      $\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n}\lambda_n^{-1}$;
    end
    $\mathsf{P}_{::n} \leftarrow \mathsf{P}_{::n}\lambda_n^{-2}$;
    $\mathsf{R}_{::n} \leftarrow \mathsf{R}_{::n}\lambda_n^{-2}$;
    $\boldsymbol{T}_n \leftarrow \boldsymbol{T}_n\lambda_n^{-2}$;
  end
until *converge*;
Calculate $\hat{\boldsymbol{y}}_{ij,n} = \boldsymbol{W}_i^{-1}\left(\boldsymbol{e}_n \circ \boldsymbol{y}_{ij}\right)$ for all i, j, and n;

Algorithm 2: Algorithm for ILRMA with partitioning function

Initialize $\boldsymbol{W}_i$ with identity matrix, $\boldsymbol{T}$ and $\boldsymbol{V}$ with nonnegative random values, and $\boldsymbol{Z}$ with random values in range $[0, 1]$;
$\boldsymbol{Z} \leftarrow \boldsymbol{Z} \circ \left(\boldsymbol{1}^{(N\times N)}\boldsymbol{Z}\right)^{\cdot -1}$;
Calculate $\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}$ for all i and j;
Calculate $\mathsf{P}_{::n} = |\mathsf{Y}_{::n}|^{\cdot 2}$ and $\mathsf{R}_{::n} = \left[\left(\boldsymbol{1}^{(I\times 1)}\boldsymbol{z}_n^{\mathrm{T}}\right) \circ \boldsymbol{T}\right] \boldsymbol{V}$ for all n, respectively;
repeat
  for $n = 1$ **to** N **do**
    $\boldsymbol{b}_n^{(\boldsymbol{Z})} = \left(\frac{\left\{\left[\boldsymbol{T}^{\mathrm{T}}\left(\mathsf{P}_{::n} \circ \mathsf{R}_{::n}^{\cdot -2}\right)\right] \circ \boldsymbol{V}\right\}\boldsymbol{1}^{(J\times 1)}}{\left[\left(\boldsymbol{T}^{\mathrm{T}}\mathsf{R}_{::n}^{\cdot -1}\right) \circ \boldsymbol{V}\right]\boldsymbol{1}^{(J\times 1)}}\right)^{\cdot \frac{1}{2}}$;
  end
  $\boldsymbol{Z} \leftarrow \max\left(\boldsymbol{Z} \circ \boldsymbol{B}^{(\boldsymbol{Z})}, \varepsilon\right)$, where $\boldsymbol{B}^{(\boldsymbol{Z})} = (\boldsymbol{b}_1^{(\boldsymbol{Z})} \ \cdots \ \boldsymbol{b}_N^{(\boldsymbol{Z})})^{\mathrm{T}}$;
  $\boldsymbol{Z} \leftarrow \boldsymbol{Z} \circ \left(\boldsymbol{1}^{(N\times N)}\boldsymbol{Z}\right)^{\cdot -1}$;
  Calculate $\mathsf{R}_{::n} = \left[\left(\boldsymbol{1}^{(I\times 1)}\boldsymbol{z}_n^{\mathrm{T}}\right) \circ \boldsymbol{T}\right] \boldsymbol{V}$ for all n;
  for $i = 1$ **to** I **do**
    $\boldsymbol{b}_i^{(\boldsymbol{T})} = \left(\frac{\left\{\left[\boldsymbol{V}\left(\mathsf{P}_{i::} \circ \mathsf{R}_{i::}^{\cdot -2}\right)\right] \circ \boldsymbol{Z}^{\mathrm{T}}\right\}\boldsymbol{1}^{(N\times 1)}}{\left[\left(\boldsymbol{V}\mathsf{R}_{i::}^{\cdot -1}\right) \circ \boldsymbol{Z}^{\mathrm{T}}\right]\boldsymbol{1}^{(N\times 1)}}\right)^{\cdot \frac{1}{2}}$;
  end
  $\boldsymbol{T} \leftarrow \max\left(\boldsymbol{T} \circ \boldsymbol{B}^{(\boldsymbol{T})}, \varepsilon\right)$, where $\boldsymbol{B}^{(\boldsymbol{T})} = (\boldsymbol{b}_1^{(\boldsymbol{T})} \ \cdots \ \boldsymbol{b}_I^{(\boldsymbol{T})})^{\mathrm{T}}$;
  Calculate $\mathsf{R}_{::n} = \left[\left(\boldsymbol{1}^{(I\times 1)}\boldsymbol{z}_n^{\mathrm{T}}\right) \circ \boldsymbol{T}\right] \boldsymbol{V}$ for all n;
  for $j = 1$ **to** J **do**
    $\boldsymbol{b}_j^{(\boldsymbol{V})} = \left(\frac{\left\{\left[\boldsymbol{T}^{\mathrm{T}}\left(\mathsf{P}_{:j:} \circ \mathsf{R}_{:j:}^{\cdot -2}\right)\right] \circ \boldsymbol{Z}^{\mathrm{T}}\right\}\boldsymbol{1}^{(N\times 1)}}{\left[\left(\boldsymbol{T}^{\mathrm{T}}\mathsf{R}_{:j:}^{\cdot -1}\right) \circ \boldsymbol{Z}^{\mathrm{T}}\right]\boldsymbol{1}^{(N\times 1)}}\right)^{\cdot \frac{1}{2}}$;
  end
  $\boldsymbol{V} \leftarrow \max\left(\boldsymbol{V} \circ \boldsymbol{B}^{(\boldsymbol{V})}, \varepsilon\right)$, where $\boldsymbol{B}^{(\boldsymbol{V})} = (\boldsymbol{b}_1^{(\boldsymbol{V})} \ \cdots \ \boldsymbol{b}_J^{(\boldsymbol{V})})$;
  Calculate $\mathsf{R}_{::n} = \left[\left(\boldsymbol{1}^{(I\times 1)}\boldsymbol{z}_n^{\mathrm{T}}\right) \circ \boldsymbol{T}\right] \boldsymbol{V}$ for all n;
  for $n = 1$ **to** N **do**
    for $i = 1$ **to** I **do**
      $\boldsymbol{\mathcal{U}}_{i,n} = \frac{1}{J}\left\{\mathsf{X}_{i::}^{\mathrm{H}}\left[\mathsf{X}_{i::} \circ \left(\mathsf{R}_{i:n}^{-1}\boldsymbol{1}^{(1\times M)}\right)\right]\right\}$;
      $\boldsymbol{w}_{i,n} \leftarrow (\boldsymbol{W}_i\boldsymbol{\mathcal{U}}_{i,n})^{-1}\boldsymbol{e}_n$;
      $\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n}(\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{\mathcal{U}}_{i,n}\boldsymbol{w}_{i,n})^{-\frac{1}{2}}$;
    end
  end
  Calculate $\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}$ for all i and j;
  Calculate $\mathsf{P}_{::n} = |\mathsf{Y}_{::n}|^{\cdot 2}$ for all n;
  for $n = 1$ **to** N **do**
    $\lambda_n = \sqrt{\frac{1}{IJ}\sum_{i,j} p_{ij,n}}$;
    for $i = 1$ **to** I **do**
      $\boldsymbol{w}_{i,n} \leftarrow \boldsymbol{w}_{i,n}\lambda_n^{-1}$;
    end
    $\mathsf{P}_{::n} \leftarrow \mathsf{P}_{::n}\lambda_n^{-2}$;
    $\mathsf{R}_{::n} \leftarrow \mathsf{R}_{::n}\lambda_n^{-2}$;
  end
  Calculate $t_{ik} \leftarrow t_{ik}\sum_n z_{nk}\lambda_n^{-2}$ for all i and k;
  Calculate $z_{nk} \leftarrow z_{nk}\frac{\lambda_n^{-2}}{\sum_{n'} z_{n'k}\lambda_{n'}^{-2}}$ for all n and k;
until *converge*;
Calculate $\hat{\boldsymbol{y}}_{ij,n} = \boldsymbol{W}_i^{-1}\left(\boldsymbol{e}_n \circ \boldsymbol{y}_{ij}\right)$ for all i, j, and n;

3.4 Relationship between IVA, MNMF, and ILRMA

In NMF-based source separation, the decomposed bases and activations must be clustered in every source to achieve source separation. To solve this problem, MNMF has been proposed [77, 78, 80, 6, 81, 82, 83, 84]. In particular, MNMF methods [6, 81, 82, 84] treat convolutive mixtures similarly to FDICA, IVA, and ILRMA and estimate a mixing system for the sources, which is utilized for the clustering of bases. In these MNMFs, the spatial covariance [195, 196], which is the covariance matrix of a zero-mean multivariate Gaussian distribution, has been utilized to model the mixing conditions of the recording environment. In this section, the relationships between time-varying Gaussian IVA, ILRMA, and MNMF are revealed from the viewpoint of their assumed generative models.

3.4.1 Generative Model in MNMF and Spatial Covariance

In MNMF [6, 81, 82, 84] and its related methods [195, 196], the probability distribution of multichannel STFT coefficients $\boldsymbol{x}_{ij}$ is modeled by a circularly symmetric multivariate complex Gaussian distribution with a time-frequency-variant covariance matrix as follows:

$$p(\boldsymbol{x}_{ij}) = \frac{1}{\pi^M \det \boldsymbol{R}_{ij}^{(\boldsymbol{x})}} \exp\left(-\boldsymbol{x}_{ij}^{\mathrm{H}} \boldsymbol{R}_{ij}^{(\boldsymbol{x})^{-1}} \boldsymbol{x}_{ij}\right), \tag{3.74}$$

where $\boldsymbol{R}_{ij}^{(\boldsymbol{x})}$ is called the spatial covariance [195, 196] of the observed multichannel signal $\boldsymbol{x}_{ij}$, namely, $\boldsymbol{R}_{ij}^{(\boldsymbol{x})} = \mathrm{E}[\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{H}}]$. This spatial covariance can be decomposed into the time-invariant source covariance $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$, the time-variant scalar variance $r_{ij,n}$, and the time-invariant noise covariance $\boldsymbol{R}_{i}^{(\boldsymbol{n})}$ that contributes to additional noise $\boldsymbol{n}_{ij}$, as

$$\boldsymbol{R}_{ij}^{(\boldsymbol{x})} = \sum_n r_{ij,n} \boldsymbol{R}_{i,n}^{(\boldsymbol{s})} + \boldsymbol{R}_{i}^{(\boldsymbol{n})}. \tag{3.75}$$

The spatial covariance $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$ represents the spatial position and the spatial spread of the nth source. In particular, if the mixing system can be modeled by the mixing matrix $\boldsymbol{A}_i$ as (2.7) with a noiseless assumption, the spatial covariance $\boldsymbol{R}_{i,n}^{(\boldsymbol{s})}$

is equal to the rank-1 matrix

$$\boldsymbol{R}_{i,n}^{(s)} = \boldsymbol{a}_{i,n}\boldsymbol{a}_{i,n}^{\mathrm{H}}. \tag{3.76}$$

This mixing model is the called *rank-1 spatial model*, which is identical to the assumption of an instantaneous mixture in the frequency domain. In contrast, if the mixing system cannot be modeled by (2.7) owing to, for example, strong reverberation in the recording environment, the rank of $\boldsymbol{R}_{i,n}^{(s)}$ increases so that it becomes a full-rank spatial covariance [195, 196].

3.4.2 Existing MNMF Models

Existing MNMF models and their related works can be characterized in terms of two features: models of spatial covariance $\boldsymbol{R}_{ij}^{(x)}$ and source spectrograms. Table 3.1 summarizes the existing methods. The models proposed in [195, 196] have the most general representations. Several types of $\boldsymbol{R}_{i,n}^{(s)}$ have been investigated including rank-1 and full-rank matrices. MNMF in [6] (hereafter referred to as *Ozerov's MNMF*) was the first method to model a power spectrogram $r_{ij,n}$ using NMF decomposition. In this method, the sourcewise spatial covariance $\boldsymbol{R}_{i,n}^{(s)}$ is constrained by a rank-1 matrix, and an additive noise component $\boldsymbol{n}_{ij}$ is also assumed. The update rules of the variables based on both EM and MU algorithms have been derived. Ozerov's MNMF was extended to a full-rank spatial model in [81]. Also, a more flexible source model with a partitioning function z_{nk} was introduced in [82]. As another optimization scheme, an MU algorithm based on an auxiliary function technique was proposed in [84] (hereafter referred to as *Sawada's MNMF*). It also employs the full-rank $\boldsymbol{R}_{i,n}^{(s)}$ and the flexible source model with z_{nk} and NMF variables. Note that all the existing MNMFs estimate the sourcewise mixing system $\boldsymbol{R}_{i,n}^{(s)}$ to achieve separation via multichannel Wiener filtering (MWF) [197], whereas ILRMA estimates the demixing matrix $\boldsymbol{W}_i$.

Table 3.1: Models of mixing system, spatial covariance, power spectrogram, and their optimization in each method

Literature	Model of $\boldsymbol{R}_{ij}^{(x)}$	Spatial covariance	Power spectrogram	Optimization
Ozerov and Févotte [6]	$\sum_{n,l} t_{il,n} v_{lj,n} \boldsymbol{R}_{i,n}^{(s)} + \boldsymbol{R}_i^{(n)}$ $(\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} + \boldsymbol{n}_{ij})$	Rank-1 matrix $\boldsymbol{R}_{i,n}^{(s)}$ and diagonal matrix $\boldsymbol{R}_i^{(n)}$	NMF w/o partitioning function	EM and MU for $\boldsymbol{A}_i$, $\boldsymbol{R}_i^{(n)}$, $\boldsymbol{T}_n$, and $\boldsymbol{V}_n$
Arberet et al. [81]	$\sum_{n,l} t_{il,n} v_{lj,n} \boldsymbol{R}_{i,n}^{(s)} + \boldsymbol{R}_i^{(n)}$	Full-rank matrix $\boldsymbol{R}_{i,n}^{(s)}$ and diagonal matrix $\boldsymbol{R}_i^{(n)}$	NMF w/o partitioning function	EM for $\boldsymbol{R}_{i,n}^{(s)}$, $\boldsymbol{R}_i^{(n)}$, $\boldsymbol{T}_n$, and $\boldsymbol{V}_n$
Duong et al. [196]	$\sum_n r_{ij,n} \boldsymbol{R}_{i,n}$	Several types of $\boldsymbol{R}_{i,n}^{(s)}$ including rank-1 and full-rank matrices	$r_{ij,n}$ (w/o NMF)	EM for $\boldsymbol{R}_{i,n}^{(s)}$
Ozerov et al. [82]	$\sum_n \boldsymbol{R}_{i,n}^{(s)} \sum_k z_{nk} t_{ik} v_{kj} + \boldsymbol{R}_i^{(n)}$ $(\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} + \boldsymbol{n}_{ij})$	Rank-1 matrix $\boldsymbol{R}_{i,n}^{(s)}$ and diagonal matrix $\boldsymbol{R}_i^{(n)}$	NMF with partitioning function	EM and MU for $\boldsymbol{A}_i$, $\boldsymbol{R}_i^{(n)}$, $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$
Sawada et al. [84]	$\sum_n \boldsymbol{R}_{i,n}^{(s)} \sum_k z_{nk} t_{ik} v_{kj}$	Full-rank matrix $\boldsymbol{R}_{i,n}^{(s)}$	NMF with partitioning function	MU for $\boldsymbol{R}_{i,n}^{(s)}$, $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$
Kitamura et al. [194]	$\sum_n \boldsymbol{R}_{i,n}^{(s)} \sum_k z_{nk} t_{ik} v_{kj}$ $(\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij})$	Rank-1 matrix $\boldsymbol{R}_{i,n}^{(s)}$	NMF with partitioning function	IP for $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1}$ and MU for $\boldsymbol{Z}$, $\boldsymbol{T}$, and $\boldsymbol{V}$

3.4.3 Equivalence between ILRMA and MNMF with Rank-1 Spatial Model

From (3.74), the likelihood function of the observed spatial covariance $\boldsymbol{R}^{(x)} = \{\boldsymbol{R}_{ij}^{(x)} | i = 1, \cdots, I; j = 1, \cdots, J\}$ is given as

$$\mathcal{L}(\boldsymbol{R}^{(x)}) = \prod_{i,j} p(\boldsymbol{x}_{ij} | \boldsymbol{R}_{ij}^{(x)}) \tag{3.77}$$

$$= \prod_{i,j} \frac{1}{\pi^M \det \boldsymbol{R}_{ij}^{(x)}} \exp\left(-\boldsymbol{x}_{ij}^{\mathrm{H}} \boldsymbol{R}_{ij}^{(x)^{-1}} \boldsymbol{x}_{ij}\right), \tag{3.78}$$

and the negative log-likelihood function is

$$-\log \mathcal{L}(\boldsymbol{R}^{(x)}) = \sum_{i,j} \left[M \log \pi + \log \det \boldsymbol{R}_{ij}^{(x)} + \boldsymbol{x}_{ij}^{\mathrm{H}} \boldsymbol{R}_{ij}^{(x)^{-1}} \boldsymbol{x}_{ij} \right] \tag{3.79}$$

$$= \text{const.} + \sum_{i,j} \left[\log \det \boldsymbol{R}_{ij}^{(x)} + \mathrm{tr}\left(\boldsymbol{X}_{ij} \boldsymbol{R}_{ij}^{(x)^{-1}} \right) \right], \tag{3.80}$$

where $\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{H}}$ is an observed instantaneous covariance matrix. Similar to ISNMF in Sect. 3.2.3, the ML estimation based on (3.80) is identical to the multichannel IS divergence $\mathcal{D}_{\mathrm{MIS}}$ [84], which is known as Stein's loss [198] in the statistics field or the log-determinant divergence [199] in the machine learning

field:

$$\sum_{i,j} \mathcal{D}_{\mathrm{MIS}}(\boldsymbol{X}_{ij} \| \boldsymbol{R}_{ij}^{(\boldsymbol{x})}) = \sum_{i,j} \left[\mathrm{tr}\left(\boldsymbol{X}_{ij} \boldsymbol{R}_{ij}^{(\boldsymbol{x})^{-1}} \right) - \log\det \boldsymbol{X}_{ij} \boldsymbol{R}_{ij}^{(\boldsymbol{x})^{-1}} - M \right] \tag{3.81}$$

$$= \mathrm{const.} + \sum_{i,j} \left[\log\det \boldsymbol{R}_{ij}^{(\boldsymbol{x})} + \mathrm{tr}\left(\boldsymbol{X}_{ij} \boldsymbol{R}_{ij}^{(\boldsymbol{x})^{-1}} \right) \right]. \tag{3.82}$$

In FDICA, IVA, and ILRMA, the mixing model (2.7) with a noiseless assumption is used, which results in the rank-1 spatial model (3.76). On the basis of this assumption, the covariance matrix $\boldsymbol{R}_{ij}^{(\boldsymbol{x})}$ can be rewritten using the mixing matrix $\boldsymbol{A}_i$ as

$$\boldsymbol{R}_{ij}^{(\boldsymbol{x})} = \sum_n r_{ij,n} \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{H}} \tag{3.83}$$

$$= \boldsymbol{A}_i \boldsymbol{D}_{ij} \boldsymbol{A}_i^{\mathrm{H}}, \tag{3.84}$$

where

$$\boldsymbol{D}_{ij} = \begin{pmatrix} r_{ij,1} & 0 & \cdots & 0 \\ 0 & r_{ij,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & r_{ij,N} \end{pmatrix}. \tag{3.85}$$

If we substitute (3.84) into the cost function in MNMF (3.80), we obtain

$$-\log \mathcal{L}(\boldsymbol{R}^{(x)}) = \text{const.} + \sum_{i,j} \left[\log\det \boldsymbol{A}_i \boldsymbol{D}_{ij} \boldsymbol{A}_i^{\mathrm{H}} + \mathrm{tr}\left(\boldsymbol{X}_{ij} \left(\boldsymbol{A}_i^{\mathrm{H}} \right)^{-1} \boldsymbol{D}_{ij}^{-1} \boldsymbol{A}_i^{-1} \right) \right] \tag{3.86}$$

$$= \text{const.} + \sum_{i,j} \Big[\log(\det \boldsymbol{A}_i)(\det \boldsymbol{D}_{ij})(\det \boldsymbol{A}_i)^{\mathrm{H}} + \mathrm{tr}\left(\boldsymbol{W}_i^{-1} \boldsymbol{y}_{ij} \boldsymbol{y}_{ij}^{\mathrm{H}} \left(\boldsymbol{W}_i^{-1} \right)^{\mathrm{H}} \boldsymbol{W}_i^{\mathrm{H}} \boldsymbol{D}_{ij}^{-1} \boldsymbol{W}_i \right) \Big] \tag{3.87}$$

$$= \text{const.} + \sum_{i,j} \left[\log |\det \boldsymbol{A}_i|^2 + \log\det \boldsymbol{D}_{ij} + \mathrm{tr}\left(\boldsymbol{W}_i \boldsymbol{W}_i^{-1} \boldsymbol{y}_{ij} \boldsymbol{y}_{ij}^{\mathrm{H}} \boldsymbol{D}_{ij}^{-1} \right) \right] \tag{3.88}$$

$$= \text{const.} - 2J \sum_i \log |\det \boldsymbol{W}_i| + \sum_{i,j} \left[\log \prod_n r_{ij,n} + \mathrm{tr}\left(\boldsymbol{y}_{ij} \boldsymbol{y}_{ij}^{\mathrm{H}} \boldsymbol{D}_{ij}^{-1} \right) \right] \tag{3.89}$$

$$= \text{const.} - 2J \sum_i \log |\det \boldsymbol{W}_i| + \sum_{i,j,n} \left[\log r_{ij,n} + \frac{|y_{ij,n}|^2}{r_{ij,n}} \right], \tag{3.90}$$

where we used $\boldsymbol{x}_{ij} = \boldsymbol{W}_i^{-1} \boldsymbol{y}_{ij}$ and $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1}$ to transform the variables. Thus, it is revealed that the cost function in MNMF with the rank-1 spatial model is identical to (3.46), the cost function in ILRMA, because the same spatial and source models are assumed.

Figure 3.7 shows the relationship between IVA, ILRMA, and MNMF. MNMF with a rank-1 spatial model, which assumes an instantaneous mixture in the frequency domain, is essentially equivalent to ILRMA, which is IVA with a flexible source model using NMF decomposition. Therefore, ILRMA can be considered as an intermediate model between IVA and MNMF in terms of the model flexibility. From the IVA side, we introduced the source model using NMF with bases to capture the specific spectral patterns, and from the MNMF side, a rank-1 spatial model was introduced to transform the variable $\boldsymbol{A}_i$ into $\boldsymbol{W}_i$ and to make the optimization more efficient.

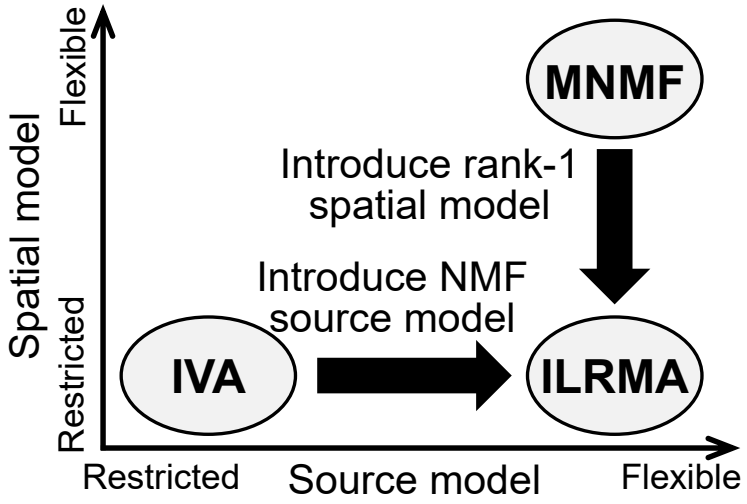

Figure 3.7: Relationship between IVA, ILRMA, and MNMF from viewpoint of flexibility of spatial and source models.

3.5 Experimental Analysis of ILRMA using Artificial Observation

In this section, I discuss the inherent difference between FDICA, Laplace IVA, and ILRMA. In addition, I evaluate them via BSS experiments using artificial sources and show that ILRMA possesses better flexibility than FDICA and Laplace IVA for both the source and spatial models.

3.5.1 Difference between Assumption in Source Model

In IVA, as already discussed in Sects. 3.2.2 and 3.2.4, we generally introduce the spherically symmetric multivariate distribution to ensure the higher-order correlation between frequency bins, where the stationary distribution is assumed in Laplace IVA and the time-varying variance is introduced in time-varying Gaussian IVA. As shown in Fig. 3.6 (a), all the frequency components are assumed to have the same activation (time-varying gain) in time-varying Gaussian IVA. This simple and nonflexible source model can be interpreted as a specific NMF that has only one frequency-uniform (flat) basis for each source. Therefore, the

number of bases of a model spectrogram in time-varying Gaussian IVA always becomes one.

On the other hand, conventional MNMF and proposed ILRMA independently assume circularly symmetric complex Gaussian distribution for each time-frequency slot [125] because their cost functions are based on IS divergence. Therefore, the estimated variances $r_{ij,n}$ can explicitly express a model spectrogram via NMF decomposition with an arbitrary number of bases (see the spectrogram in Fig. 3.6 (b)). For this reason, the source model in ILRMA is more flexible than that in IVA. In addition, time-varying Gaussian IVA can be thought of as a special case of ILRMA. If the number of bases for each source is set to one and the basis is fixed to a flat spectrum, both methods become essentially equivalent.

For conventional FDICA, its source model depends on how the permutation problem is solved. The permutation solver utilizing the correlation between frequency bins [179] is an essentially equivalent approach to IVA. However, the permutation solver based on DOA of each source [178] is a different approach. Hereafter, I refer to the combined method of FDICA and DOA-based permutation solver as FDICA+DOA. FDICA+DOA uses the estimated steering vectors (estimated spatial model), and there is no explicit source model except for non-Gaussianity in the time series for each frequency bin.

3.5.2 Difference between Assumption in Spatial Model

In IVA and ILRMA, there is no explicit assumption in the spatial model (mixing system) except for the rank-1 spatial model (3.76). Both methods only use the statistical independence between source models (model spectrograms) and the observed multichannel mixtures to estimate the demixing matrix.

In contrast, FDICA+DOA directly uses the difference between the estimated spatial conditions for each source to solve the permutation problem. Therefore, the separation performance of FDICA+DOA is sensitive to the spatial setup of the sources; if the positions of the sources become close or the reverberation becomes strong, the error of the permutation solver may increase. In summary, FDICA+DOA is severely affected by the spatial conditions rather than source modeling, whereas IVA and ILRMA are not.

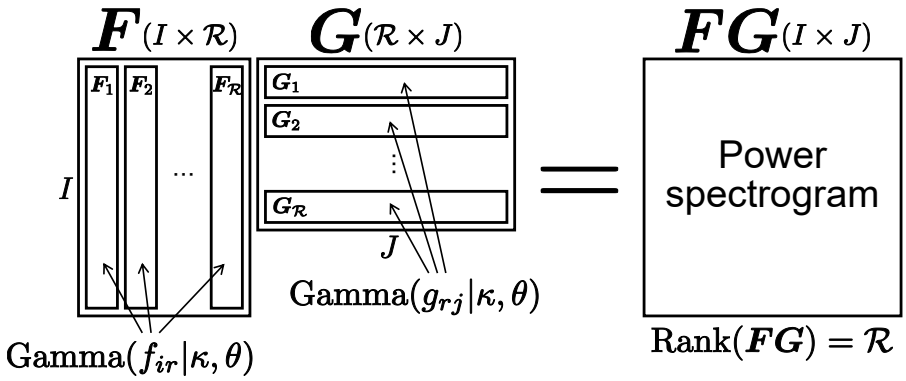

Figure 3.8: Artificial source that consists of R bases.

3.5.3 Experimental Validation

I validate the difference between both the source and spatial models among IVA, FDICA+DOA, and ILRMA. In this validation, for simplicity, the numbers of sources and microphones are set to two, namely, $N = M = 2$.

Design of Artificial Spectrograms with $\mathcal{R}$ Bases

From the difference between the source models of IVA and ILRMA, we can expect that the number of bases (rank) in the power spectrogram will affect the separation performance for IVA. If the power spectrogram of each source consists of only one basis, both IVA and ILRMA can separate the sources with high accuracy. However, if the sources have more complicated power spectrograms, the source model in IVA cannot represent them in principle, and the separation performance may decrease.

To investigate this issue, in this experiment, I produce artificial sources whose power spectrograms can be represented by $\mathcal{R}$ bases. Figure 3.8 shows the power spectrogram that I produced. To simulate a nonnegative sparse spectrogram, I generate nonnegative random values f_{ir} and g_{rj} that obey independent and identically distributed (i.i.d.) gamma distributions, where $r = 1, \cdots, \mathcal{R}$ is the integral index of the basis in matrices $\boldsymbol{F}$ and $\boldsymbol{G}$. The power spectrogram is a product of $\boldsymbol{F}$ and $\boldsymbol{G}$ and its size is $I \times J$. The gamma distribution can be

Table 3.2: Estimated values of shape parameter κ so that kurtosis of $\boldsymbol{FG}$ is adjusted to 50 for each $\mathcal{R}$

Number of bases $\mathcal{R}$	Shape parameter κ
1	0.83809
2	0.54962
3	0.43450
4	0.36929
5	0.32617
6	0.29504
7	0.27124
8	0.25231

represented as

$$\mathrm{Gamma}(\chi|\kappa,\theta) = \chi^{\kappa-1}\frac{1}{\Gamma(\kappa)\theta^{\kappa}}\exp\left(-\frac{\chi}{\theta}\right), \tag{3.91}$$

where χ is a random variable, and κ and θ are shape and scale parameters, respectively. After producing the power spectrogram $\boldsymbol{FG}$, I add random phases that obey a uniform distribution in the range $[0, 2\pi]$ to $\boldsymbol{FG}$, and the produced complex spectrogram ($\boldsymbol{FG}$ with random phases) is used as an artificial source whose power spectrogram has $\mathcal{R}$ bases. Therefore, in this procedure, I simulate the variances of zero-mean spherical complex Gaussian distributions with an outer product of variables that obey i.i.d. gamma distributions and their linear combination.

In this artificial source, it is important to set κ to an appropriate value. For example, when κ is set to a constant value regardless of $\mathcal{R}$, the random values in the power spectrogram $\boldsymbol{FG}$ become close to a Gaussian distribution.This is because the kurtosis of the element $\sum_{r=1}^{\mathcal{R}} f_{ir}g_{rj}$ in $\boldsymbol{FG}$ converges to three by the central limit theorem when $\mathcal{R}$ increases. For this reason, the separation accuracy of ICA-based methods decreases as $\mathcal{R}$ increases. To avoid this influence, I adjust the shape parameter κ for each value of $\mathcal{R}$ so that the kurtosis of $\boldsymbol{FG}$ is always the same value regardless of $\mathcal{R}$. Such a κ can be derived using the moment-cumulant transform [200] (see Appendix A). The following equation

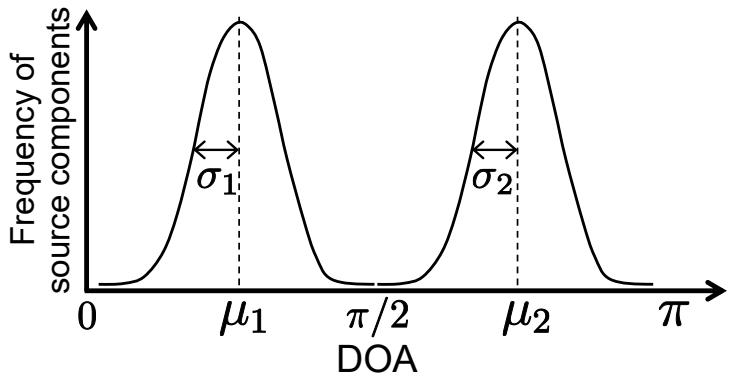

Figure 3.9: Artificial DOA with Gaussian distributions.

gives the shape parameter κ used to adjust the kurtosis of $\boldsymbol{FG}$:

$$\frac{\zeta(\kappa, \mathcal{R})}{\xi(\kappa, \mathcal{R})} - \text{kurt} = 0, \tag{3.92}$$

where kurt is the intended value for the kurtosis of $\boldsymbol{FG}$ and

$$\begin{aligned}\zeta(\kappa, \mathcal{R}) &= 84\kappa^3 + 174\kappa^2 + 132\kappa + 36 \\ &\quad + \mathcal{R}\left(52\kappa^4 + 60\kappa^3 + 19\kappa^2\right) + \mathcal{R}^2\left(12\kappa^5 + 6\kappa^4\right) + \mathcal{R}^3\kappa^6, \qquad (3.93)\\ \xi(\kappa, \mathcal{R}) &= \mathcal{R}\left(4\kappa^4 + 4\kappa^3 + \kappa^2\right) + \mathcal{R}^2\left(4\kappa^5 + 2\kappa^4\right) + \mathcal{R}^3\kappa^6. \qquad (3.94)\end{aligned}$$

Since no closed-form solution exists that satisfies (3.92), I calculate the optimal κ by a greedy search. Table 3.2 shows the estimated shape parameter values when kurt = 50. I experimentally confirmed that the kurtosis of the produced power spectrogram $\boldsymbol{FG}$ is always controlled to be approximately 50 using these shape parameters.

Design of Artificial Mixing Systems

For a mixing system, I designed an artificial DOA that consists of $N = 2$ Gaussian distributions, as shown in Fig. 3.9, where μ_n and σ_n are the mean value (position of the source) and standard deviation of the nth source, respectively. This

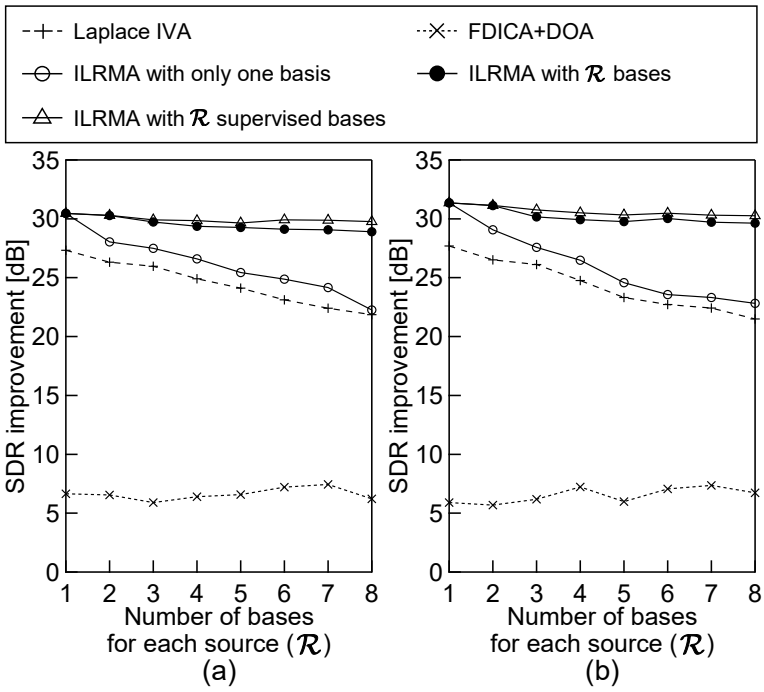

Figure 3.10: Separation results of (a) first source and (b) second source for various numbers of bases.

modeling mimics an actual acoustical phenomenon in which the DOAs of wavefronts in different frequencies i are randomly distributed owing to the room reverberation effect. I produced steering vectors $\boldsymbol{a}_{i,n}$ that obey the Gaussian distributions in Fig. 3.9 and prepared an artificial mixing matrix $\boldsymbol{A}_i$. Finally, I produced an artificial observed signal $\boldsymbol{x}_{ij}$ with artificial sources and an artificial mixing system using (2.7).

Experiment on Variational Artificial Spectrogram

In this experiment, I assume the following conditions: $I = J = 257$, kurt = 50, $\theta = 1$, $\mu_1 = 5\pi/12$, $\mu_2 = 7\pi/12$, $\sigma_1^2 = \sigma_2^2 = 0.05$, and the interelement spacing of microphones is set to 5.66 cm. Figure 3.10 shows the improvement of the signal-to-distortion ratio (SDR) [201] for various numbers of bases $\mathcal{R}$, where SDR indicates the total separation performance and the improvement in the SDR is the increment from the SDR value of the observed signal. For ILRMA, I

use a simple formulation without the partitioning function. Also, I evaluate three patterns, namely, the case of $L = 1$ (ILRMA with only one basis), the case of a suitable number of spectral bases $L = \mathcal{R}$ (ILRMA with $\mathcal{R}$ bases), and the case of a supervised source model by setting $\boldsymbol{T} = \boldsymbol{F}$ and $\boldsymbol{V} = \boldsymbol{G}$ (ILRMA with $\mathcal{R}$ supervised bases). From Fig. 3.10, the separation scores of Laplace IVA and ILRMA with only one basis decrease when the number of bases of each source, $\mathcal{R}$, increases because they cannot capture the exact power spectrograms. In contrast, ILRMA with $\mathcal{R}$ bases can maintain high SDR values because the power spectrogram of each source can be represented by a model spectrogram using $\mathcal{R}$ spectral bases in $\boldsymbol{T}$. This clearly demonstrates the flexibility of the source model in ILRMA.

Experiment on Variational Artificial Mixing Systems

From the difference between the spatial models in FDICA+DOA and ILRMA, we can expect that the mixing system (spatial conditions of each source) will affect the separation performance for FDICA+DOA. If the source positions are close or the variance of the DOAs is large, a large error of DOA clustering occurs in FDICA+DOA, resulting in marked degradation of the separation. However, since IVA and ILRMA do not use the explicit properties of the mixing condition (spatial model), we can expect that their separation performance will not strongly depend on the source positions or the variance of the DOAs. To investigate this issue, in this experiment, I produce observed signals with various mixing conditions and evaluate the separation performance. I use the artificial sources described in Sect. 3.5.3, where the power spectrograms of these sources are generated with kurt = 50 and $\mathcal{R} = 1$. The mixing system is produced by the artificial DOA shown in Fig. 3.9 with various μ_1, μ_2, σ_1^2, and σ_2^2. Note that the experiment in which σ_1^2 and σ_2^2 are changed does not simulate a change in the reverberation time. It only controls the variance of the DOAs over the frequencies, and the length of the impulse response does not change. Therefore, even when using larger σ_1^2 and σ_2^2, the rank-1 spatial model is always valid in this simulation. For ILRMA, the number of bases L is set to one, which is equal to $\mathcal{R}$. The other conditions are the same as those in Sect. 3.5.3.

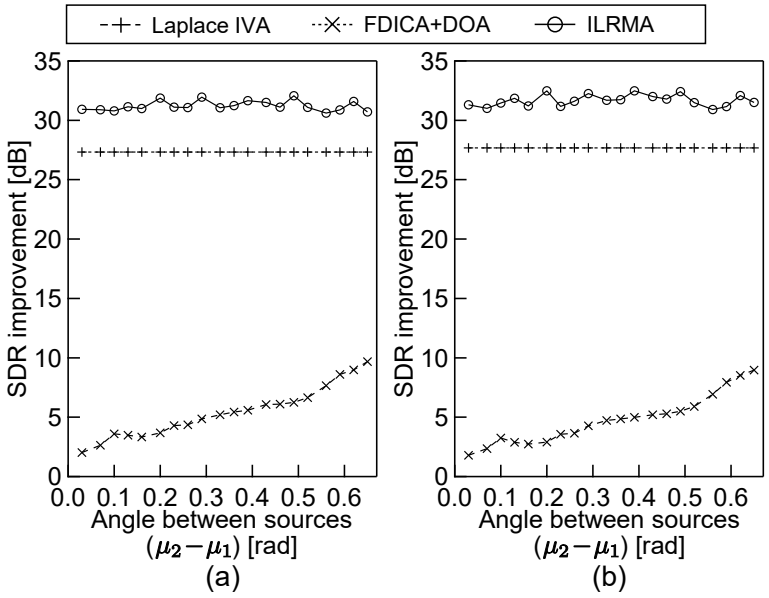

Figure 3.11: Separation results of (a) first source and (b) second source for various angles.

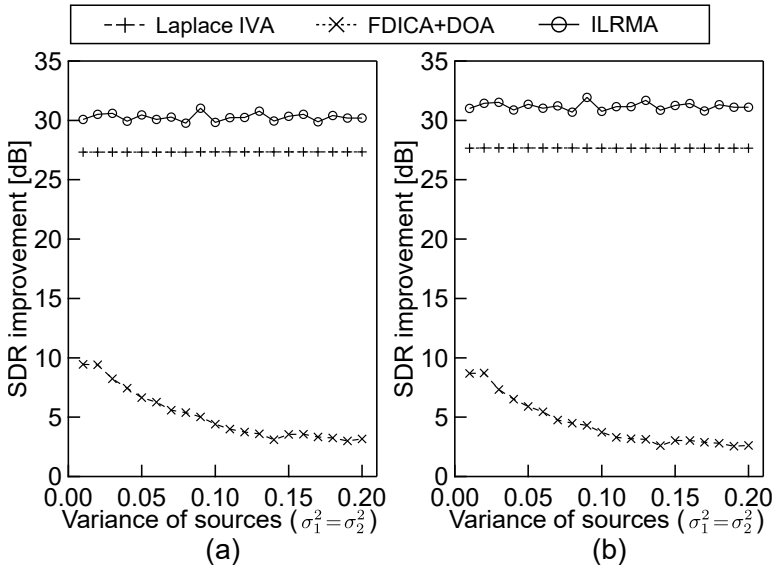

Figure 3.12: Separation results of (a) first source and (b) second source for various variances.

Figure 3.11 shows the SDR results for various positions of the sources (μ_1 and μ_2), where the horizontal axis indicates the angle between the two sources, $\mu_2 - \mu_1$, and the variances are fixed to $\sigma_1^2 = \sigma_2^2 = 0.05$. Also, Fig. 3.12 shows the SDR results for various variances (σ_1^2 and σ_2^2), where σ_1^2 and σ_2^2 are always set to the same value and the positions of the sources are fixed to $\mu_1 = 5\pi/12$ and $\mu_2 = 7\pi/12$. From these results, we can confirm that the separation performance of FDICA+DOA is sensitive to the mixing system. In particular, when the source positions become close (around 0.0 on the horizontal axis in Fig. 3.11) or the variance of the DOAs. is large (around 0.20 on the horizontal axis in Fig. 3.12), the permutation solver using the DOA cannot cluster the sources correctly, resulting in large permutation errors. In contrast, Laplace IVA and ILRMA achieve good performance regardless of the mixing system because these methods do not have explicit spatial constraints. This shows the flexibility of the spatial model in ILRMA.

3.6 Comparison of Speech and Music Separation Performance

In this section, I confirm the efficacy of ILRMA for determined BSS task by comparing the separation performance of many techniques.

3.6.1 Datasets

In this experiment, I investigated two cases: speech signal and music signal cases. In the speech signal case, I used live recorded mixture signals obtained from an underdetermined BSS task in SiSEC2011 [122]. This dataset includes 12 mixture signals (*dev1* and *dev2* datasets) with female and male speech, where the reverberation time is 130 ms/250 ms and the microphone spacing is 1 m/5 cm. Details of the other conditions for this dataset can be found in [122]. Note that since this dataset is for underdetermined BSS, three sources ($N = 3$) are provided as stereo recordings ($M = 2$). In this experiment, I used only the first and second speech sources to make the task determined ($M = N = 2$). In the music signal

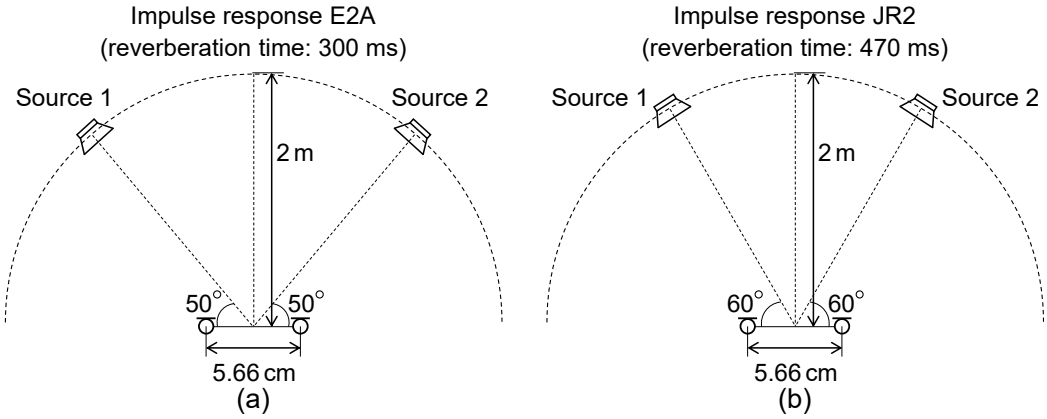

Figure 3.13: Recording conditions of impulse responses (a) E2A and (b) JR2 for two-source case.

Table 3.3: Music sources for two-source case

ID	Song name	Source (1/2)
1	bearlin-roads	acoustic_guit_main/vocals
2	another_dreamer-the_ones_we_love	guitar/vocals
3	fort_minor-remember_the_name	violins_synth/vocals
4	ultimate_nz_tour	guitar/synth

case, the observed signals were produced by convoluting the impulse response *E2A* or *JR2*, which was obtained from the RWCP database [202], with each source. Figure 3.13 shows the recording conditions of impulse responses E2A and JR2. As the music sources, I used professionally produced music obtained from a music separation task in SiSEC2011. The titles of the music and the instruments used are shown in Table 3.3.

3.6.2 Experimental Analysis of Optimal Number of Bases

In this subsection, I give an experimental analysis of the optimal number of bases in ILRMA. Since NMF decomposition is more suitable for music than speech because of the stable pitch of instruments, we expect that the optimal number of bases will be different between them. For this reason, I evaluated the

Table 3.4: Experimental conditions

Sampling frequency	16 kHz
Window length in STFT	256 ms in speech signal case and 512 ms in music signal case
Window function	Hamming window
Window shift length	128 ms in both speech and music signal cases
Initialization	$\boldsymbol{W}_i$: identity matrix NMF variables: uniform random values $[0, 1]$
Number of iterations	200

separation performance of ILRMA without a partitioning function using various numbers of bases for each source, where this method models all the sources with the same fixed number of bases L. The experimental conditions used are shown in Table 3.4. As the evaluation score, I used the improvement of SDR.

Figures 3.14 and 3.15 show the average SDR improvements and their deviations in 10 trials with different various pseudorandom seeds, where the speech signal (Fig. 3.14) is a female speech from the dev1 dataset with 130 ms reverberation time and 1 m microphone spacing, and the music signal (Fig. 3.15) is song ID4 with impulse response E2A. From these results, we confirm that ILRMA cannot achieve a good separation performance for speech signals when the number of bases is large. This is due to the structural complexity of the speech spectrogram. Figure 3.16 shows cumulative singular values of each source spectrogram in the speech and music signals. The speech sources require more than 50 bases to represent the spectrogram while the music sources are saturated with 25 bases. Because of the time-varying pitch, it is difficult to capture speech spectrograms using NMF decomposition. If ILRMA fails to capture the correct spectrogram of each speech in the optimization, the demixing matrix will be trapped at a poor solution (local minimum). On the other hand, owing to the low rank of music spectrograms, ILRMA gives a better performance for music separation even if the number of bases increases.

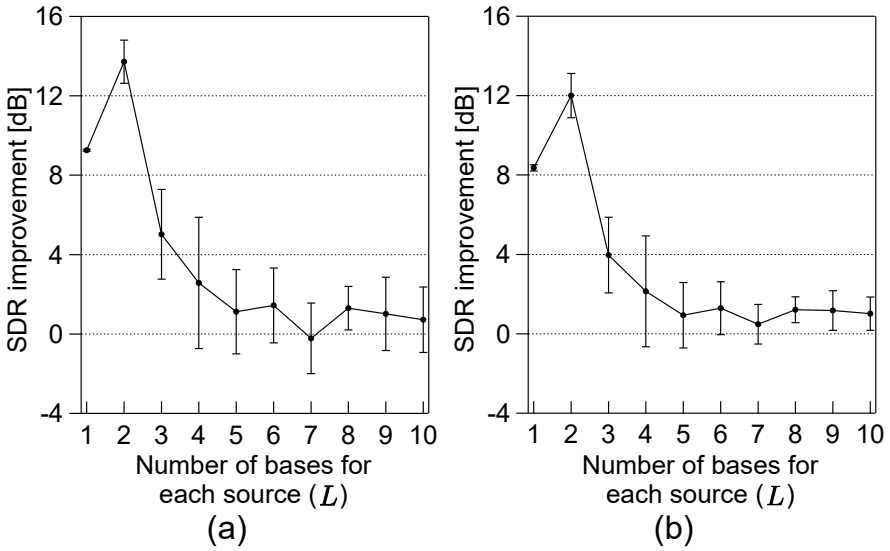

Figure 3.14: Average SDR improvements for female speech (dev1) with 1 m microphone spacing and 130 ms reverberation time: (a) first speaker and (b) second speaker.

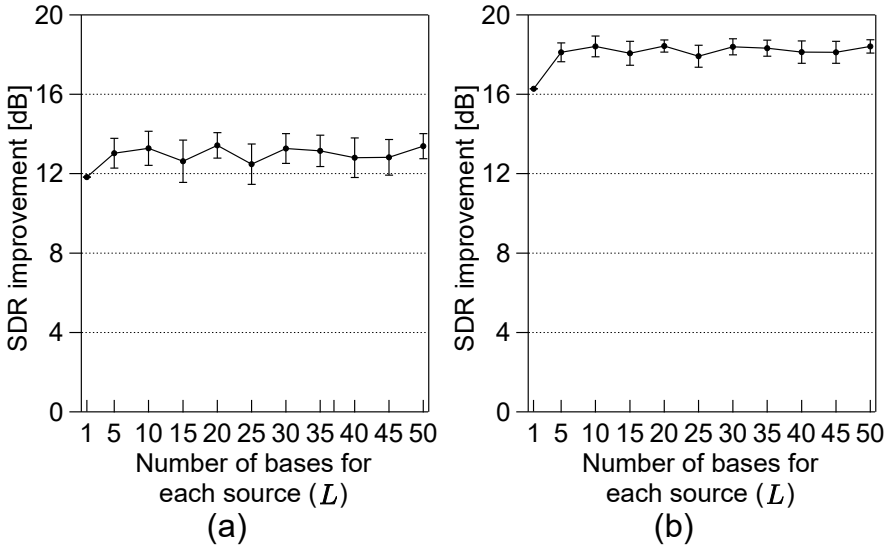

Figure 3.15: Average SDR improvements for song ID4 with impulse response E2A: (a) guitar and (b) synth.

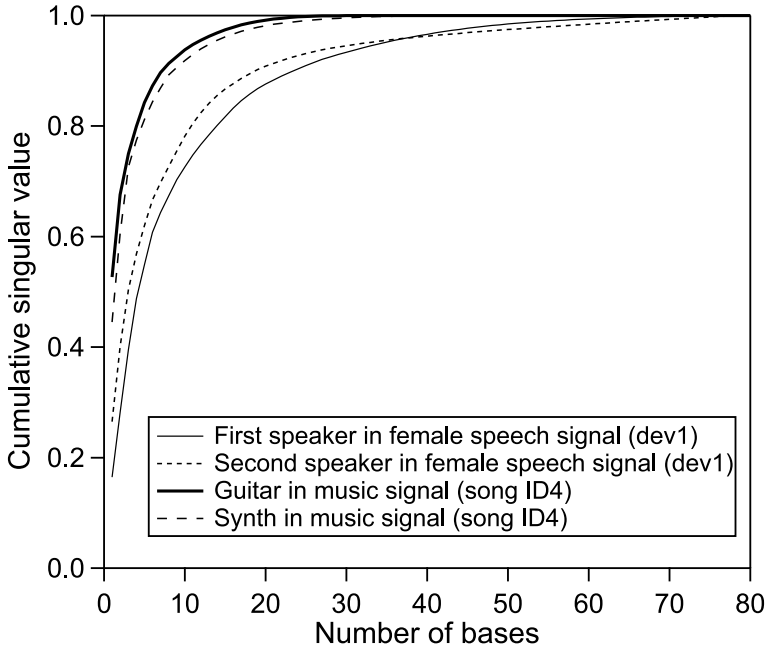

Figure 3.16: Cumulative singular values of each source spectrogram in dev1 female speech and song ID4 music, where all sources are truncated to be the same signal length.

3.6.3 Comparison of Separation Performance

Experimental Conditions

I next compare the separation performance of eight methods, namely, *directional clustering (DC)* [60], *Laplace IVA*, *Ozerov's MNMF*, *Ozerov's MNMF with random initialization*, *Sawada's MNMF*, *ILRMA w/o partitioning function*, *ILRMA with partitioning function*, and *Sawada's MNMF initialized by ILRMA*. DC is a simple separation technique, which clusters all the STFT coefficients into specific sources using both powers and phases. In this experiment, I use k-means clustering in directional clustering, which corresponds to a double-disjoint assumption, namely, we assume that each time-frequency slot has only one source component. In Ozerov's MNMF, I used the experimental conditions described in [6], as shown in Table 3.5, where the mixing matrices and the source models are initialized by estimation using Soft-LOST [203] with the permutation solver [181].

Table 3.5: Experimental conditions used in Ozerov's MNMF

Sampling frequency	16 kHz
Window length in STFT	128 ms
Window function	Hamming window
Window shift length	64 ms
Number of bases	10 bases for each speech source and 4 bases for each music source
Initialization of mixing matrices	Mixing matrices estimated by Soft-LOST [203] and permutation solver [181]
Initialization of source models (NMF variables)	Pretrained bases and activations using simple NMF based on KL divergence with sources estimated by Soft-LOST and [181]
Annealing for EM algorithm	Annealing with noise injection proposed in [6]
Number of iterations	500

Also, Ozerov's MNMF with random initialization has the same conditions as Ozerov's MNMF except for the initialization, namely, the mixing matrices and the source models are initialized by the identity matrix and the uniform random values $[0, 1]$, respectively. In the other methods, the experimental conditions shown in Table 3.4 were used. In ILRMA with partitioning function, I only set the total number of bases, K, and the sources are flexibly modeled with the optimal number of bases using the partitioning function $\boldsymbol{Z}$. Sawada's MNMF initialized by ILRMA has the same algorithm as Sawada's MNMF, but the initial values of the spatial covariance matrix $\boldsymbol{R}_{i,n}^{(s)}$ are given by (3.76), where the steering vector $\boldsymbol{a}_{i,n}$ is calculated from the inverse of the demixing matrix $\boldsymbol{W}_i$ estimated by ILRMA w/o partitioning function.

On the basis of the results in Sect. 3.6.2, I set the number of bases of each source to $L = 2$ for the speech signals and $L = 30$ for the music signals in ILRMA w/o partitioning function. In ILRMA with partitioning function and Sawada's MNMF, I set the total number of bases to $K = 2 \times N$ for the speech signals and $K = 30 \times N$ for the music signals. The number of bases used in Ozerov's MNMF

is shown in Table 3.5.

Results

Figures 3.17 and 3.18 respectively show examples of results for speech signals given by the average SDR improvements and their deviations in 10 trials with different pseudorandom seeds. Also, Figs. 3.19 and 3.20 show examples of results for music signals. The total average scores are shown in Tables 3.6 and 3.7. From these results, we confirm that DC cannot separate the sources because of the imperfect double-disjoint assumption and the deviation of the DOAs in reverberant environments. Also, Laplace IVA cannot achieve satisfactory separation because the source model in Laplace IVA is not flexible as described in Sect. 3.5. Ozerov's MNMF outperforms Laplace IVA for the music signals, but the separation performance for speech signals is inferior to that of Laplace IVA. In addition, Ozerov's MNMF with random initialization cannot solve the BSS problem. This method must be initialized by other methods to find a good solution. The results of Sawada's MNMF have large error bars, namely, this method is also sensitive to initial values. However, for the music signals, Sawada's MNMF gives better performance than Laplace IVA and Ozerov's MNMF. ILRMA-based methods achieve a high and stable performance. For the speech signals, ILRMA w/o partitioning function is preferable to ILRMA with partitioning function. This might be due to the sensitivity of the performance to the number of bases, as discussed in Sect. 3.6.2. In contrast, for the music signals, ILRMA with partitioning function exhibits slightly higher performance than ILRMA w/o partitioning function. This improvement is achieved by modeling the sources with the optimal number of bases using the partitioning function z_{nk}. Figure 3.21 shows an example of the convergence of the partitioning function z_{1k} from $k = 1$ to $k = K$ in the music signal case. These values indicate whether the kth basis contributes to only source one ($z_{1k} = 1$) or only source two ($z_{1k} = 0$). We can confirm that almost partitioning functions converge to one or zero, but several ones converge to intermediate values. This is because similar or the same spectral patterns appear in both two sources. Thanks to the partitioning function, all the sources can effectively be modeled with the optimal number of bases.

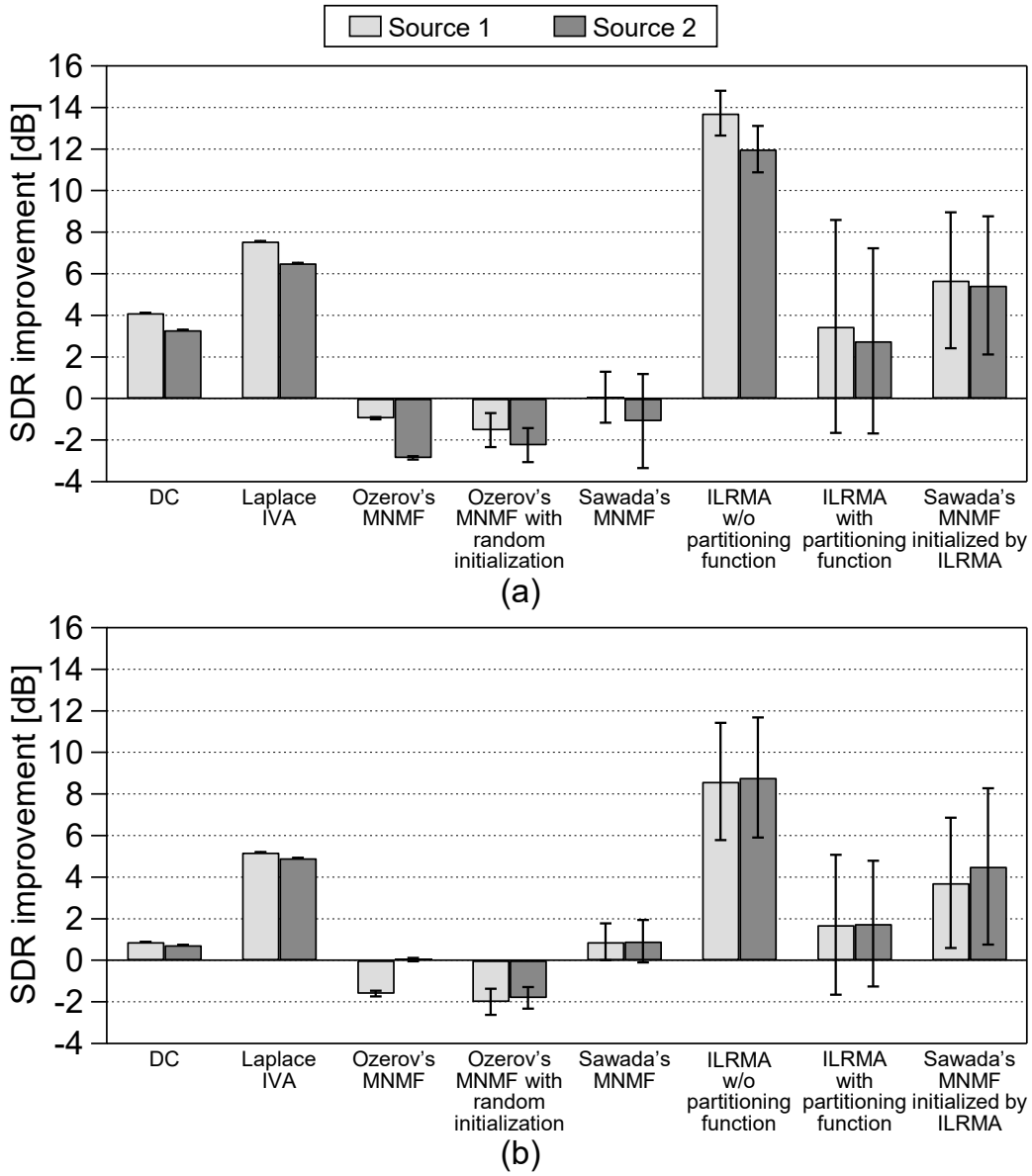

Figure 3.17: Average SDR improvements for female speech (dev1) with 1 m microphone spacing, where reverberation time is (a) 130 ms and (b) 250 ms.

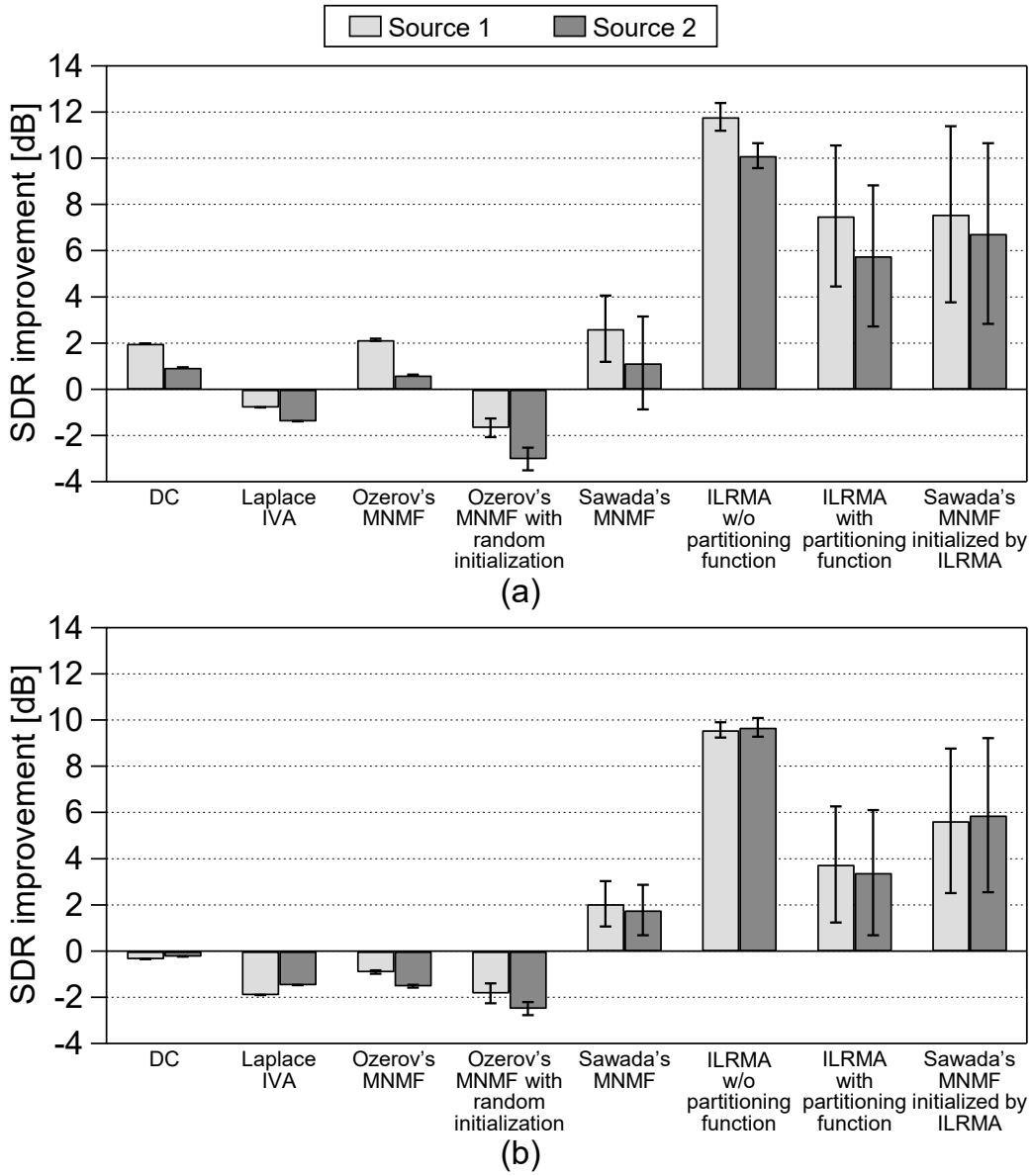

Figure 3.18: Average SDR improvements for male speech (dev1) with 1 m microphone spacing, where reverberation time is (a) 130 ms and (b) 250 ms.

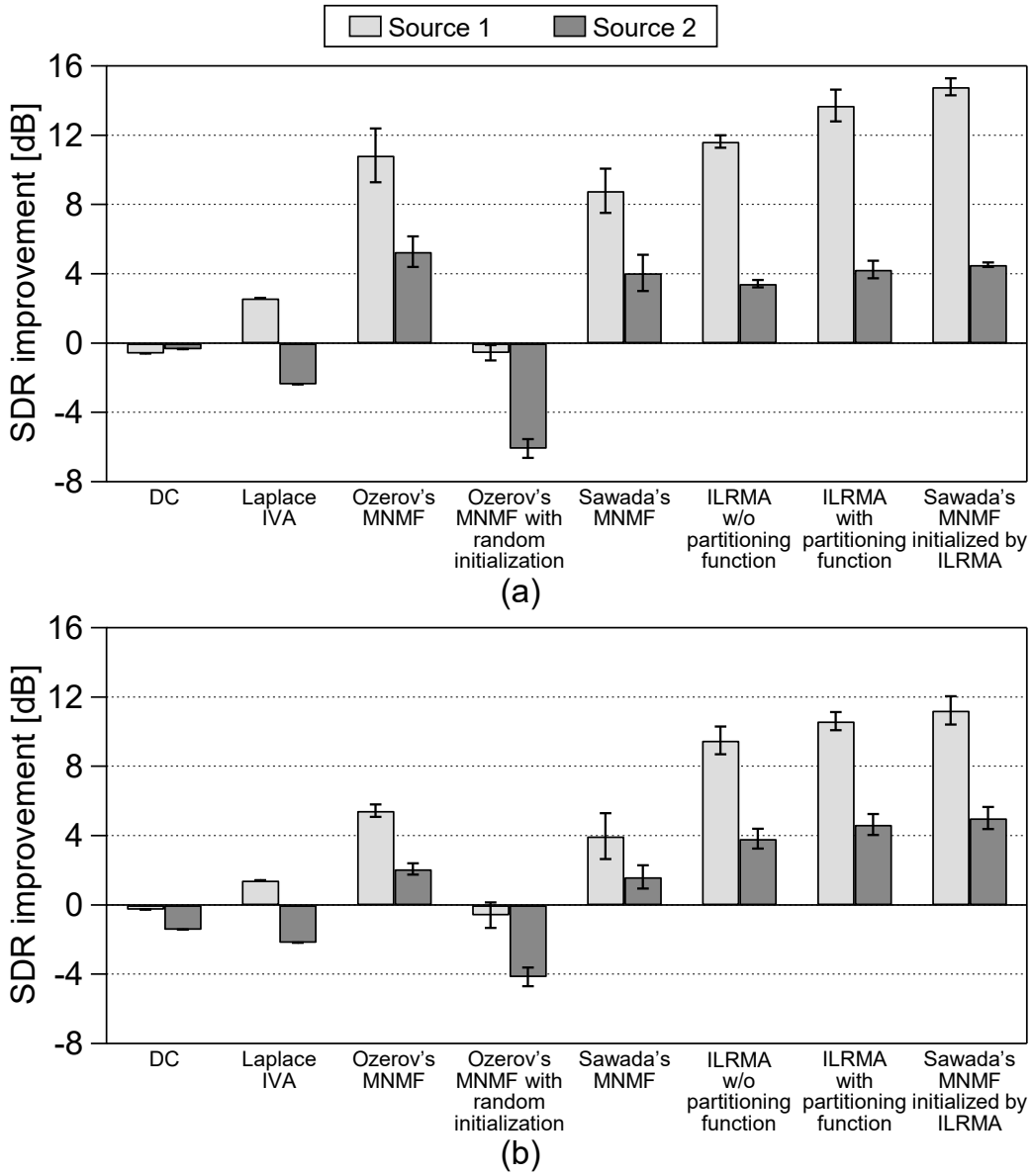

Figure 3.19: Average SDR improvements for music signal song ID3 with impulse response (a) E2A and (b) JR2.

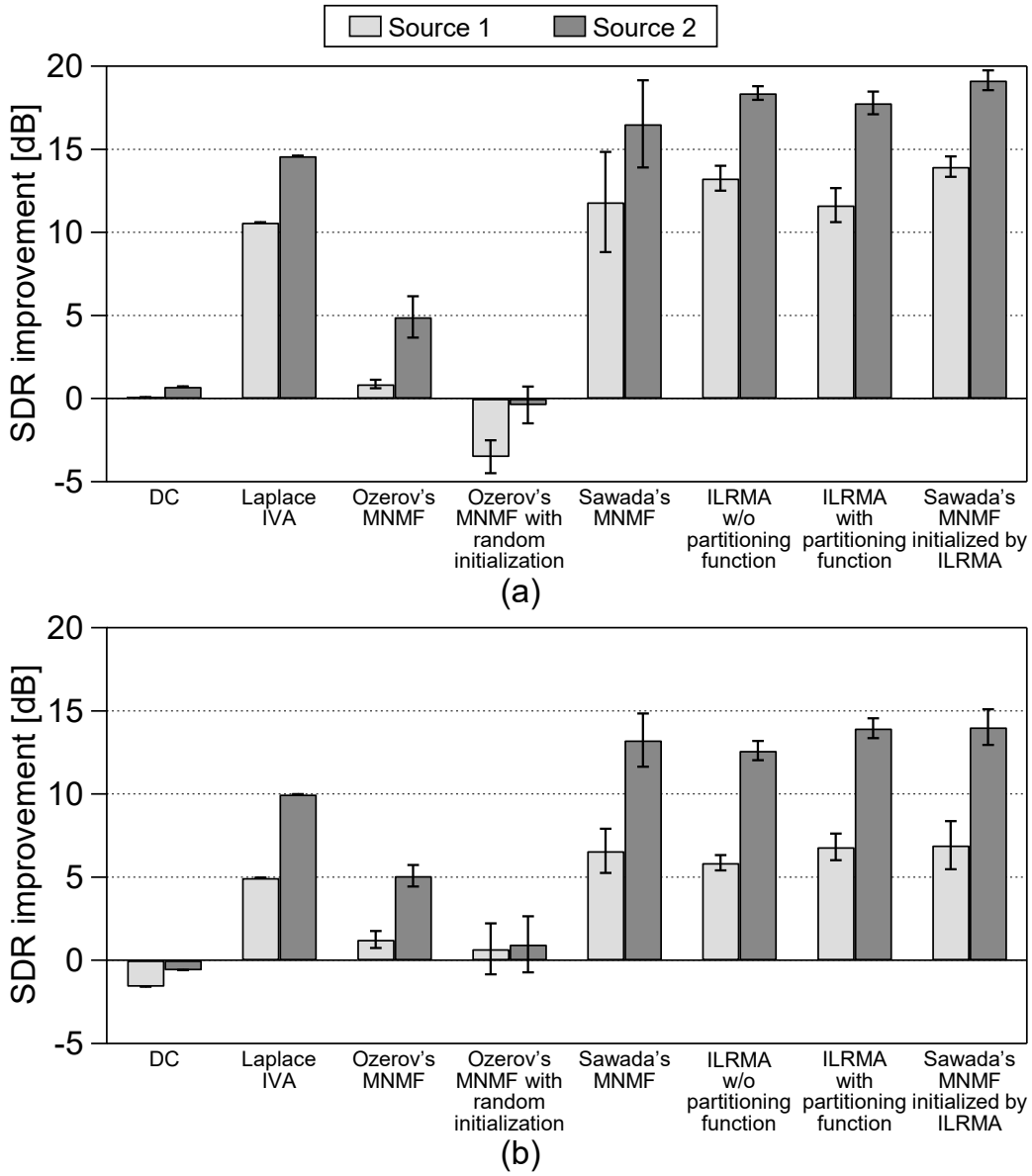

Figure 3.20: Average SDR improvements for music signal song ID4 with impulse response (a) E2A and (b) JR2.

Table 3.6: Averaged SDR improvements over various speech signals and sources with same recording conditions in two-source case

Conditions (rev. time and mic. spacing)	DC	Laplace IVA	Ozerov's MNMF	Ozerov's MNMF with random initialization	Sawada's MNMF	ILRMA w/o partitioning function	ILRMA with partitioning function	Sawada's MNMF initialized by ILRMA
130 ms & 1 m	2.59	2.98	1.35	-2.11	0.68	**11.91**	4.88	6.36
130 ms & 5 cm	-1.51	2.86	2.13	-0.22	1.13	**8.97**	3.48	5.60
250 ms & 1 m	0.14	2.03	0.49	-2.02	0.48	**7.34**	2.09	4.19
250 ms & 5 cm	-1.56	2.43	0.91	-1.06	0.47	**6.43**	1.91	3.95

Table 3.7: Averaged SDR improvements over various music signals and sources with same impulse response in two-source case

Impulse response	DC	Laplace IVA	Ozerov's MNMF	Ozerov's MNMF with random initialization	Sawada's MNMF	ILRMA w/o partitioning function	ILRMA with partitioning function	Sawada's MNMF initialized by ILRMA
E2A	-0.73	5.72	5.73	-2.70	10.32	12.29	12.29	**14.41**
JR2	-1.18	1.77	2.37	0.75	6.11	6.62	7.40	**9.06**

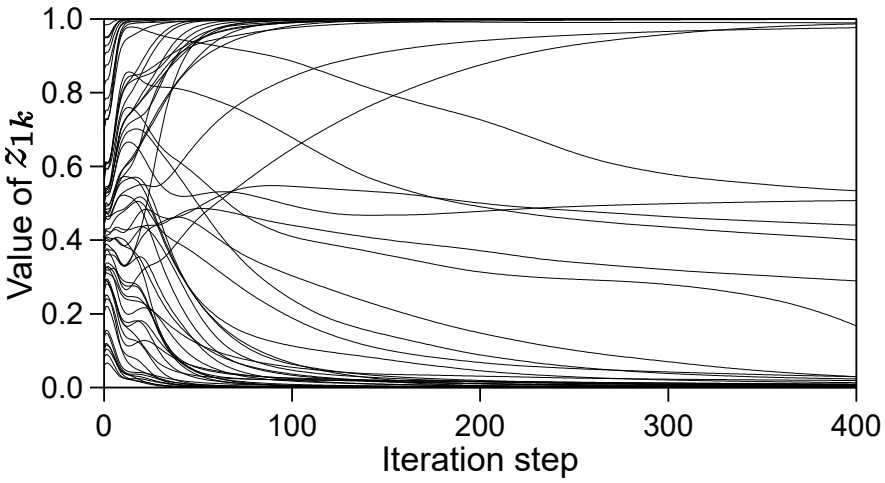

Figure 3.21: Convergence of z_{1k} from $k = 1$ to $k = K$ in music signal case.

The deviations of the ILRMA-based methods are smaller than those of Ozerov's and Sawada's MNMFs, which is particularly evident in ILRMA w/o partitioning function. This is because the optimization of the demixing matrix using the IVA update rules results in a stable separation performance. In fact, I experimentally confirmed that the initialization using Soft-LOST [203] and

the permutation solver [181], which was employed in Ozerov's MNMF, did not improve the separation performance of ILRMA w/o partitioning function. This fact means that ILRMA is robust against the initial values. For music signals with impulse response JR2 (Figs. 3.19 (b) and 3.20 (b)), the SDRs of the ILRMA-based methods are markedly degraded compared with those with impulse response E2A because the reverberation time is longer than impulse response E2A and is close to the length of the window function in the STFT. Even if Sawada's MNMF has the potential to model such a mixing system by employing a full-rank spatial model, it is a very difficult problem to find the optimal $\boldsymbol{R}_{i,n}^{(s)}$. However, Sawada's MNMF initialized by ILRMA can achieve high and very stable separation performance even with impulse response JR2. This means that the demixing matrix estimated by ILRMA can be a good initial value of the spatial model $\boldsymbol{R}_{i,n}^{(s)}$ in order to find the full-rank spatial covariance.

Figure 3.22 shows an example of the SDR convergence for each method in music signal case. Both Laplace IVA and the proposed methods show much faster convergence than Sawada's MNMF. Also, the numbers of required iterations in Sawada's MNMF is greatly reduced by the initialization of the rank-1 spatial covariance. This result shows the difficulty of optimizing the full-rank spatial covariance $\boldsymbol{R}_{i,n}^{(s)}$.

Figure 3.23 shows a result of a subjective evaluation, where I presented 48 pairs of separated speech and 48 pairs of separated music signals in random order to 14 examinees, who selected which signal they preferred from the viewpoint of the total quality of the separated sounds. Also, Fig. 3.24 shows a probability of selection regarding a difference between two subjective scores. For example, the difference of subjective scores of Laplace IVA and Ozerov's MNMF for speech signals is around 0.9, and it means that Laplace IVA is preferably selected with a 81% probability when it is compared with Ozerov's MNMF. We can confirm that Laplace IVA is better than MNMF methods for the speech signals. In contrast, Sawada's MNMF achieves a better result for music signals owing to the suitable representation using NMF. ILRMA is the most preferable method for the high-quality separation of both speech and music signals. Similarly to FDICA and Laplace IVA, ILRMA employs the demixing matrix $\boldsymbol{W}_i$ for the separation, which is essentially equivalent to the spatial linear filter [47]

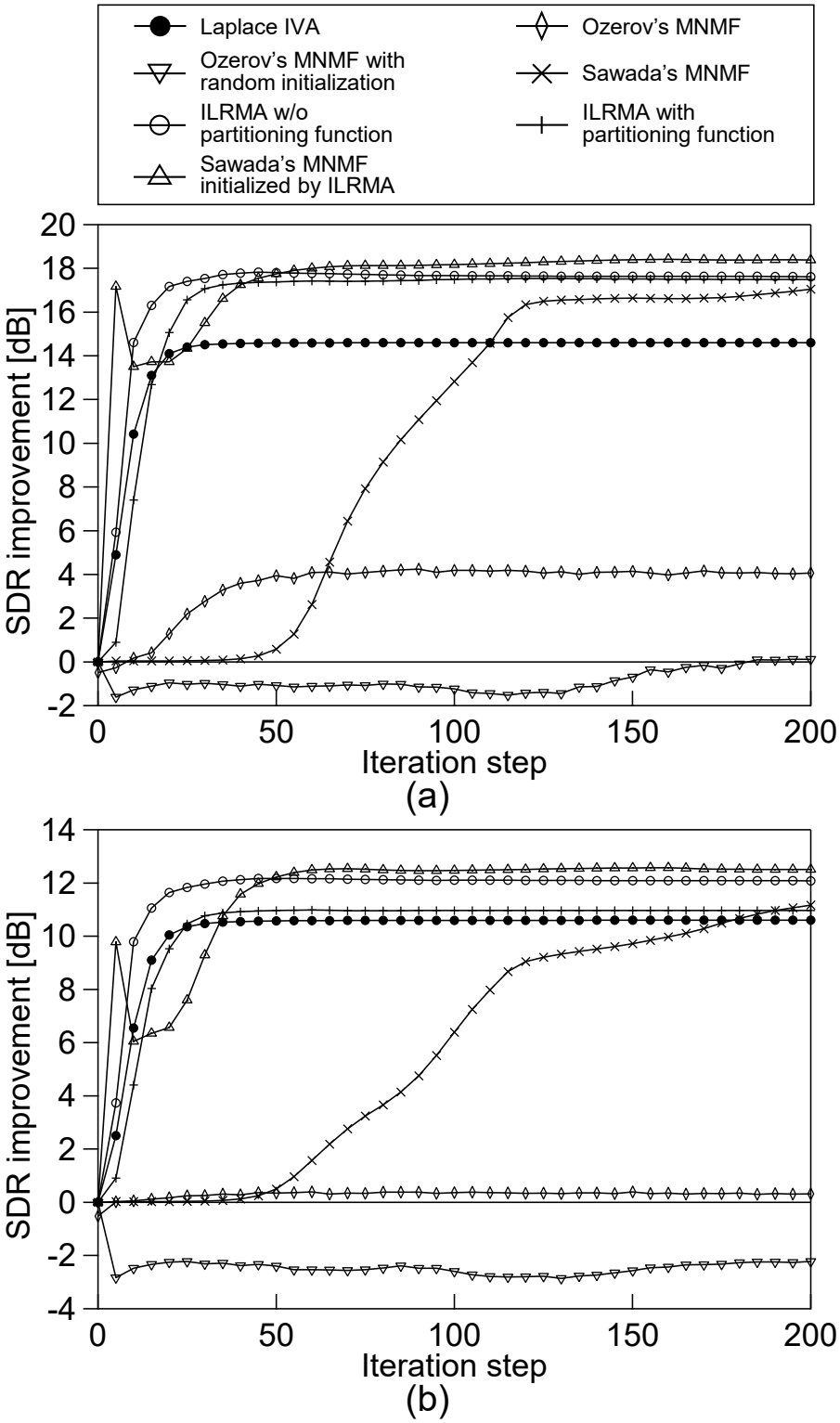

Figure 3.22: SDR convergence for music signal song ID4 with impulse response E2A: (a) guitar and (b) synth.

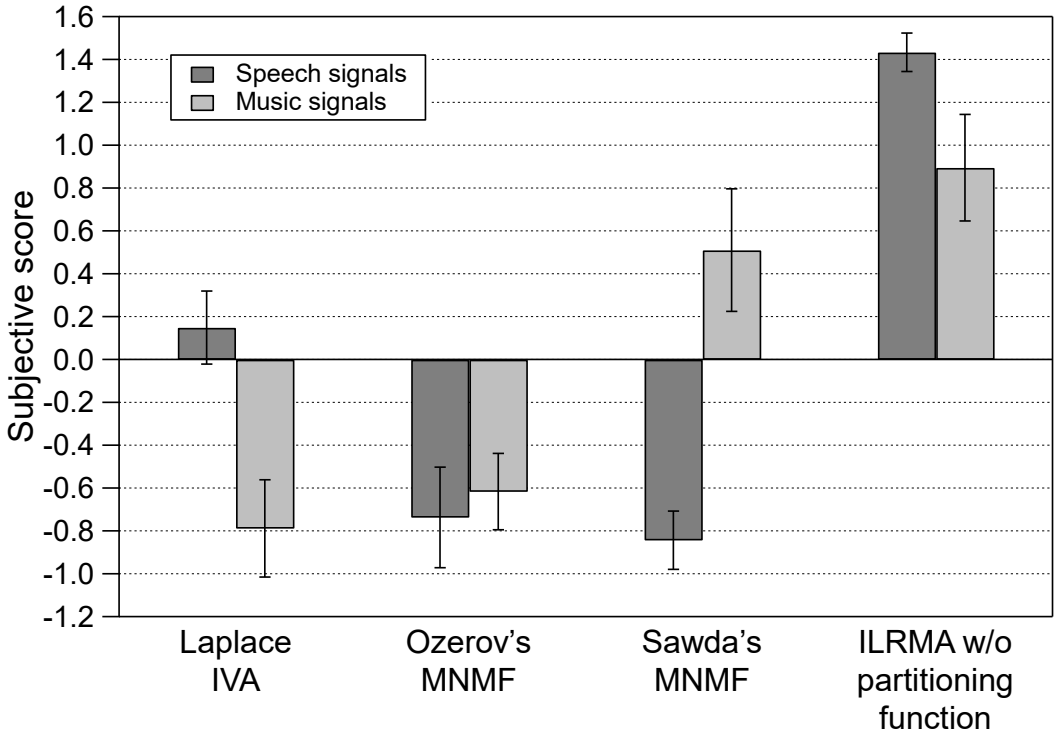

Figure 3.23: Results of subjective scores obtained by Thurstone pairwise comparison method, where 48 pairs of separated speech and 48 pairs of separated music signals are presented in random order to 14 examinees, who selected which signal they preferred from the viewpoint of total quality of separated sound. Scores show relative tendency of selection.

in beamforming techniques [204, 205], and it is more difficult for such linear filtering to generate artificial noise than for time-frequency mask separation techniques including MNMF with MWF. Thus, the quality of separated sources via ILRMA from the viewpoint of human perception might be better than that via MNMF.

3.6.4 Experiments on Three-Source Case with Music Signals

I also conducted an experiment involving three sources and three microphones ($M = N = 3$) with music signals. Similarly to the music dataset described in Sect. 3.6.1, I produced the observed signals using the same songs and the three

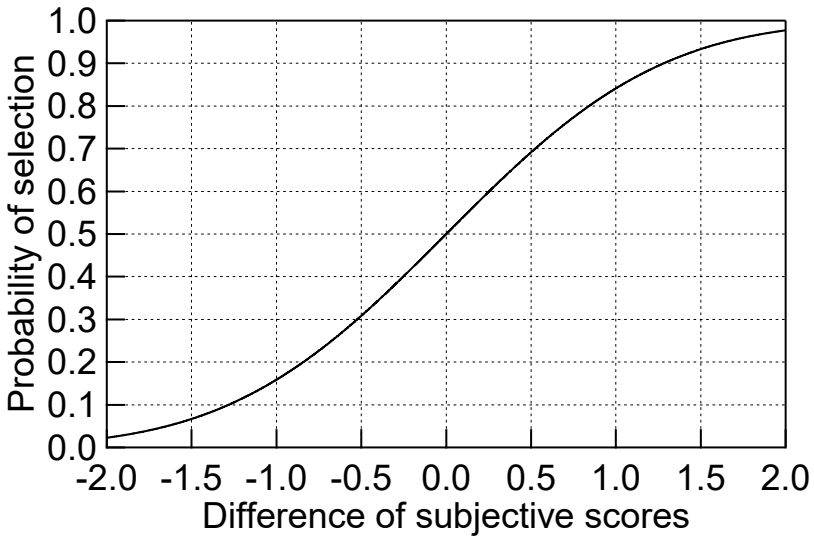

Figure 3.24: Probability of selection regarding difference between two subjective scores.

Table 3.8: Music sources for three-source case

ID	Song name	Source (1/2/3)
1	bearlin-roads	acoustic_guit_main/bass/vocals
2	another_dreamer-the_ones_we_love	drums/guitar/vocals
3	fort_minor-remember_the_name	drums/violins_synth/vocals
4	ultimate_nz_tour	guitar/synth/vocals

instruments shown in Table 3.8 with the impulse responses shown in Fig. 3.25. The experimental conditions are those in Table 3.4, where I here omit the results of DC, Ozerov's MNMF, and Ozerov's MNMF with random initialization.

Figure 3.26 shows examples of results, and Table 3.9 shows the total average scores in the three-source case. Similarly to the previous results, the proposed method achieves better and more stable performance than Sawada's MNMF, and the spatial model estimated by ILRMA provides an efficient initialization for Sawada's MNMF. Table 3.10 shows the actual computational time for each method in the three-source case, where the calculations were performed using MATLAB 8.3 (64-bit) with an Intel Core i7-4790 (3.60 GHz) CPU. The computational times of ILRMA-based methods are less than twice that of

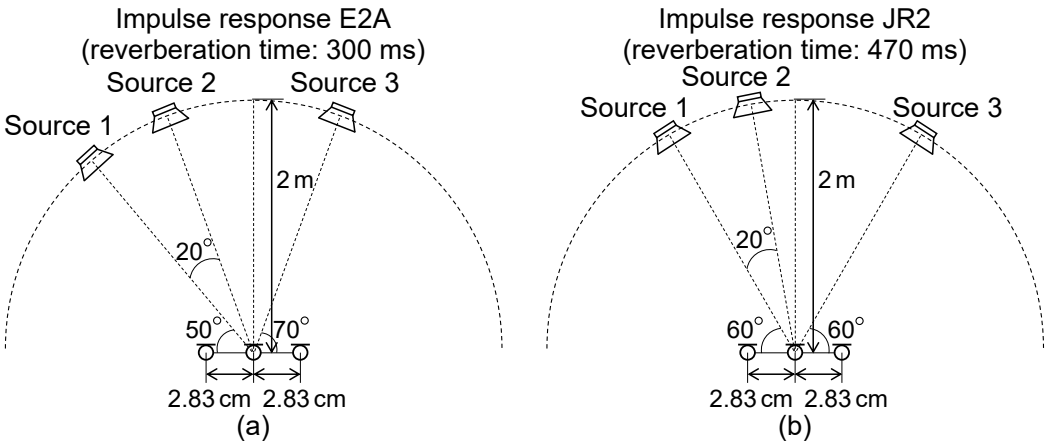

Figure 3.25: Recording conditions of impulse responses (a) E2A and (b) JR2 for three-source case.

Table 3.9: Averaged SDR improvements over various music signals and sources with same impulse response in three-source case

Impulse response	Laplace IVA	Sawada's MNMF	ILRMA w/o partitioning function	ILRMA with partitioning function	Sawada's MNMF initialized by ILRMA
E2A	3.86	7.77	8.03	6.18	**9.44**
JR2	2.81	4.44	5.03	4.11	**7.00**

Laplace IVA. Sawada's MNMF requires a longer computational time because the eigenvalue decomposition of a $2M \times 2M$ matrix is required for each update iteration of $\boldsymbol{R}_{i,n}^{(s)}$. From these results, ILRMA is advantageous in terms of the convergence speed and computational cost while maintaining comparable separation performance with Sawada's MNMF.

3.6.5 Experimental Analysis of Optimal Window Length

Since ILRMA estimates low-rank source model using NMF decomposition, a rank of an observed spectrogram directly affects the separation performance. As already shown in Sect. 3.6.2, the inherent time-frequency structure of each source

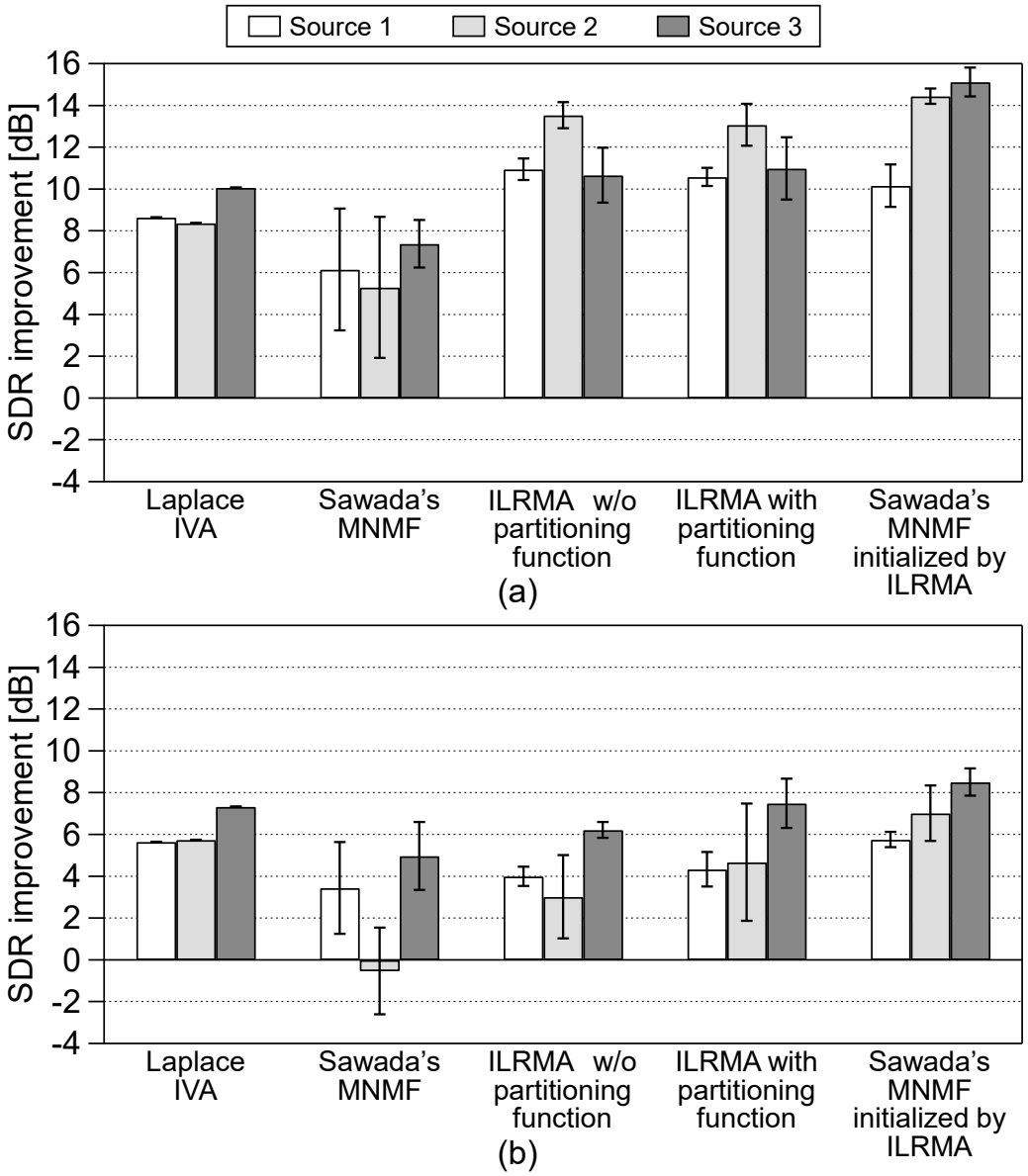

Figure 3.26: Average SDR improvements for music signal song ID4 in three-source case with impulse response (a) E2A and (b) JR2.

Table 3.10: Computational times (s) for separation of song ID1 with impulse response E2A in three-source case

Laplace IVA	Sawada's MNMF	ILRMA w/o partitioning function	ILRMA with partitioning function
91.6	4498.4	121.0	173.4

Table 3.11: Experimental conditions used in analysis of optimal window length

Sampling frequency	16 kHz
Window length in STFT	32/64/128/256/512/768/1024/1280/1536 ms
Window function	Hamming window
Window shift length	1/4 of window length
Initialization	$\boldsymbol{W}_i$: identity matrix NMF variables: uniform random values $[0, 1]$
Number of bases	$L = 5/10/20/30/40/50$ and $K = 2L$
Number of iterations	200

(e.g., speech or music) must be suitable for ILRMA to achieve a good separation, but the rank of the spectrogram also depends on a length of analysis window used in STFT. In this subsection, I compare separation performance of ILRMA with various window length and experimentally analyze its optimal condition.

Conditions

Similar to Sect. 3.6.1, I used the music dataset shown in Table 3.3, and the music sources were convolved with the impulse responses JR2 (Fig. 3.13 (b)). The other conditions are shown in Table 3.11. I compared nine lengths of analysis window, and its shift length was always set to a quarter of the window length. Also, the number of bases (L for ILRMA w/o partitioning functin and K for ILRMA with partitioning function) was set to six patterns. In this experiment, only Laplace IVA, ILRMA w/o partitioning function, and ILRMA with partitioning function were compared.

Results

Figures 3.27–3.30 show the separation results with various window lengths and number of NMF bases, where the SDR improvements are averaged for two sources and 10 trials using various pseudorandom seeds. From these results, we can confirm that the separation performance strongly depends on the window length in STFT rather than the number of NMF bases L or K. In particular, in Fig. 3.27, the long analysis window that exceeds 1.2 s provides the highest performance even though the improvements of Laplace IVA drop when the window length exceeds 1.0 s. In conventional FDICA, the separation fails if we use a too long window in STFT. This is because the sequential signals in each frequency bin close to a sinusoidal wave when the window length is too long, and their independence assumption between sources collapses in each frequency band, meaning that there is a fundamental limitation for FDICA [206]. Therefore, the separation performance has a trade-off based on the length of the analysis window in terms of the assumptions of linear time-invariant mixing and the independence of sources. In ILRMA, this trade-off might be solved owing to take the source model $r_{ij,n}$ into account for the estimation of $\boldsymbol{W}_i$ if the source model $r_{ij,n}$ could accurately capture the time-frequency structure. However, in the other results (Figs. 3.28–3.30), the optimal window length is around 0.5 s, and longer windows do not give a better separation.

Figure. 3.31 shows a number of bases in the power spectrogram of each source when its cumulative singular value reaches 80% or 90%, namely, an approximative rank of each source spectrogram. The songs ID1 and ID4 relatively have a low-rank power spectrograms. Indeed, the separation performance of these observations is better than those of the others (see Figs. 3.27 and 3.30). Thus, when we apply ILRMA to the observed signals that do not have a low-rank time-frequency structure (e.g., vocal or speech signal), the number of NMF bases should be set to a smaller value as shown in Figs. 3.15 and 3.28 (a) and a partitioning function should be omitted. For the signals that have low-rank spectrograms, we can increase the number of NMF bases to achieve more accurate separation.

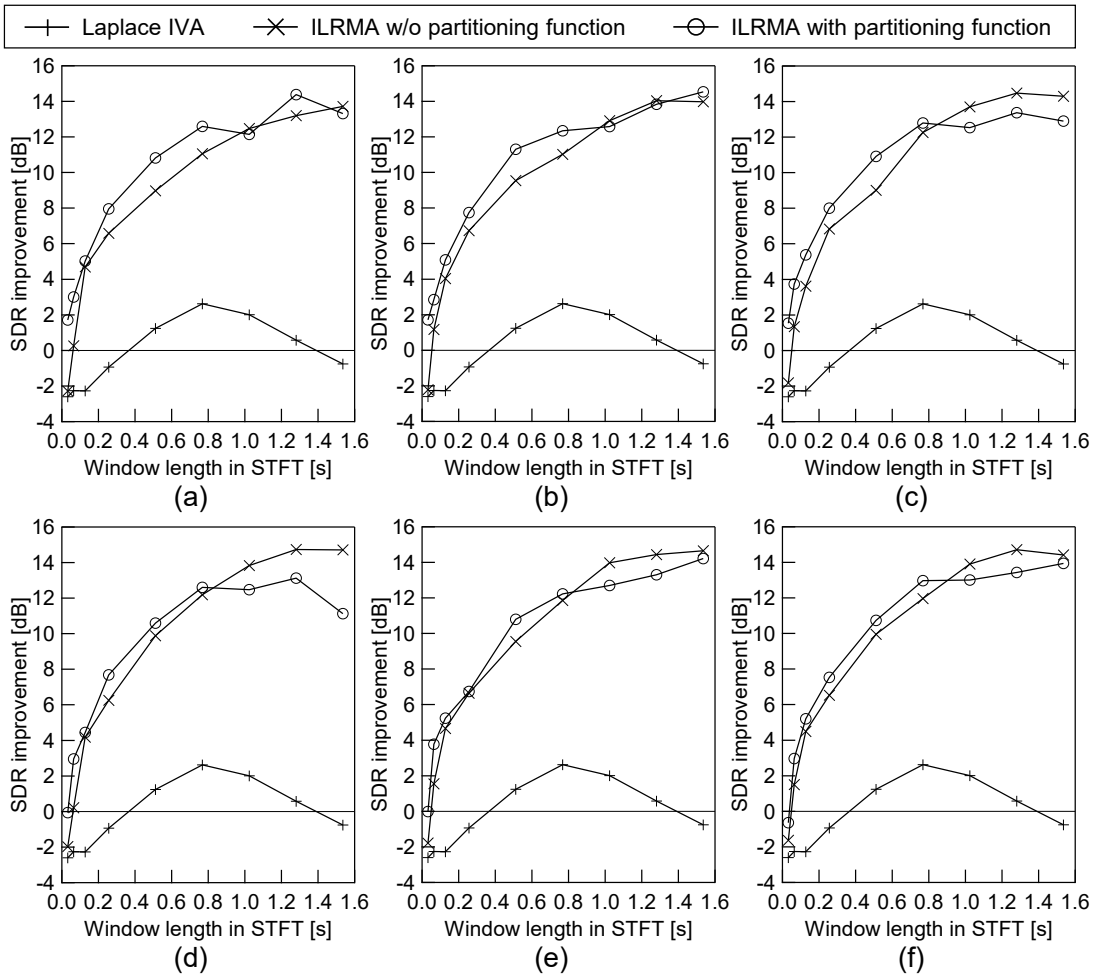

Figure 3.27: Averaged SDR improvements averaged for music signal song ID1 with impulse response JR2 (reverberation time is 470 ms): (a) $L = 5$ and $K = 10$, (b) $L = 10$ and $K = 20$, (c) $L = 20$ and $K = 40$, (d) $L = 30$ and $K = 60$, (e) $L = 40$ and $K = 80$, and (f) $L = 50$ and $K = 100$. Scores are averaged for two sources and 10 trials with different pseudorandom seeds.

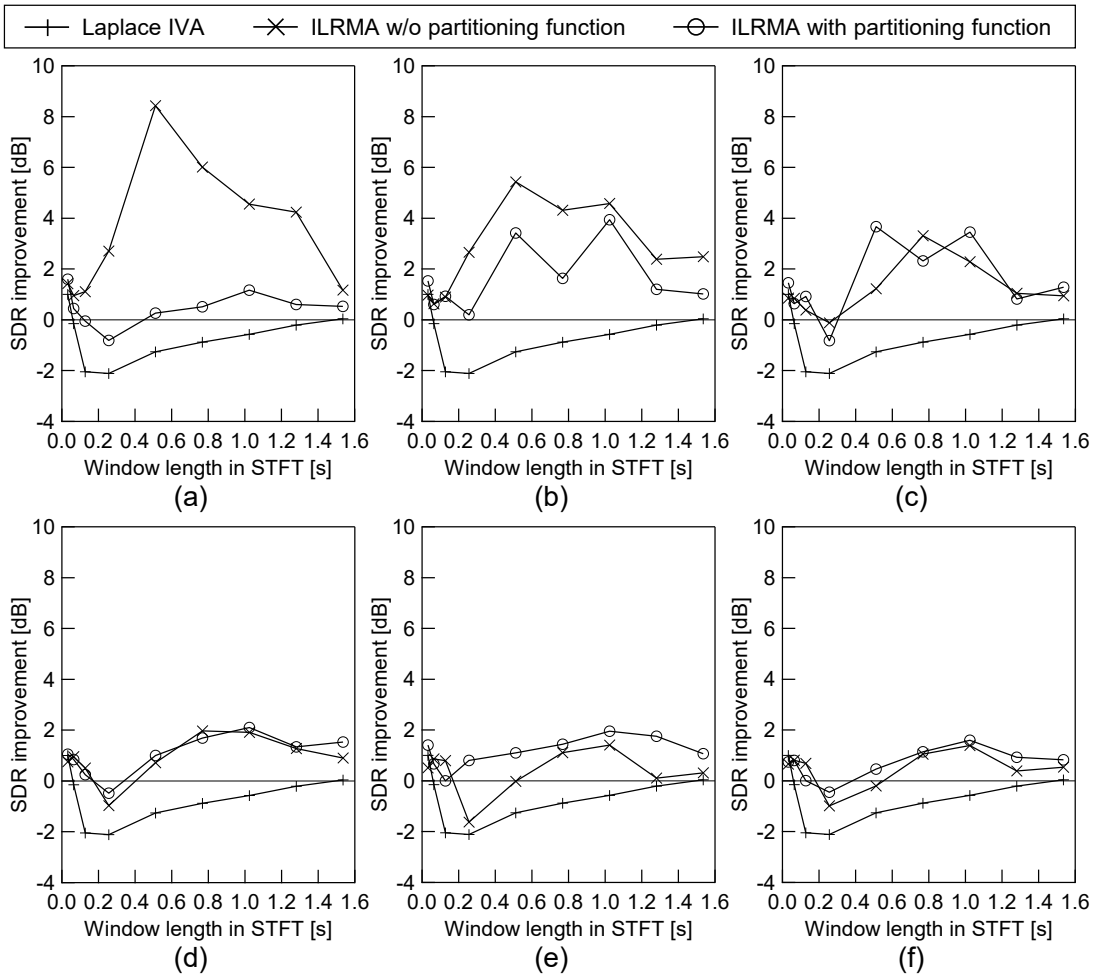

Figure 3.28: Averaged SDR improvements averaged for music signal song ID2 with impulse response JR2 (reverberation time is 470 ms): (a) $L = 5$ and $K = 10$, (b) $L = 10$ and $K = 20$, (c) $L = 20$ and $K = 40$, (d) $L = 30$ and $K = 60$, (e) $L = 40$ and $K = 80$, and (f) $L = 50$ and $K = 100$. Scores are averaged for two sources and 10 trials with different pseudorandom seeds.

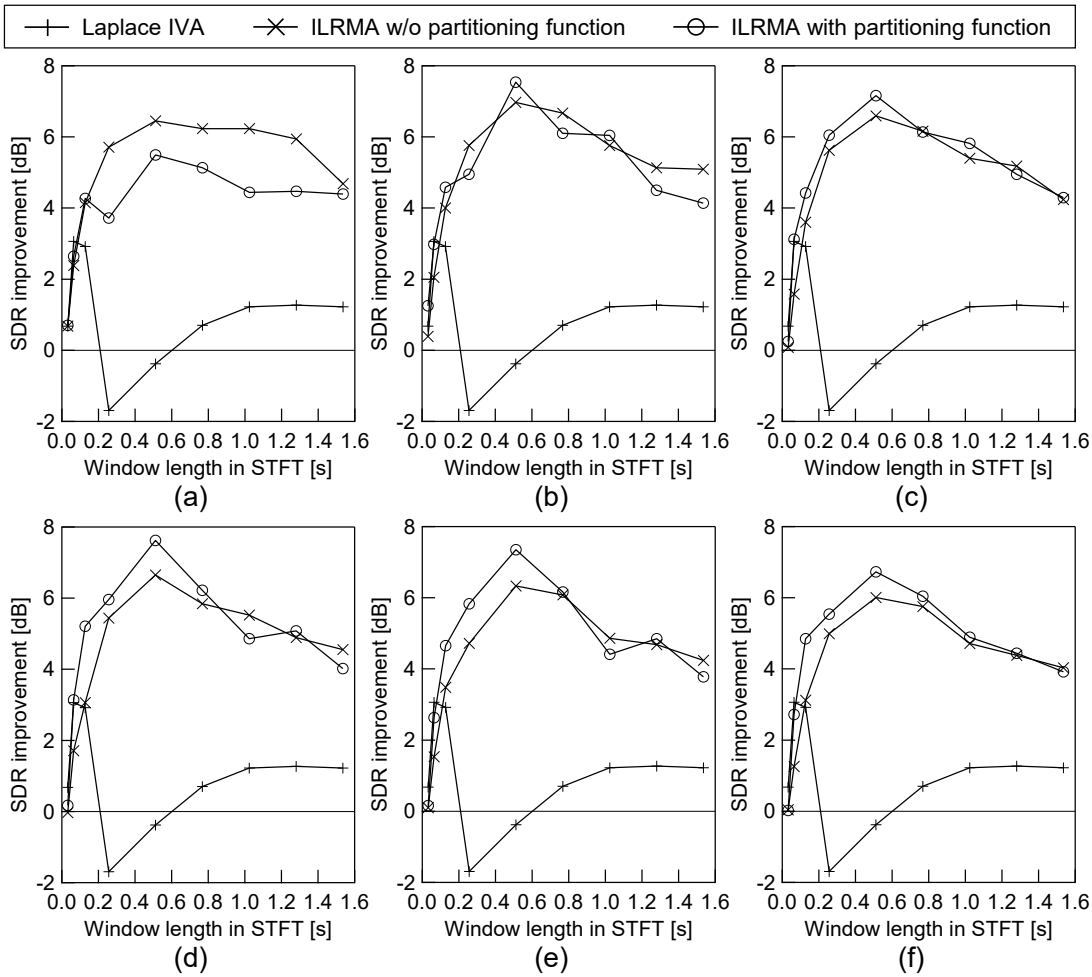

Figure 3.29: Averaged SDR improvements averaged for music signal song ID3 with impulse response JR2 (reverberation time is 470 ms): (a) $L = 5$ and $K = 10$, (b) $L = 10$ and $K = 20$, (c) $L = 20$ and $K = 40$, (d) $L = 30$ and $K = 60$, (e) $L = 40$ and $K = 80$, and (f) $L = 50$ and $K = 100$. Scores are averaged for two sources and 10 trials with different pseudorandom seeds.

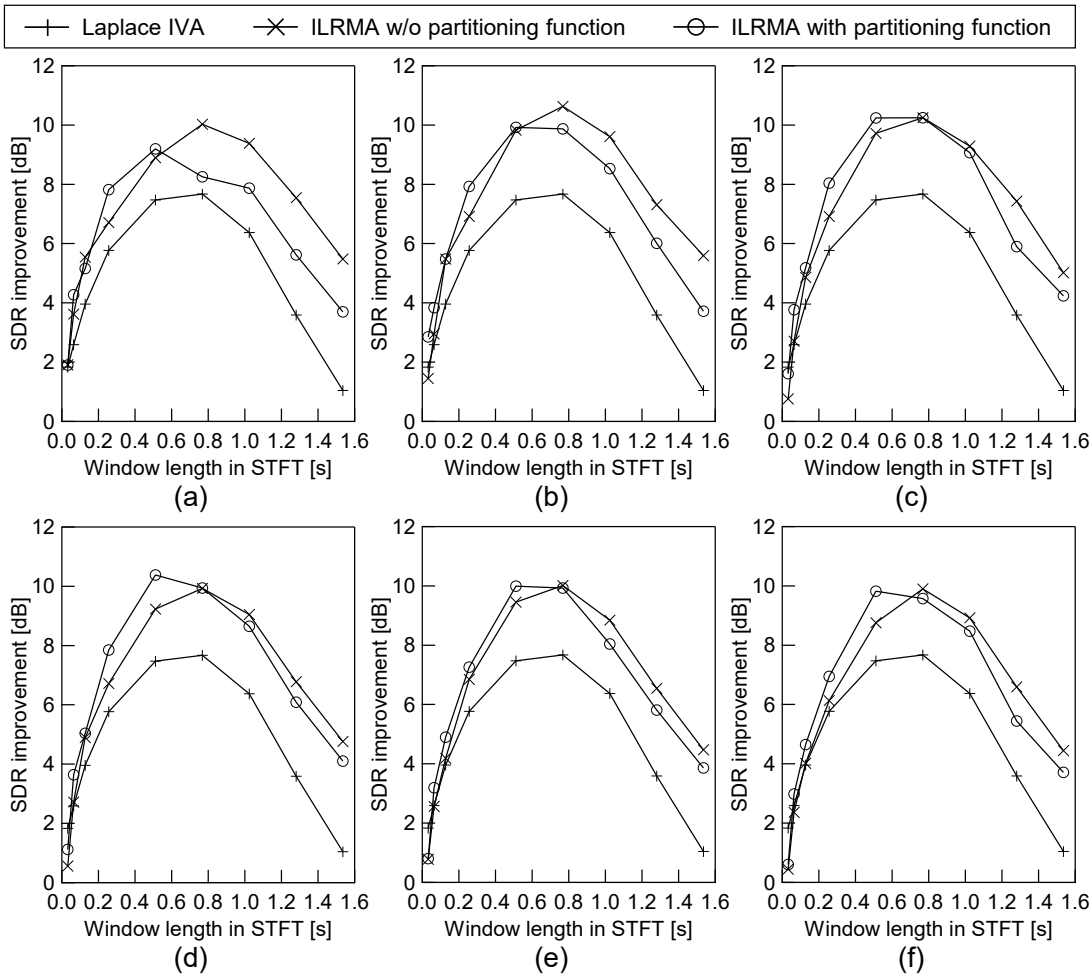

Figure 3.30: Averaged SDR improvements averaged for music signal song ID4 with impulse response JR2 (reverberation time is 470 ms): (a) $L = 5$ and $K = 10$, (b) $L = 10$ and $K = 20$, (c) $L = 20$ and $K = 40$, (d) $L = 30$ and $K = 60$, (e) $L = 40$ and $K = 80$, and (f) $L = 50$ and $K = 100$. Scores are averaged for two sources and 10 trials with different pseudorandom seeds.

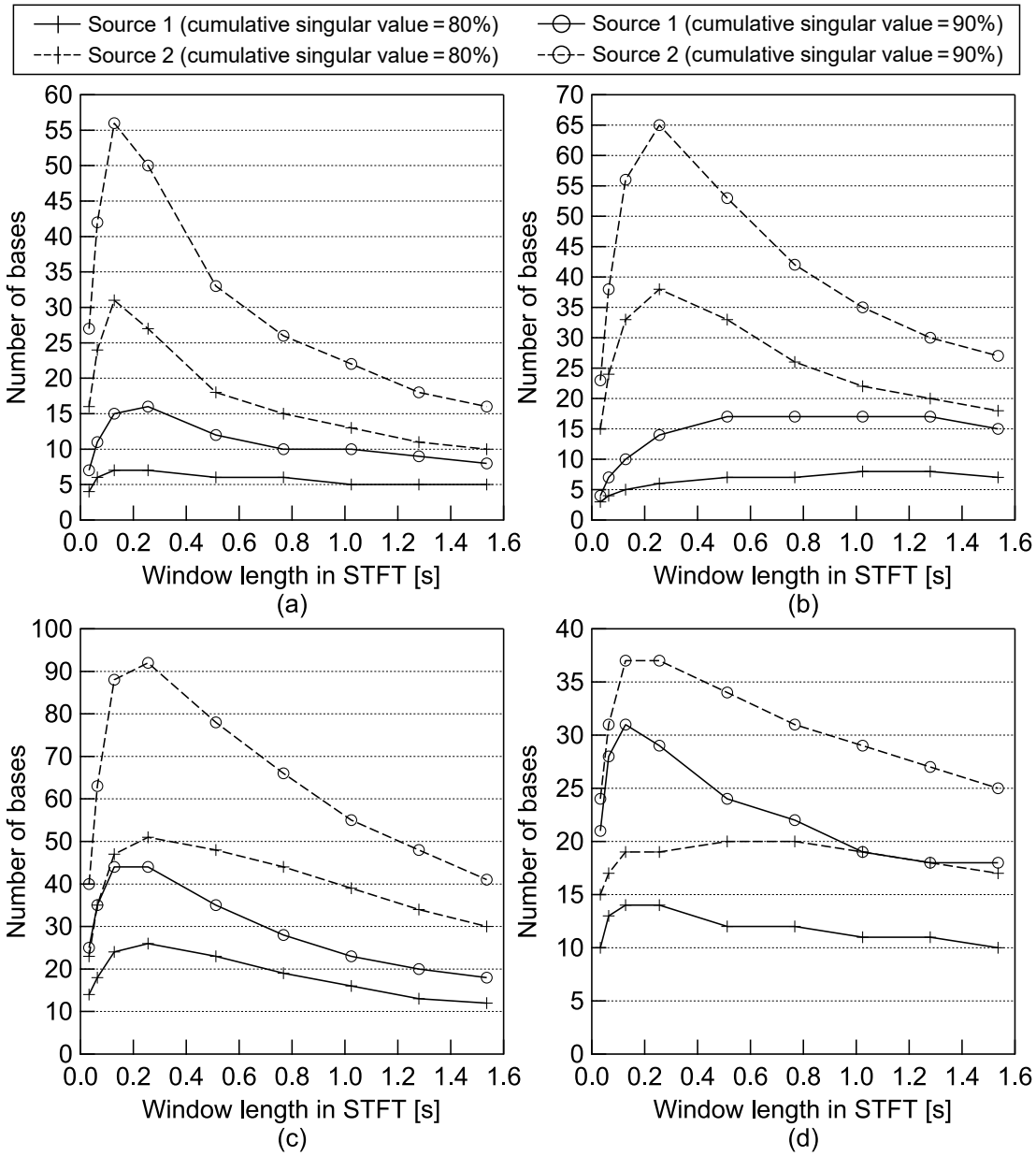

Figure 3.31: Number of bases in power spectrogram of each source when its cumulative singular value reaches 80% or 90%: (a) song ID1, guitar (source 1) and vocals (source 2), (b) song ID2, guitar (source 1) and vocals (source 2), (c) song ID3, violins synth. (source 1) and vocals (source 2), and (d) song ID4, guitar (source 1) and synth. (source 2).

3.7 Extension of ILRMA for Overdetermined and Reverberant Recording

In the discussions so far, the linear instantaneous mixture in frequency domain, (2.7), is always assumed to be valid. Figure 3.32 (a) shows the mixing system corresponding to (2.7), which is also called *linear time-invariant mixing*. In this mixing system, all the time frames are independent of other time frames, meaning that they do not affect each other. However, for the case of reverberant recording, reverberant components can leak from the previous frame as shown in Fig. 3.32 (b), and the mixed signal $\boldsymbol{x}_{ij}$ cannot be represented using only $\boldsymbol{A}_i$. Therefore, the assumption of linear time-invariant mixing holds only when the lengths of all impulse responses between the sources and microphones are sufficiently shorter than the length of the analysis window in STFT.

For this reason, in this section, I propose an extended algorithm of ILRMA particularly for overdetermined ($M > N$) and reverberant recordings. Since conventional ICA-based methods and ILRMA exploit the assumption of linear time-invariant mixing, the separation performance degrades when this assumption does not hold. The proposed new approach for overdetermined BSS enables us to achieve good separation performance even for reverberant signals. The algorithm utilizes extra observations (channels) to estimate the reverberant components of each source [207].

3.7.1 PCA for Overdetermined BSS

When $M > N$, in a typical separation method using FDICA or IVA, PCA is applied in advance and the dimension of $\boldsymbol{x}_{ij}$ is reduced so that $M = N$. This preprocessing is performed with the expectation that the reverberant components in the observed signal are eliminated by the dimensionality reduction. Therefore, PCA is applied to make the assumption of linear time-invariant mixing (2.7) valid even in a reverberant environment. However, if the purpose of source separation is to obtain each source image including the reverberation, PCA degrades the separation performance by removing the reverberation components. Moreover, if the source powers in mixtures are unbalanced (e.g., music signals), PCA can

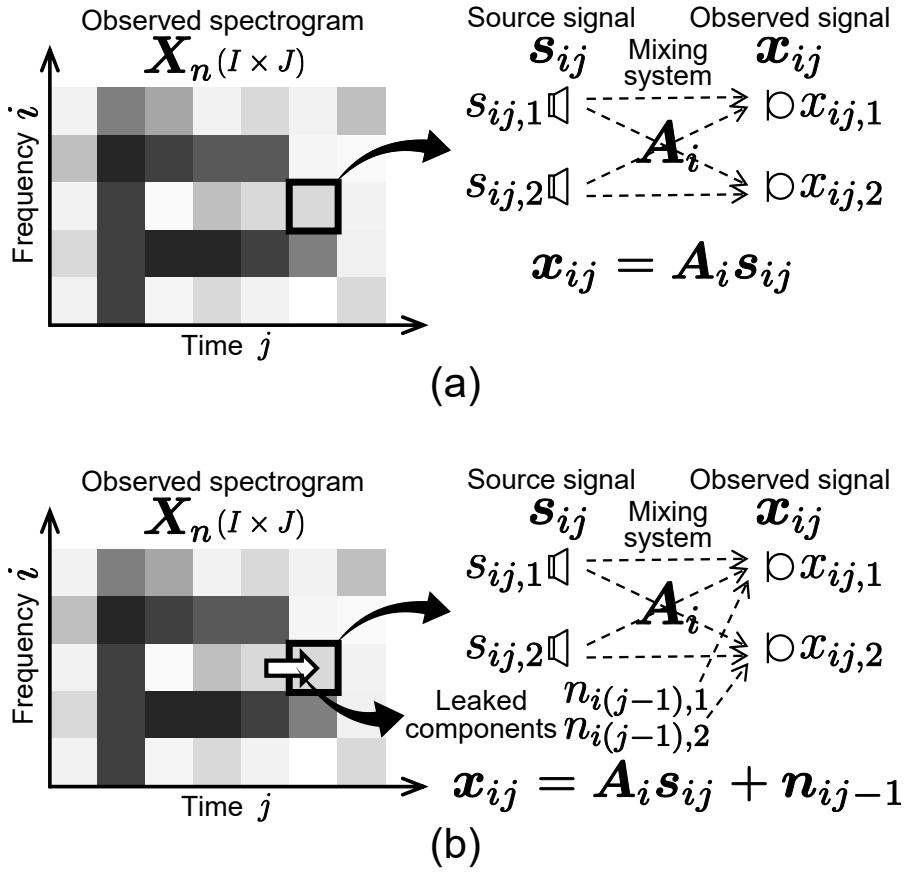

Figure 3.32: Mixing system of each spectrogram slot when $N = M = 2$; (a) has a linear time-invariant mixing system and there is no reverberation; (b) has some leaked components from the previous frame because of reverberation.

even remove direct components of weak sources, which leads to a greater risk of poor separation.

3.7.2 Relaxation of Rank-1 Spatial Model in ILRMA

To relax the constraint of the rank-1 spatial model in FDICA, IVA, and ILRMA, I propose the utilization of extra observations for modeling the reverberant components. In this method, we consider that the number of observations M is Q times the number of sources N, namely, $M = Q \times N$. In conventional

overdetermined BSS, PCA is applied before the separation so that M equals N as shown in Fig. 3.33 (a). In the proposed algorithm, we estimate M separated signals $\check{\boldsymbol{y}}$ as shown in Fig. 3.33 (b). In this approach, the leaked component from previous frames ($\boldsymbol{n}_{i(j-1)}$ in Fig. 3.32 (b)) of each source is modeled as an additional new source, namely, each original source is represented with rank-Q spatial model. To obtain an estimate of the source including both direct and reverberant components, the separated signals must be clustered using some criteria, which is a kind of permutation problem. The clustered separated signal $\check{\boldsymbol{y}}$ is represented as follows:

$$\check{\boldsymbol{y}}_{ij} = \left(\check{y}_{ij,11} \cdots \check{y}_{ij,1Q} \quad \check{y}_{ij,21} \cdots \check{y}_{ij,2Q} \cdots \check{y}_{ij,NQ}\right)^{\mathrm{T}} \tag{3.95}$$

$$\equiv \left(\check{y}_{ij,1} \cdots \check{y}_{ij,M}\right)^{\mathrm{T}}, \tag{3.96}$$

$$y_{ij,n} = \textstyle\sum_q \check{y}_{ij,nq}, \tag{3.97}$$

where $\check{y}_{ij,n1}, \ldots, \check{y}_{ij,nQ}$ correspond to the direct and reverberant components of one source n. Finally, each estimated source $y_{ij,n}$ is reconstructed by summing of the clustered components as represented by (3.97).

3.7.3 Clustering with Spectral Correlations

In Sect. 3.7.2, the complex-valued spectrograms of the sources are estimated by assuming the independence between them. However, we can expect that the power spectrograms of the direct and the reverberant components for the same source have a correlation. Based on this assumption, I propose to use cross-correlation between the power spectrograms $\check{p}_{ij,m} = |\check{y}_{ij,m}|^2$ to determine which separated signal $\check{y}_{ij,nq}$ corresponds to the direct or reverberant component of which source:

$$\mathrm{cor}(\check{\boldsymbol{P}}_m \| \check{\boldsymbol{P}}_{m'}) = \max\left(\left\{\textstyle\sum_{i,j} \check{p}_{ij,m}\, \check{p}_{i(j+\tau),m'} \mid \tau = 0, 1, \ \cdots, \tau_{\max}\right\}\right), \tag{3.98}$$

where $\boldsymbol{P}_m$ ($\in \mathbb{R}_{\geq 0}^{I \times J}$) is the power spectrograms whose element is $\check{p}_{ij,m}$ and τ is an index of the delay in the time frame. For clustering, I first calculate (3.98) between all separated signals $\check{y}_{ij,m}$. Then, the signals are merged in descending

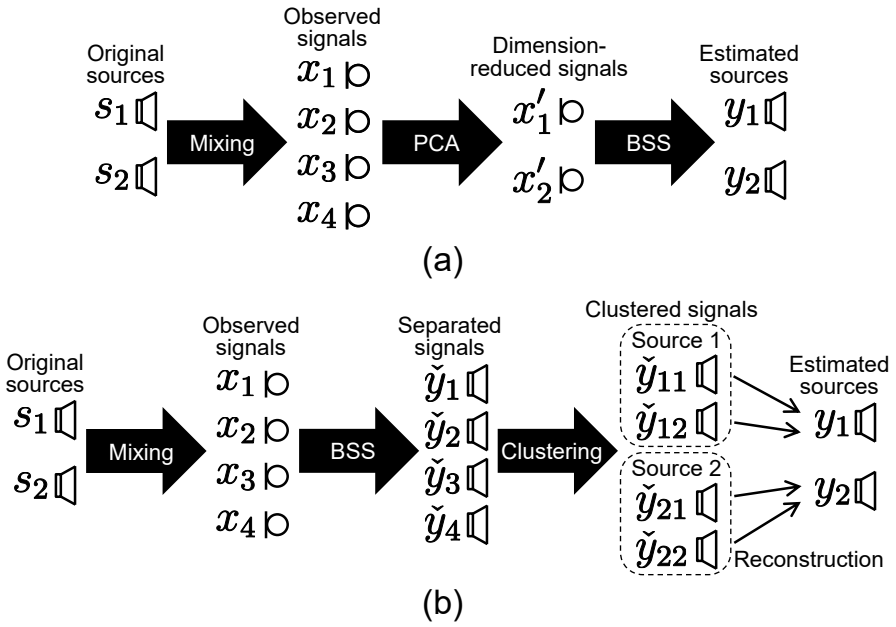

Figure 3.33: Algorithms of (a) conventional and (b) proposed methods ($N = 2$, $M = 4$, and $Q = 2$), where subscripts for i and j are omitted.

order of cor until the number of clusters becomes N, with all the clusters (signal sets) required to have the same number of signals (see Fig. 3.34).

3.7.4 Auto-Clustering with Basis-Shared ILRMA

For ILRMA, we can consider another approach for clustering the signals $\check{y}_{ij,m}$ into N sources. Since the reverberation consists of a sum of time-delayed versions of the direct component, it is represented by the convolution. Even in the power spectrogram domain, this model is approximately valid [208]. If we assume that the impulse response in the power spectrogram domain is identical over all frequency bins, the direct and reverberant components of the same source can be modeled by the same bases $\boldsymbol{T}_n$ (spectral patterns) and different activations $\boldsymbol{V}_{nq}$ (time-varying gains) as follows:

$$\check{\boldsymbol{P}}_{n1} \simeq \boldsymbol{T}_n \boldsymbol{V}_{n1}, \quad \check{\boldsymbol{P}}_{n2} \simeq \boldsymbol{T}_n \boldsymbol{V}_{n2}, \quad \cdots, \quad \check{\boldsymbol{P}}_{nQ} \simeq \boldsymbol{T}_n \boldsymbol{V}_{nQ}, \tag{3.99}$$

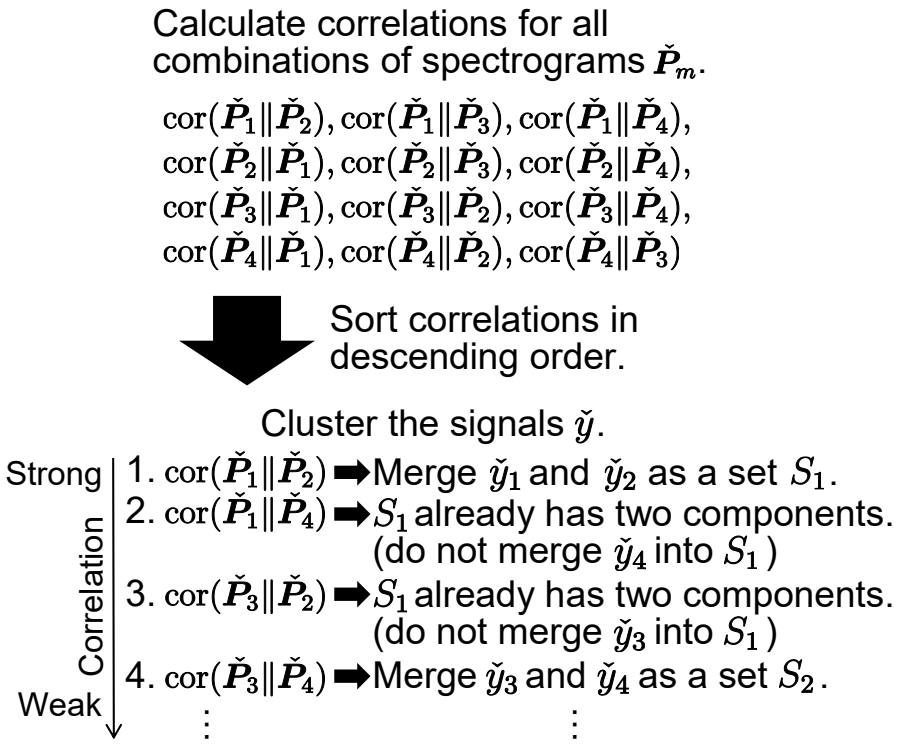

Figure 3.34: Hierarchical clustering using correlation cor ($N = 2$, $M = 4$, and $Q = 2$), where all sets must have the same number of signals and subscripts for i and j are omitted..

where $\check{\boldsymbol{P}}_{nq} (\in \mathbb{R}_{\geq 0}^{I\times J})$ is the power spectrogram of signal $\check{y}_{ij,nq}$, $\boldsymbol{T}_n$ $(\in \mathbb{R}_{\geq 0}^{I\times L})$ is a shared basis matrix whose elements are $t_{i1,n}, \cdots, t_{iL,n}$, and $\boldsymbol{V}_{nq}$ $(\in \mathbb{R}_{\geq 0}^{L\times J})$ is an activation matrix whose elements are $v_{1j,nq}, \cdots, v_{Lj,nq}$. This basis sharing leads to the separated signals $\check{y}_{ij,n1}, \ldots, \check{y}_{ij,nQ}$ representing the direct and reverberant components of one source n. The cost function of basis-shared ILRMA (BSILRMA) can be defined as [207]

$$\mathcal{L}_{\mathrm{BSILRMA}} = \mathrm{const.} - 2J \sum_i \log|\det \boldsymbol{W}_i| + \sum_{i,j,n,q} \left(\log \sum_l t_{il,n} v_{lj,nq} + \frac{|y_{ij,n}|^2}{\sum_l t_{il,n} v_{lj,nq}} \right). \tag{3.100}$$

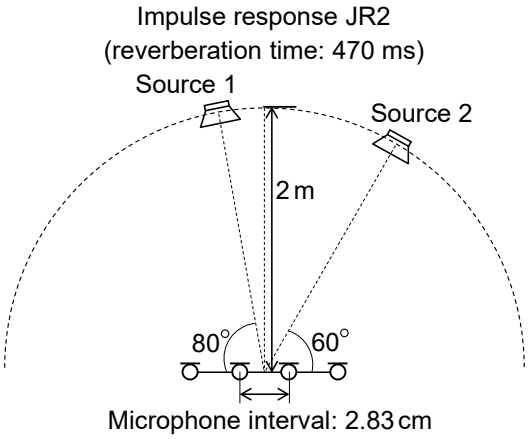

Figure 3.35: Recording condition of impulse response used in experiment of reverberant signals.

The update rules of $\boldsymbol{W}_i$ for minimizing (3.100) are the same as (3.50)–(3.52) if we consider $N \leftarrow M = NQ$ and $r_{ij,n} = \sum_l t_{il,n} v_{lj,nq}$, and the update rules of the NMF variables are obtained as follows:

$$t_{il,n} \leftarrow t_{il,n} \sqrt{\frac{\sum_{j,q} |y_{ij,nq}|^2 v_{lj,nq} \left(\sum_{l'} t_{il',n} v_{l'j,nq}\right)^{-2}}{\sum_{j,q} v_{lj,nq} \left(\sum_{l'} t_{il',n} v_{l'j,nq}\right)^{-1}}}, \tag{3.101}$$

$$v_{lj,nq} \leftarrow v_{lj,nq} \sqrt{\frac{\sum_{i} |y_{ij,nq}|^2 t_{il,n} \left(\sum_{l'} t_{il',n} v_{l'j,nq}\right)^{-2}}{\sum_{i} t_{il,n} \left(\sum_{l'} t_{il',n} v_{l'j,nq}\right)^{-1}}}. \tag{3.102}$$

However, the clustering result fluctuates depending on the initial values of the variables. To avoid this problem, I used IVA and the clustering method described in Sect. 3.7.3 to obtain initial value of demixing matrix $\boldsymbol{W}_i$.

3.7.5 Experiments and Results

Conditions

To confirm the efficacy of the proposed algorithm, similar to Sect. 3.6.3, I conducted an evaluation experiment using professional music signals. In this experiment, I produced observed signals with $M = 4$ channels and $N = 2$ sources

Table 3.12: Music sources used in experiment of reverberant signals

ID	Song name	Source (1/2)
1	bearlin-roads	acoustic_guit_main/piano
2	tamy-que_pena_tanto_faz	guitar/vocals
3	fort_minor-remember_the_name	drums/vocals
4	ultimate_nz_tour	guitar/vocals

Table 3.13: Characteristics of each method used in experiment of reverberant signals

Method	Number of filters per source	Postfilter
PCA+IVA	1	None
PCA+ILRMA	1	None
Sawada's MNMF w/o MWF	1	None
Sawada's MNMF	1	MWF
Ideal linear filter	1	None
Proposed IVA	2	None
Proposed BSILRMA	2	None

by convoluting the impulse response JR2 (see Fig. 3.35) [202] with each source. Table 3.12 shows the songs and sources used in this experiment, which were obtained from SiSEC2011 [122]. I compared IVA with PCA (*PCA+IVA*) and ILRMA with PCA (*PCA+ILRMA*), which both assume the linear time-invariant mixing (rank-1 spatial model). In addition, two types of Sawada's MNMFs [84] were also evaluated: *Sawada's MNMF w/o MWF* and *Sawada's MNMF*. In Sawada's MNMF w/o MWF, the maximum SNR BF [22], which is calculated from the estimated spatial covariance $\boldsymbol{R}_{i,n}^{(s)}$, was used for separation. Sawada's MNMF is the same method proposed in [84], which utilizes MWF [197] to enhance the estimated sources. As the proposed methods, spatial-model-relaxed IVA with the clustering method in Sect. 3.7.3 (*Proposed IVA*) and spatial-model-relaxed BSILRMA (*Proposed BSILRMA*) were evaluated, where the pretrained and clustered demixing matrix was used for the initial value in Proposed BSILRMA. Moreover, I evaluated the limit separation performance of linear filtering (*Ideal*

Table 3.14: Experimental conditions used in experiment of reverberant signals

Sampling frequency	Downsampled from 44.1 kHz to 16 kHz
Window length in STFT	128 ms
Window function	Hamming window
Window shift length	64 ms
Number of bases	$L=15$ ($K=30$)
Maximum delay in time frame	$\tau_{\max}=2$
Number of iterations	200

linear filter) as a reference performance, which is the maximum SNR BF calculated using the ideal (oracle) spatial covariances of each source. It is necessary to apply the back-projection technique (3.14), except for in Sawada's MNMF, to the estimated sources. The characteristics of each method are shown in Table 3.13 and the other conditions are described in Table 3.14. Note that I used a 128-ms-long window in the STFT for the signals that have 470-ms-long reverberation, which means that the rank-1 spatial model collapses. As the evaluation scores, I used the SDR improvement.

Results

Figures 3.36–3.39 shows the average scores and their deviations in 10 trials with different pseudorandom seeds. The methods using PCA cannot achieve good separation because they require the assumption of rank-1 spatial model. The scores of Sawada's MNMF w/o MWF indicate poor separation accuracy and strong dependence on the initial values. However, MWF with NMF variables (the scores of Sawada's MNMF) can greatly enhance the estimated sources. Proposed BSILRMA separates the sources with high accuracy. In particular, this method outperforms the limit performance of linear filtering (Ideal linear filter) as shown in Figs. 3.38 and 3.39. This is because ground truth sources include reverberations, which can span more than two dimensional space, and the proposed algorithm can effectively relax the constraint in rank-1 spatial model. However, in Fig. 3.37, the proposed methods cannot separate sources because guitar and vocals are split into one and three components, and the clustering

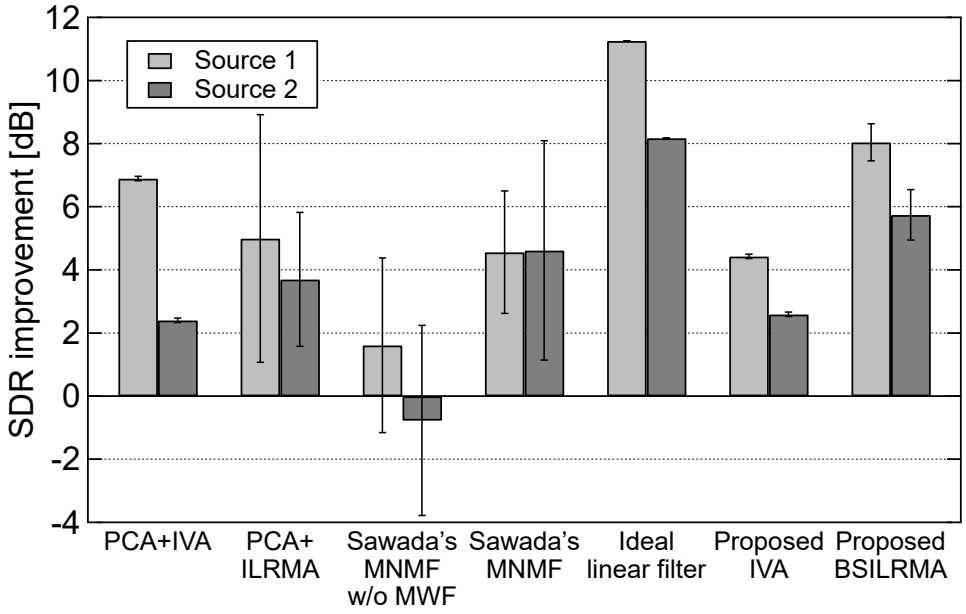

Figure 3.36: Average SDR improvements for song ID1 used in experiment of reverberant signals.

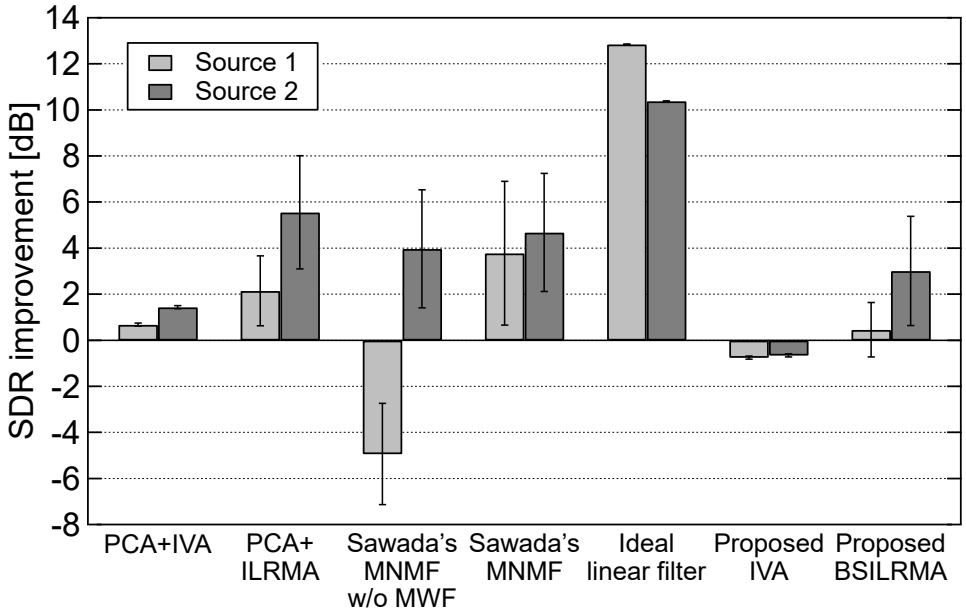

Figure 3.37: Average SDR improvements for song ID2 used in experiment of reverberant signals.

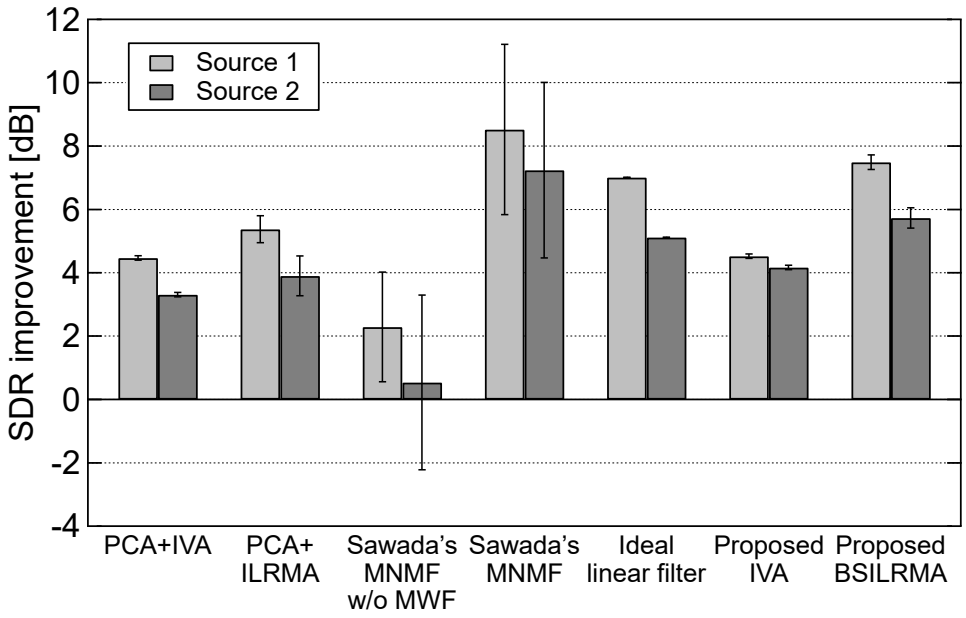

Figure 3.38: Average SDR improvements for song ID3 used in experiment of reverberant signals.

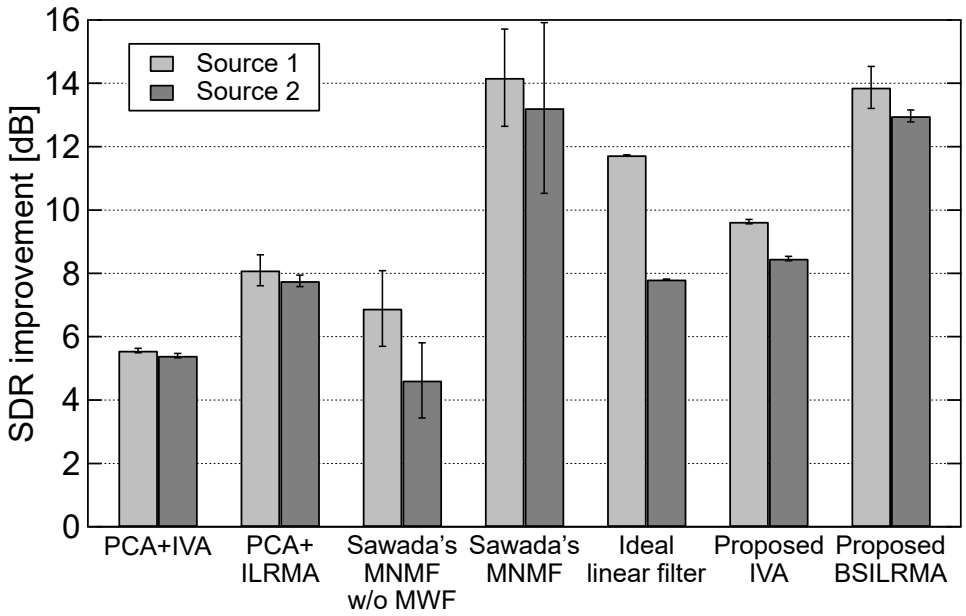

Figure 3.39: Average SDR improvements for song ID4 used in experiment of reverberant signals.

Table 3.15: Computational times for separation of song ID3 used in experiment of reverberant signals (s)

PCA+IVA	PCA+ILRMA	Sawada's MNMF	Proposed IVA	Proposed BSILRMA
23.4	29.4	3611.8	60.1	143.9

cannot divide these components as the individual sources.

Table 3.15 shows actual computational times for the separation of song ID3, where the calculations were performed using the same environment as the results in Table 3.10. The computational time of Proposed BSILRMA includes the initialization time for $\boldsymbol{W}_i$, which is the same as that of Proposed IVA. We confirm that Proposed BSILRMA can maintain efficient optimization and achieve good separation performance.

3.8 Summary

In this chapter, I proposed a new efficient determined BSS technique, ILRMA, that extends a source model in IVA from a vector to a low-rank matrix using the NMF representation. Also, the relationship between conventional MNMF and IVA was revealed: ILRMA is equivalent to MNMF with a rank-1 spatial model, and time-varying Gaussian IVA can be thought of as a special case of ILRMA, namely, ILRMA can be thought of as IVA with increased flexibility of the model. ILRMA can be optimized using fast update rules based on the auxiliary function technique. The experimental results show that ILRMA achieves faster convergence and better results than the conventional BSS techniques. In addition, an extension of ILRMA for overdetermined BSS with reverberant observation was proposed. This algorithm utilizes extra observations for simultaneously modeling both direct and reverberant components in each source. This method can be considered as an effective relaxation of the constraint in rank-1 spatial model. Owing to the source modeling using extra observations, the proposed method can exceed a limit performance of linear separation filter

for the reverberant signals in some cases.

4

Single-Channel Semi-Supervised Source Separation Based on Discriminative Nonnegative Matrix Factorization

4.1 Introduction

In this chapter, I address the single-channel supervised source separation, and propose a new algorithm for discriminative basis training in a semi-supervised situation. First, I review some existing methods for single-channel source separation and explain a problem in supervised NMF methods, which is related to discriminancy of supervised NMF bases. Next, conventional approached for discriminative basis training in a full-supervised situation are reviewed,

then the motivation and strategy for discriminative training in semi-supervised situation are clarified. After giving an explanation of the proposed algorithm, its performance of music source separation is experimentally confirmed. Finally, the whole contents in this chapter are summarized.

4.2 Existing NMF-Based Single-Channel Source Separation

A single-channel source separation task is one the most important problem in acoustic signal processing field because it can be applied to a front-end system for almost all acoustic applications. In contrast to the source separation for multichannel signals, in this problem, spatial information cannot be used, and a clue for solving the separation is only the difference of spectral features in each source, e.g., spectral patterns, spectral envelopes (linear prediction cepstral coefficients [209, 210, 211]), and mel-frequency cepstrum coefficients (MFCC) [212, 213]. The problem can be formulated as a clustering or classification of these extracted features.

In the past decade, NMF becomes the most popular approach for single-channel source separation. Many techniques based on NMF have been proposed and investigated so far, and it is still a growing research topic. In [214, 68, 70], some regularizations including sparseness, temporal continuity (smoothness), lower complexity, and better predictability, were introduced to the cost function in NMF. These regularized cost function can be derived from a maximum a posteriori estimation with a prior distribution of parameters, and it was generalized as NMF with Bayesian estimation in [72]. We may solve the separation problem only when such prior models fit to the inherent nature of the sources. As another approach, in some literatures, the structures of sources were modeled with NMF decomposition. In [67, 75], shifted NMF was proposed. They assumed that the timbre of a musical note produced by the same instrumental source is constant for the entire range of pitch, and the sourcewise NMF basis can be shifted to represent all the notes of the same source. Similar idea was introduced in [124], but they utilized average spectral patterns for the

clustering criteria of NMF bases. Also, in [73, 128, 76], a source filter model or MFCC and its estimation were exploited for the clustering. In [69], the NMF bases were extended to have two dimensions (frequency bins × very short time frames) as spectral fragments. Since such two-dimensional bases should be convolved to approximately represent the observed matrix, this method is called nonnegative matrix factor deconvolution. In [71, 74, 126, 127], a Markov chain model was introduced to NMF bases. These methods mainly aim to model the temporal (continuous) fluctuation or variation. For example, a piano note is accurately characterized by a succession of several spectral patterns corresponding to "attack," "sustain," "decay," and "release" segments. Also, a musical vibrato can simply be represented by the Markov chain model.

To learn these sourcewise structures, patterns, or natures, a supervised approach is very effective and has a big potential to achieve better separation performance. The simplest way of supervised method is to prepare the sourcewise basis matrix, which can be obtained by independently applying simple NMF to the training signal for each source. This approach is called supervised NMF [7, 131, 98], and for music signals, sample sequential notes (tones) with wide range of pitch (e.g., two or more octaves) can be a good training signal for the instrumental sources. In particular, supervised NMF using training signals for all the sources is called FSNMF, and the other approach (e.g., preparing the training signals for only the target source) is called SSNMF. The detailed algorithms of these methods are described in the following section.

4.3 Conventional Supervised NMF and Discriminative Training of Supervised Bases

4.3.1 Conventional Supervised NMF

The algorithms in FSNMF and SSNMF are depicted in Fig. 4.1. In FSNMF, we prepare the supervised basis matrix for all the sources in the observed mixture. Let $\boldsymbol{\Delta} \in \mathbb{R}_{\geq 0}^{\Phi \times \Psi}$ be a power or an amplitude spectrogram that includes two sources, $\boldsymbol{S}_1 \in \mathbb{R}_{\geq 0}^{\Phi \times \Psi}$ and $\boldsymbol{S}_2 \in \mathbb{R}_{\geq 0}^{\Phi \times \Psi}$. Here, note that two mixture spectrograms $\boldsymbol{\Delta}$

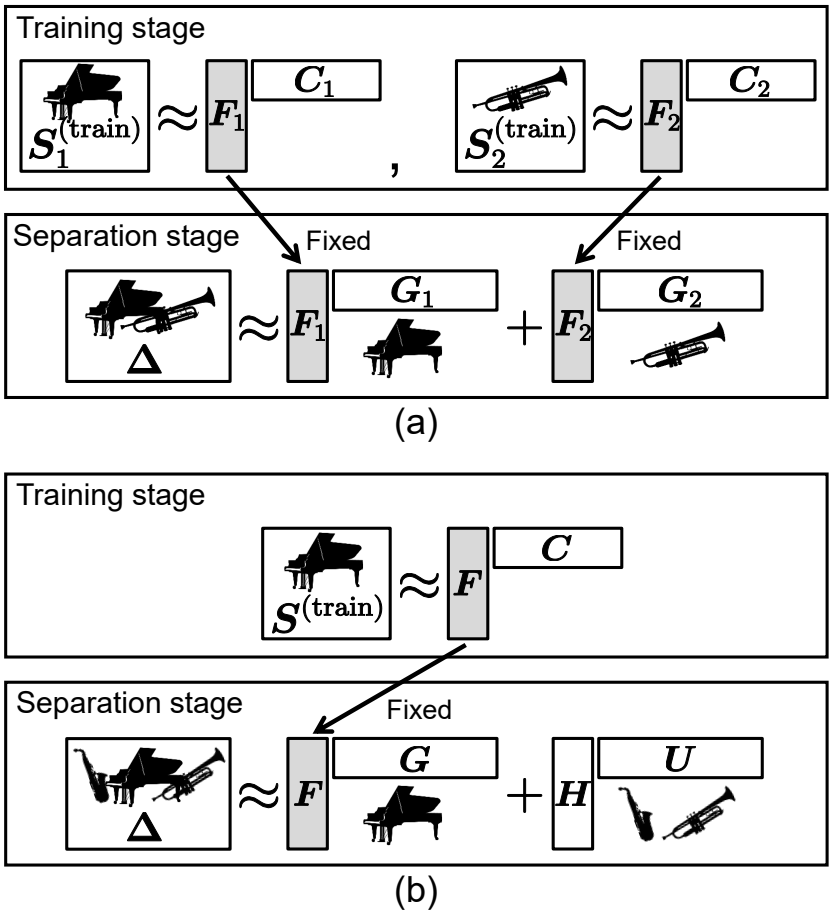

Figure 4.1: Training and separation stages in (a) FSNMF and (b) SSNMF.

and $\boldsymbol{S}_1 + \boldsymbol{S}_2$ are generally not identical because they are the amplitude or power spectrograms, and the addition $\boldsymbol{S}_1 + \boldsymbol{S}_2$ do not take the phase cancellation into account. $\boldsymbol{\Delta}$ is the amplitude or the power spectrogram of the waveform that is an addition of two time-domain waveforms of $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. Now, we prepare the training signals for both sources as $\boldsymbol{S}_1^{(\text{train})} \in \mathbb{R}_{\geq 0}^{\Phi \times \Psi_1}$ and $\boldsymbol{S}_2^{(\text{train})} \in \mathbb{R}_{\geq 0}^{\Phi \times \Psi_2}$. In the training stage, we estimate the sourcewise supervised basis matrices $\boldsymbol{F}_1 \in \mathbb{R}_{\geq 0}^{\Phi \times K_1}$ and $\boldsymbol{F}_2 \in \mathbb{R}_{\geq 0}^{\Phi \times K_2}$ (dictionaries of spectral patterns for each source) by performing

the following NMFs:

$$(\boldsymbol{F}_1, \boldsymbol{C}_1) = \underset{\boldsymbol{F}_1,\, \boldsymbol{C}_1}{\arg\min}\, \mathcal{D}_\beta\left(\boldsymbol{S}_1^{(\text{train})} \| \boldsymbol{F}_1\boldsymbol{C}_1\right), \tag{4.1}$$

$$(\boldsymbol{F}_2, \boldsymbol{C}_2) = \underset{\boldsymbol{F}_2,\, \boldsymbol{C}_2}{\arg\min}\, \mathcal{D}_\beta\left(\boldsymbol{S}_2^{(\text{train})} \| \boldsymbol{F}_2\boldsymbol{C}_2\right), \tag{4.2}$$

where these minimization are carried out under a nonnegative constraint for all the variables (hereafter, all the minimizations in this chapter include the nonnegative constraint). The activation matrices $\boldsymbol{C}_1 \in \mathbb{R}_{\geq 0}^{K_1 \times \Psi_1}$ and $\boldsymbol{C}_2 \in \mathbb{R}_{\geq 0}^{K_2 \times \Psi_2}$ can be discarded after the training stage. In the separation stage, we decompose the mixture $\boldsymbol{\Delta}$ using fixed $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ as

$$(\boldsymbol{G}_1, \boldsymbol{G}_2) = \underset{\boldsymbol{G}_1,\, \boldsymbol{G}_2}{\arg\min}\, \mathcal{D}_\beta(\boldsymbol{\Delta} \| \boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{F}_2\boldsymbol{G}_2). \tag{4.3}$$

Therefore, we expect that the sources in $\boldsymbol{\Delta}$ will be separated by the estimated activation matrices $\boldsymbol{G}_1 \in \mathbb{R}_{\geq 0}^{K_1 \times \Psi}$ and $\boldsymbol{G}_2 \in \mathbb{R}_{\geq 0}^{K_2 \times \Psi}$ as $\boldsymbol{S}_1 \approx \boldsymbol{F}_1\boldsymbol{G}_1$ and $\boldsymbol{S}_2 \approx \boldsymbol{F}_2\boldsymbol{G}_2$. Similarly to (2.24), the update rules for these activations can be derived from the minimization of (4.3) as

$$\boldsymbol{G}_1 \leftarrow \boldsymbol{G}_1 \circ \left\{ \frac{\boldsymbol{F}_1^{\mathrm{T}} \left[\boldsymbol{\Delta} \circ (\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{F}_2\boldsymbol{G}_2)^{\cdot -2}\right]}{\boldsymbol{F}_1^{\mathrm{T}} (\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{F}_2\boldsymbol{G}_2)^{\cdot -1}} \right\}^{\cdot \varphi(\beta)}, \tag{4.4}$$

$$\boldsymbol{G}_2 \leftarrow \boldsymbol{G}_2 \circ \left\{ \frac{\boldsymbol{F}_2^{\mathrm{T}} \left[\boldsymbol{\Delta} \circ (\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{F}_2\boldsymbol{G}_2)^{\cdot -2}\right]}{\boldsymbol{F}_2^{\mathrm{T}} (\boldsymbol{F}_1\boldsymbol{G}_1 + \boldsymbol{F}_2\boldsymbol{G}_2)^{\cdot -1}} \right\}^{\cdot \varphi(\beta)}. \tag{4.5}$$

In SSNMF, we only focus on the extraction of the target source from the mixture. Similarly to FSNMF, the supervised basis matrix $\boldsymbol{F} \in \mathbb{R}_{\geq 0}^{\Phi \times K_{\mathrm{T}}}$ is obtained using a training signal of the target source $\boldsymbol{S}^{(\text{train})} \in \mathbb{R}_{\geq 0}^{\Phi \times \Psi'}$ as

$$(\boldsymbol{F}, \boldsymbol{C}) = \underset{\boldsymbol{F},\, \boldsymbol{C}}{\arg\min}\, \mathcal{D}_\beta\left(\boldsymbol{S}^{(\text{train})} \| \boldsymbol{F}\boldsymbol{C}\right). \tag{4.6}$$

In the separation stage, $\boldsymbol{\Delta}$ is decomposed using fixed $\boldsymbol{F}$ as

$$(\boldsymbol{G}, \boldsymbol{H}, \boldsymbol{U}) = \underset{\boldsymbol{G},\, \boldsymbol{H},\, \boldsymbol{U}}{\arg\min}\, \mathcal{D}_{\beta}(\boldsymbol{\Delta} \| \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}), \tag{4.7}$$

where $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{\Phi \times K_{\mathrm{N}}}$ and $\boldsymbol{U} \in \mathbb{R}_{\geq 0}^{K_{\mathrm{N}} \times \Psi}$ are the basis and activation matrices for the non-target sources. Thus, the target source and non-target sources are ideally separated into $\boldsymbol{F}\boldsymbol{G}$ and $\boldsymbol{H}\boldsymbol{U}$, respectively. The update rules for $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ are obtained as follows:

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \circ \left\{ \frac{\boldsymbol{F}^{\mathrm{T}} \left[\boldsymbol{\Delta} \circ (\boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U})^{\cdot -2} \right]}{\boldsymbol{F}^{\mathrm{T}} (\boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U})^{\cdot -1}} \right\}^{\cdot \varphi(\beta)}, \tag{4.8}$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \circ \left\{ \frac{\left[\boldsymbol{\Delta} \circ (\boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U})^{\cdot -2} \right] \boldsymbol{U}^{\mathrm{T}}}{(\boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U})^{\cdot -1} \boldsymbol{U}^{\mathrm{T}}} \right\}^{\cdot \varphi(\beta)}, \tag{4.9}$$

$$\boldsymbol{U} \leftarrow \boldsymbol{U} \circ \left\{ \frac{\boldsymbol{H}^{\mathrm{T}} \left[\boldsymbol{\Delta} \circ (\boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U})^{\cdot -2} \right]}{\boldsymbol{H}^{\mathrm{T}} (\boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U})^{\cdot -1}} \right\}^{\cdot \varphi(\beta)}. \tag{4.10}$$

General NMF-based source separation including FSNMF and SSNMF does not ensure the physically accurate signal decomposition because the additivity of the amplitude or the power spectrograms is not valid and the NMF decomposition is an approximation. Also, the phase information of each source cannot be obtained, and the observed phase (noisy phase spectrogram) is often utilized to perform an inverse STFT. However, even though there are such drawbacks, FSNMF and SSNMF are still a powerful method and are used for many situations.

As another approach for obtaining the supervised basis matrices, exemplar-based NMF was proposed [215, 216, 217]. In this method, every supervised basis corresponds to an observation in the training data, namely, the training stage is just a sampling of several spectra from different time frames for each source. This approach becomes popular for a large-scale training of audio signals because it does not require a large computational cost in the training stage. However, for the music signals, it is hard to prepare such large-scale training dataset for all the possible instruments and vocals. Thus, the separation technique that requires

only the small dataset is desired for supervised music source separation.

4.3.2 Drawback in Supervised NMF and Motivation for Discriminative Basis Training

These supervised methods have the potential to achieve high separation performance. However, in both FSNMF and SSNMF, the cost function (4.3) or (4.7) in the separation stage represents how well the NMF model approximates $\mathbf{\Lambda}$, which does not include any criteria regarding the degree of separation, and the pretrained supervised bases may represent not only the relevant source but also a part of the other sources. For example, in SSNMF, $\boldsymbol{F}$ may represent a part of the non-target source spectra that should originally be absorbed in $\boldsymbol{H}$, and this phenomenon markedly degrades the separation quality. The dominant cause of this problem is that the supervised bases are independently trained using isolated training signals of each source as (4.1) and (4.1), or (4.6), and there may be some spectral overlaps between these training signals even if they are inherently different sources.

In the context of FSNMF, several methods have been proposed to solve this problem [134, 135, 8, 9, 10]. In [134], the cross-coherence of the bases was added to the cost function. In [135, 8, 9, 133], a sample mixture signal obtained by mixing the sample sounds of each source was utilized in the training stage to estimate more discriminative sourcewise basis matrices. Also, the authors in [10] proposed a joint optimization of NMF and a classification problem, where the NMF variables are optimized so that each basis is classified into one source during the training. Since these algorithms aim to estimate or train a discriminative supervised bases, they are called *discriminative NMF*. However, the conventional approaches are only applicable to the full-supervised situation because it is difficult to train such discriminative bases in semi-supervised situation. In SSNMF, a penalized SSNMF that forces the non-target bases $\boldsymbol{H}$ to be different from the target bases $\boldsymbol{F}$ was proposed [218, 98][1], but the target bases $\boldsymbol{F}$ may still represent non-target source components. For this reason,

[1] After [218] was submitted, a similar penalty was independently proposed in [134] for FSNMF, but [218, 98] also include another type of penalty.

in this dissertation, I address the discriminative training of supervised basis that is applicable not only for FSNMF but also for SSNMF, and propose a new algorithm for achieving the same objective in the conventional discriminative NMF.

4.3.3 Algorithm of Discriminative Basis Training for FSNMF

We here review the discriminative NMF in the full-supervised situation. As discriminative basis training for FSNMF, the following bilevel optimization [219] has been proposed [8, 9] (in which a two-source case was considered):

$$(\boldsymbol{C}_1^*, \boldsymbol{C}_2^*) = \underset{\boldsymbol{C}_1,\, \boldsymbol{C}_2}{\arg\min}\, \mathcal{D}_\beta\left(\boldsymbol{S}_{\text{mix}}^{(\text{train})} \| \boldsymbol{F}_1\boldsymbol{C}_1 + \boldsymbol{F}_2\boldsymbol{C}_2\right) + \text{Reg}(\boldsymbol{C}_1, \boldsymbol{C}_2), \tag{4.11}$$

$$(\boldsymbol{F}_1^*, \boldsymbol{F}_2^*) = \underset{\boldsymbol{F}_1,\, \boldsymbol{F}_2}{\arg\min}\, \mathcal{D}_\beta\left(\boldsymbol{S}_1^{(\text{train})} \| \boldsymbol{F}_1\boldsymbol{C}_1^*\right) + \mathcal{D}_\beta\left(\boldsymbol{S}_2^{(\text{train})} \| \boldsymbol{F}_2\boldsymbol{C}_2^*\right), \tag{4.12}$$

where $\boldsymbol{S}_{\text{mix}}^{(\text{train})} \approx \boldsymbol{S}_1^{(\text{train})} + \boldsymbol{S}_2^{(\text{train})}$ (addition in the time domain) and $\text{Reg}(\boldsymbol{C}_1, \boldsymbol{C}_2)$ is a regularization term for the activations corresponding to, for example, sparseness criteria. Note that the cost (4.12) depends on the minimizers $\boldsymbol{C}_1^*$ and $\boldsymbol{C}_2^*$ of (4.11); thus, they are functions of $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$. Hence, (4.12) is a bilevel optimization problem since the basis matrices appear in both levels. This optimization finds the basis matrices by taking into account the reconstruction of each source, $\boldsymbol{S}_1^{(\text{train})}$ and $\boldsymbol{S}_2^{(\text{train})}$. Therefore, the obtained bases tend to be discriminative.

The authors in [9] also mentioned that the basis matrices $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ in (4.11) and (4.12) do not have to be the same, making the method in [9] a generalized discriminative NMF. However, it is more challenging to simultaneously optimize the different bases in (4.11) and (4.12). Thus, they simply used the independently trained $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ in (4.11) and obtained different $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ in (4.12).

4.4 New Algorithm for Discriminative Basis Training

4.4.1 Strategy

In the proposed method, I only focus on SSNMF. With a similar motivation to [9], I propose the training of two types of supervised basis matrix for one source without bilevel optimization because it is difficult to find supervised bases that have *reconstructive* and *discriminative* spectra simultaneously. To maximize the discrimination from other sources, the supervised bases should consist of unique spectral components of the target source, which ideally do not overlap with those of the other sources. For example, inharmonic components are significant cues in distinguishing piano spectra from other instrumental sounds. Such discriminative bases (hereafter, referred to as *discriminative bases* $\boldsymbol{F}'$) should only be used to estimate the activations $\boldsymbol{G}$ of the target source in the separation stage. After the separation, the target source can be reconstructed with $\boldsymbol{G}$ and the reconstructive bases (hereafter, referred to as *reconstructive bases* $\boldsymbol{F}$).

Figure 4.2 depicts the conceptual difference between conventional and proposed algorithms of separation. In this figure, the target source (colored in black) is subjected to interference by the non-target source (colored in gray), where the fundamental frequency components overlap. When the discriminative basis $\boldsymbol{F}'$ consists of unique components of the target source as shown in Fig. 4.2 (b), the discrimination (finding the correct activation $\boldsymbol{G}$ of the target source) becomes easier than in the case when the supervised basis consists of all the components of the target source. The fundamental components in the target source, which are missing in $\boldsymbol{F}'$, are represented by another non-target basis $\boldsymbol{H}_2$ in the NMF decomposition. Since the activation vectors $\boldsymbol{G}$ are correctly estimated in the proposed algorithm, we can reconstruct the target source using reconstructive basis $\boldsymbol{F}$ as $\boldsymbol{FG}$.

In conventional SSNMF, the supervised bases $\boldsymbol{F}$ for the target source $\boldsymbol{S}$ are trained using a sample sound $\boldsymbol{S}^{(\mathrm{train})}$ of the target source. However, to train the discriminative bases $\boldsymbol{F}'$, we here prepare a simulative mixture as $\boldsymbol{S}^{(\mathrm{train})}_{\mathrm{mix}} \approx \boldsymbol{S}^{(\mathrm{train})} + \boldsymbol{N}^{(\mathrm{train})}$, where $\boldsymbol{N}^{(\mathrm{train})}$ is a sample sound of the non-target

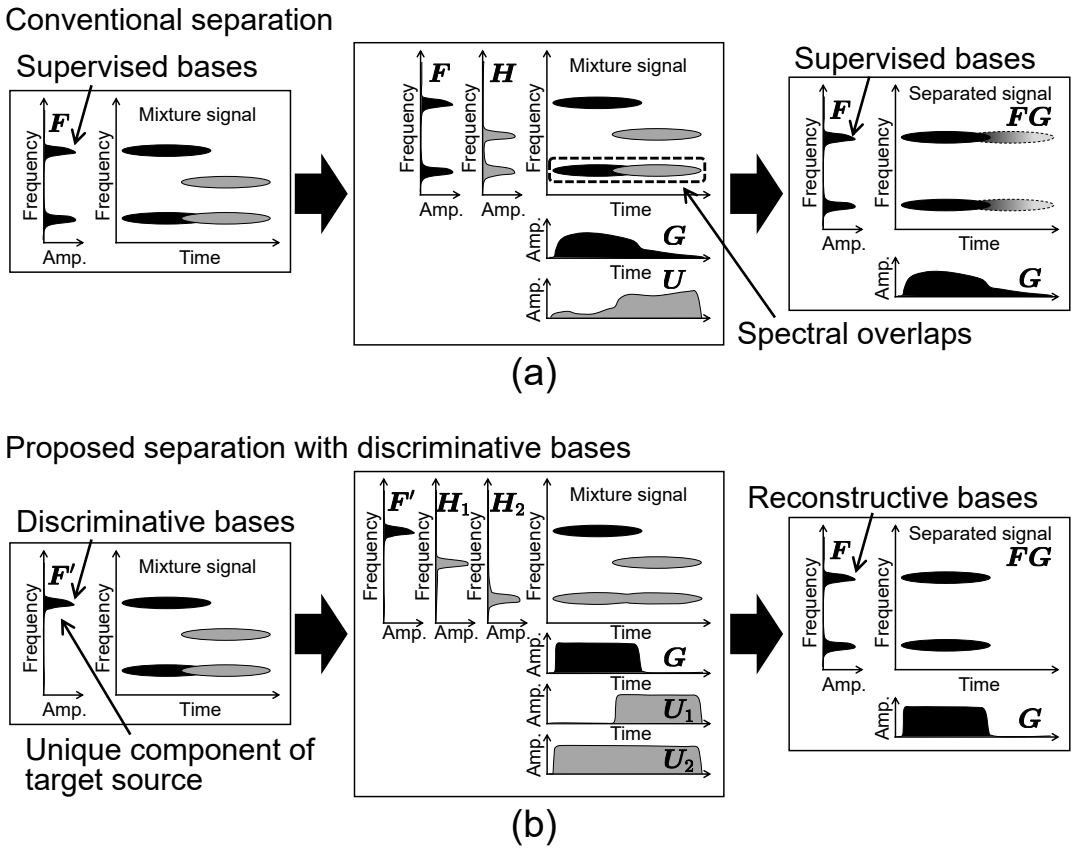

Figure 4.2: Difference between (a) conventional and (b) proposed algorithms of separation, where black components correspond to target source and gray components correspond to interfering source. Proposed method utilizes discriminative bases ($\boldsymbol{F}'$) that has unique component of target source for separation, and target source is synthesized using reconstructive bases ($\boldsymbol{F}$) that has complete spectral components.

sources and the addition is performed in the time domain. Although we do not know the non-target sources in the semi-supervised scenario, we can collect possible candidates such as different instruments for the target source. Note that the mixture $\boldsymbol{S}_{\mathrm{mix}}^{(\mathrm{train})}$ is utilized to train the discriminative bases $\boldsymbol{F}'$, namely, for the estimation of the unique component in the target source spectra, and the bases for $\boldsymbol{N}^{(\mathrm{train})}$ are never used in the separation stage.

4.4.2 Discriminative and Reconstructive Basis Training in SS-NMF

Since the role of the discriminative bases $\boldsymbol{F}'$ is to obtain the accurate activations, we would like to find $\boldsymbol{F}'$ such that the following $\boldsymbol{C}^*$ and $\boldsymbol{C}'$ become equivalent:

$$\boldsymbol{C}^* = \underset{\boldsymbol{F},\,\boldsymbol{C}}{\arg\min}\, \mathcal{D}_\beta\left(\boldsymbol{S}^{(\text{train})} \| \boldsymbol{F}\boldsymbol{C}\right), \tag{4.13}$$

$$\boldsymbol{C}' = \underset{\boldsymbol{C},\,\boldsymbol{T},\boldsymbol{V}}{\arg\min}\, \mathcal{D}_\beta\left(\boldsymbol{S}^{(\text{train})}_{\text{mix}} \| \boldsymbol{F}'\boldsymbol{C} + \boldsymbol{T}\boldsymbol{V}\right). \tag{4.14}$$

This is also a bilevel optimization problem, and it is challenging to optimize $\boldsymbol{F}'$ using this criterion. Instead, I here propose the following simple optimization for training both the discriminative and reconstructive bases $\boldsymbol{F}'$ and $\boldsymbol{F}$:

$$(\boldsymbol{F}, \boldsymbol{C}^*) = \underset{\boldsymbol{F},\,\boldsymbol{C}}{\arg\min}\, \mathcal{D}_\beta\left(\boldsymbol{S}^{(\text{train})} \| \boldsymbol{F}\boldsymbol{C}\right), \tag{4.15}$$

$$(\boldsymbol{F}', \boldsymbol{T}, \boldsymbol{V}) = \underset{\boldsymbol{F}',\,\boldsymbol{T},\,\boldsymbol{V}}{\arg\min}\, \mathcal{D}_\beta\left(\boldsymbol{S}^{(\text{train})}_{\text{mix}} \| \boldsymbol{F}'\boldsymbol{C}^* + \boldsymbol{T}\boldsymbol{V}\right), \tag{4.16}$$

where $\boldsymbol{T}$ and $\boldsymbol{V}$ are new NMF matrices representing the non-target sources included in $\boldsymbol{S}^{(\text{train})}_{\text{mix}}$. The reconstructive bases $\boldsymbol{F}$ and their activations $\boldsymbol{C}^*$ are obtained by (4.15). Then, we initialize the discriminative bases as $\boldsymbol{F}' \leftarrow \boldsymbol{F}$ and calculate the optimization (4.16) with fixed $\boldsymbol{C}^*$. Since $\boldsymbol{C}^*$ is the true activation of the target source in $\boldsymbol{S}^{(\text{train})}_{\text{mix}}$, $\boldsymbol{F}'$ is trained such that the activations of the target source obtained by NMF with $\boldsymbol{F}'$ applied to the mixture are as close to $\boldsymbol{C}^*$ as possible. In the separation stage, the mixture $\boldsymbol{\Lambda}$ is decomposed using fixed $\boldsymbol{F}'$ as

$$(\boldsymbol{G}, \boldsymbol{H}, \boldsymbol{U}) = \underset{\boldsymbol{G},\,\boldsymbol{H},\,\boldsymbol{U}}{\arg\min}\, \mathcal{D}_\beta(\boldsymbol{\Lambda} \| \boldsymbol{F}'\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}). \tag{4.17}$$

The update rules for (4.15)–(4.17) are the same as those described in Sects. 2.5 and 4.3.1 with the replacement of the corresponding variables. After the separation stage, the estimated target source $\hat{\boldsymbol{S}}$ can be reconstructed as $\boldsymbol{F}\boldsymbol{G}$ or

using Wiener filtering as

$$\hat{\boldsymbol{S}} \approx \frac{\boldsymbol{F}\boldsymbol{G}}{\boldsymbol{F}'\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}} \circ \boldsymbol{\Lambda}. \tag{4.18}$$

In [9], the authors refined the bases used for reconstruction after first fixing the bases used in the separation. In this paper, we refine the discriminative bases using the fixed reconstructive bases.

4.5 Experiments

4.5.1 Simple Experiment Using Piano and Flute Tones

In this subsection, I confirm how the proposed algorithm works using simple audio data of piano and flute tones provided by the MIDI tone generator *Garritan Personal Orchestra 4*. The sample sound $\boldsymbol{S}^{(\mathrm{train})}$ contains only a piano tone (C5) and $\boldsymbol{N}^{(\mathrm{train})}$ contains only a flute tone (C6). Figure 4.3 shows the NMF bases, which are independently trained for the piano and flute tones using a single basis.

After initializing $\boldsymbol{F}'$ as shown in Fig. 4.3 (a), I calculated (4.16) with $\boldsymbol{C}^*$ and randomized $\boldsymbol{T}$ and $\boldsymbol{V}$, where the number of bases in $\boldsymbol{T}$ was set to two. Figure 4.4 shows the estimated $\boldsymbol{F}'$ and the bases in $\boldsymbol{T}$ after 50 iterations of the update rules for optimizing (4.16). From Fig. 4.4 (a), we can confirm that spectral notches appear at the second, sixth, and eighth peaks in $\boldsymbol{F}'$ and do not overlap with the peaks in the flute tone, and such lost peak components are compensated by the other basis as shown in Fig. 4.4 (c). From these results, I consider that the proposed algorithm can train the unique components of the target source, as shown in Fig. 4.2, to some extent.

4.5.2 Music Source Separation

Conditions

We compare the separation performance between simple SSNMF and the proposed method in the music separation task. We used three songs and six

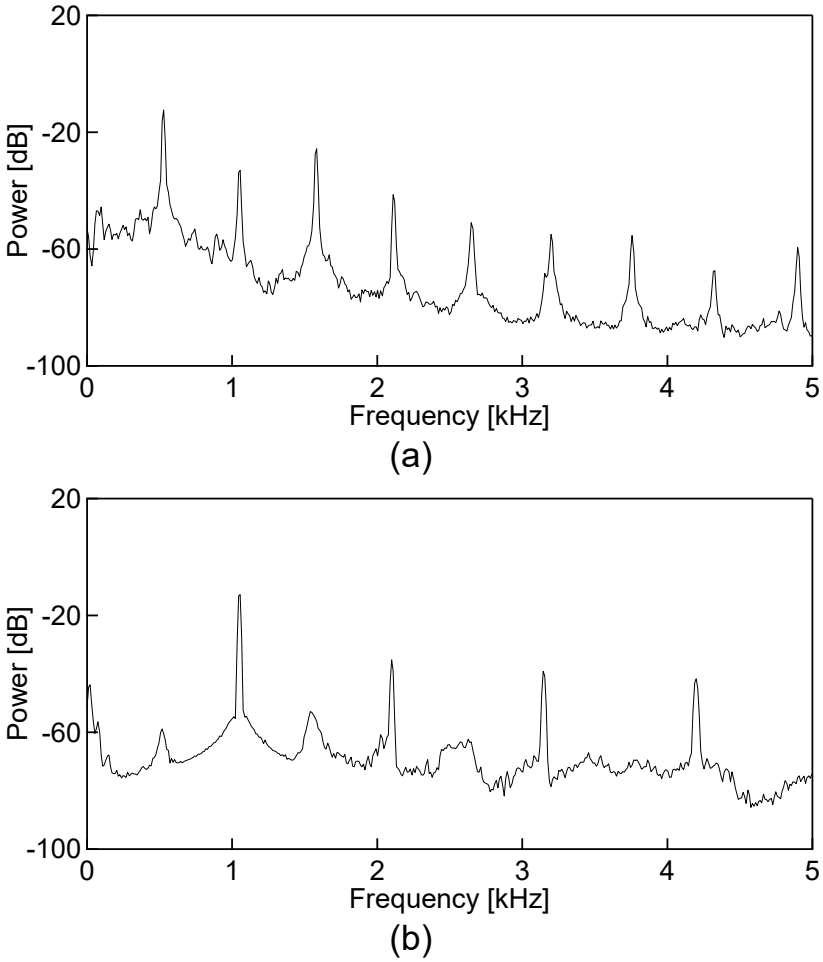

Figure 4.3: Spectral bases obtained from simple NMF: (a) C5 piano tone and (b) C6 flute tone.

mixtures obtained from SiSEC2011 [122] (see Table 4.1). Each song consists of two sources, and 4-fold cross-validation was applied to each song to obtain a training sound of the target source $\boldsymbol{S}^{(\text{train})}$ and a test mixture $\boldsymbol{\Lambda} \approx \boldsymbol{S}^{(\text{test})} + \boldsymbol{N}^{(\text{test})}$, where addition was performed in the time domain. More precisely, the training sound of the target source $\boldsymbol{S}^{(\text{train})}$ was obtained from three-quarters of a source in a song, and the remaining quarter $\boldsymbol{S}^{(\text{test})}$ was used to obtain a mixture $\boldsymbol{\Lambda}$. The sample sound of the non-target source $\boldsymbol{N}^{(\text{train})}$ was obtained from a different song as shown in Table 4.2. For example, in the ID3 data, the mixture $\boldsymbol{\Lambda}$ was part of the song "Que pena tanto faz" and include the classic guitar sound

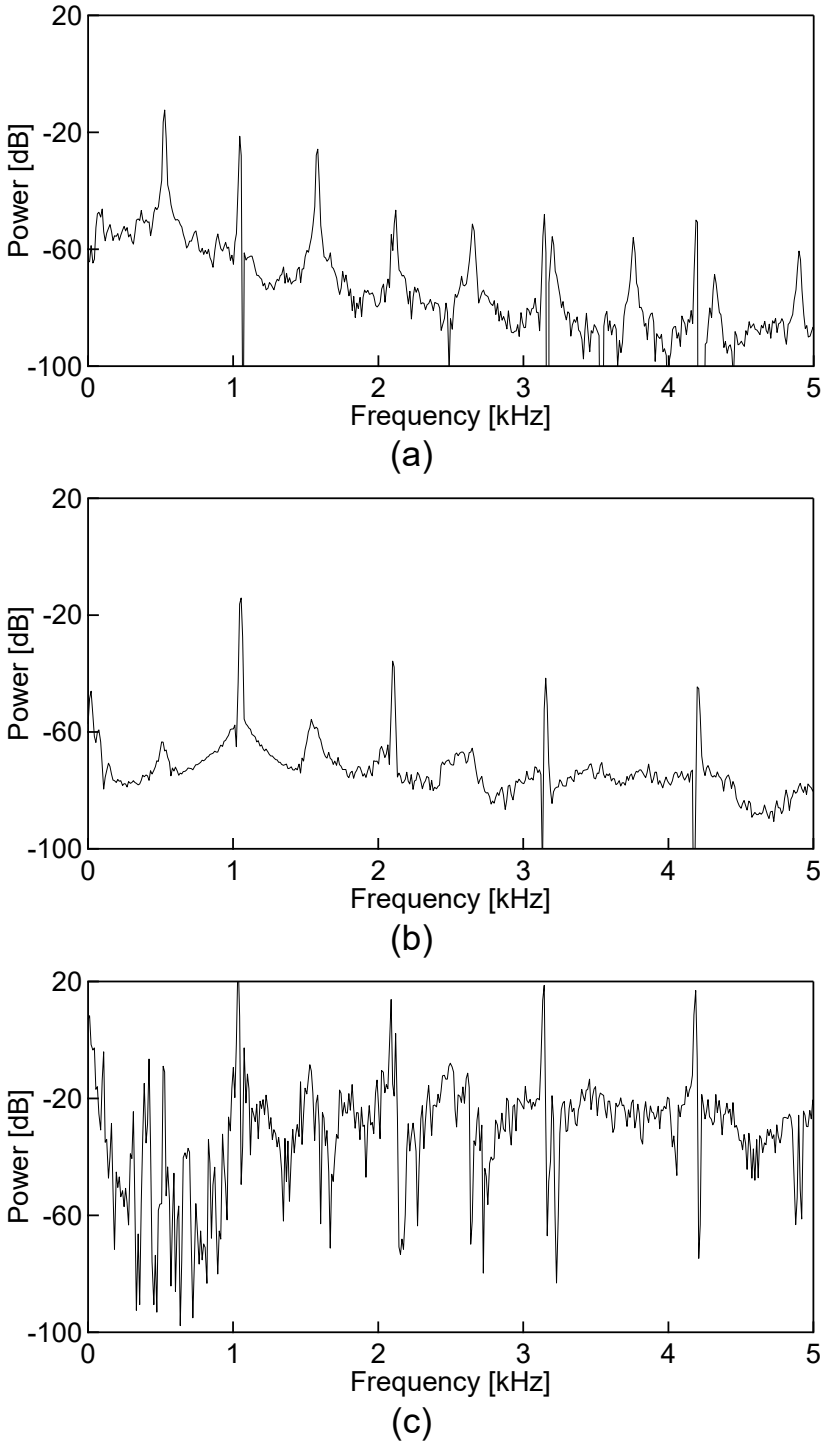

Figure 4.4: Spectral bases obtained from (4.16), where (a) is discriminative basis $\boldsymbol{F}'$ of piano tone and (b) and (c) are the other bases in $\boldsymbol{T}$.

Table 4.1: Mixture $\boldsymbol{\Lambda}$ with target and non-target sources

ID	Song name of mixture $\boldsymbol{\Lambda}$	Target source $\boldsymbol{S}^{(\text{test})}$ and $\boldsymbol{S}^{(\text{train})}$	Non-target source $\boldsymbol{N}^{(\text{test})}$
1	Roads	Acoustic guitar	Drums
2	Roads	Drums	Acoustic guitar
3	Que pena tanto faz	Classic guitar	Female vocals
4	Que pena tanto faz	Female vocals	Classic guitar
5	Ultimate NZ tour	Electric guitar	Synthesizer
6	Ultimate NZ tour	Synthesizer	Electric guitar

Table 4.2: Sample sounds of non-target source $\boldsymbol{N}^{(\text{train})}$ for preparing $\boldsymbol{S}_{\text{mix}}^{(\text{train})}$

ID	Song name	Sample sound $\boldsymbol{N}^{(\text{train})}$
1	The ones we love	Drums
2	The ones we love	Acoustic guitar
3	Remember the name	Male vocals
4	Ultimate NZ tour	Electric guitar
5	Remember the name	Synthetic violins
6	Roads	Acoustic guitar

$\boldsymbol{S}^{(\text{test})}$ and female vocals $\boldsymbol{N}^{(\text{test})}$. The sample sound $\boldsymbol{S}^{(\text{train})}$ was the same classic guitar sound obtained from a different segment of $\boldsymbol{S}^{(\text{test})}$ in the same song, and the sample sound $\boldsymbol{N}^{(\text{train})}$ was obtained from the different song "Remember the name." In addition, since $\boldsymbol{N}^{(\text{train})}$ is a male vocal sound, FSNMF does not work well in this situation. However, this experimental setting is convenient for evaluating the proposed method because we assumed that a similar type (but not exactly the same) of source to $\boldsymbol{N}$ was available as the sample sound $\boldsymbol{N}^{(\text{train})}$. In the proposed algorithm, preparing $\boldsymbol{N}^{(\text{train})}$ or $\boldsymbol{S}_{\text{mix}}^{(\text{train})}$ is a very important issue, and this experiment simulates the case that we know only the *type* of non-target source in the observed mixture $\boldsymbol{\Lambda}$.

The spectrograms were computed by an STFT with a 92-ms-long Hamming window and half-size shifting. In this experiment, all NMF decomposition is performed with amplitude spectrograms obtained via STFT, and as the

divergence criterion in NMF, we set β to 1, namely, NMF based on KL divergence (KLNMF). The numbers of bases in $\boldsymbol{F}$, $\boldsymbol{F}'$, $\boldsymbol{T}$, and $\boldsymbol{H}$ were set to 35, and the numbers of iterations in the training ((4.6) or (4.15)) and separation ((4.7) or (4.17)) stages were set to 1000. As the number of iterations in (4.16), we investigated numbers from 0 to 50, where 0 corresponds to simple SSNMF because the discriminative bases $\boldsymbol{F}'$ are equal to the reconstructive bases $\boldsymbol{F}$. As the separation performance, we used the improvement of SDR.

Results

Figure 4.5 shows the average SDR improvements of the target source. The zero point of the horizontal axis corresponds to simple SSNMF. As shown in Fig. 4.5, the proposed algorithm converges after about 20 iterations and the separation performance was improved at the point of convergence in all cases except ID6. Moreover, most of the improvement was obtained in the first four iterations. This indicates that better discriminative bases $\boldsymbol{F}'$ exist than $\boldsymbol{F}$ but they are not obtained at the point of convergence of the proposed method. We consider that this may be caused by the fact that we do not solve the bilevel optimization described in the beginning of Sect. 4.4.2. In a future work, we will investigate how to obtain better discriminative bases as a result of optimization.

4.6 Summary

In this chapter, I proposed a new basis training method for SSNMF. The proposed algorithm estimates both types of supervised basis matrix, namely, discriminative and reconstructive bases, for the target source and can be considered as an approximation to solving bilevel optimization to find discriminative bases. The efficacy of the proposed method was confirmed by performing a music separation.

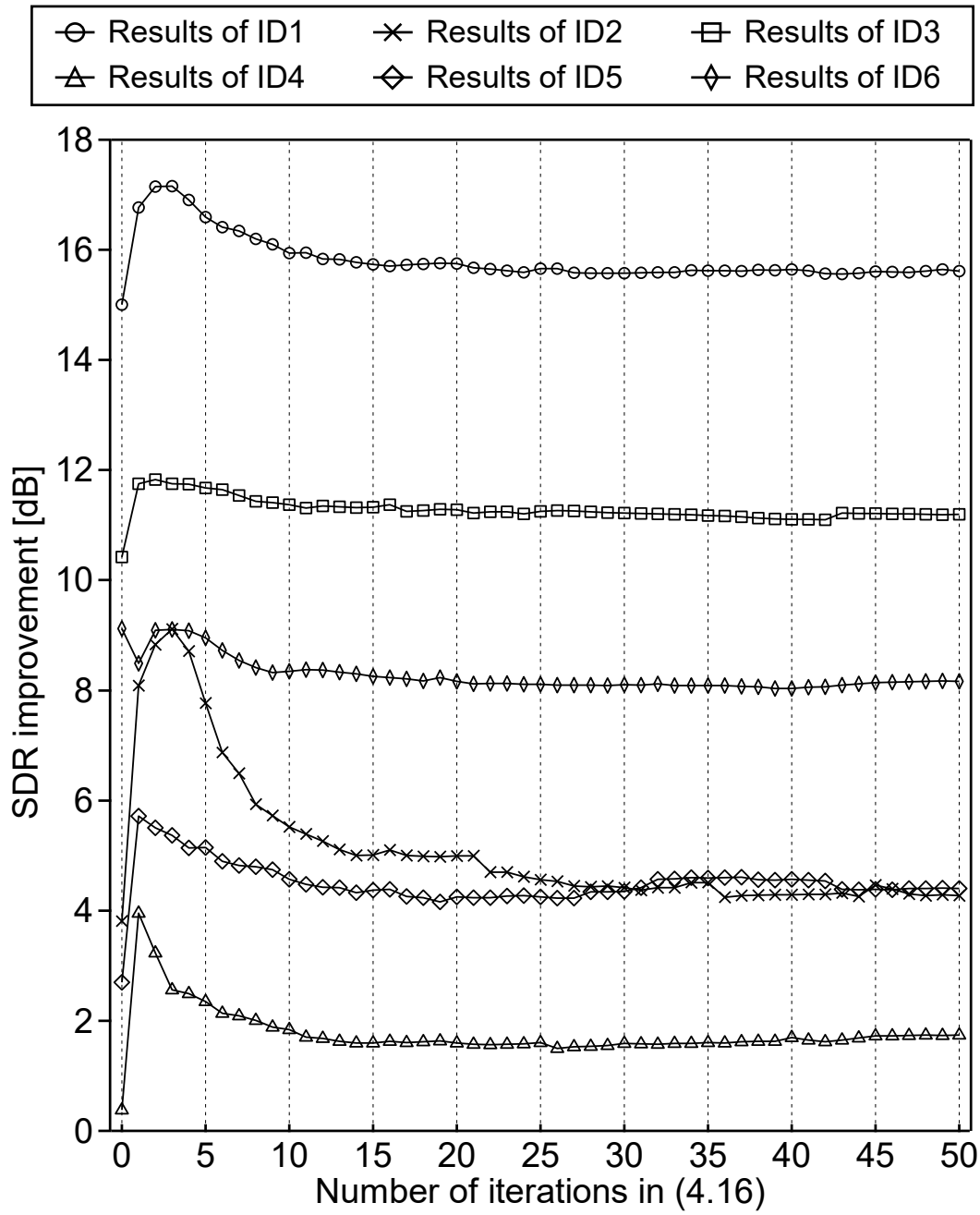

Figure 4.5: Average SDR improvement for each mixture and each number of iterations in (4.16).

5 Effective Initialization for Nonnegative Matrix Factorization Based on Statistical Independence

5.1 Introduction

In this chapter, I deal with an initialization problem for NMF. As already discussed so far, NMF is a powerful unsupervised learning method that extracts meaningful nonnegative features from an observed nonnegative matrix, which is not only the audio spectrogram but also many data. Various applications using NMF have been proposed as described in Sect. 2.5. However, the result of such applications always depends on the initial values of the NMF variables because of the existence of local minima. To solve this problem, in this chapter, I propose new initialization methods based on statistical independence between NMF components. First, existing conventional initializations for NMF are

introduced with the viewpoint of presence of random values. Next, after the motivation is clarified, the efficient initialization method based on ICA. To take the nonnegativity in NMF into account, I propose two types of algorithms: (i) applying nonnegative ICA (NICA) [220, 221, 222] to the observed data matrix; (ii) applying simple ICA with zero-mean Laplace prior to the differential of observed data matrix using nonnegative projection in each update. The convergence speed and the converged value of NMF cost function is compared with conventional and proposed methods. Also, the availability of the proposed initialization for source separation is experimentally investigated via FSNMF, BSS based on ILRMA proposed in Chap. 3, and discriminative SSNMF proposed in Chap. 4. Finally, the contents in this chapter are summarized.

5.2 Conventional NMF Initializations

Similar to Sect. 2.5, let $\boldsymbol{\Delta} \in \mathbb{R}_{\geq 0}^{\Phi \times \Psi}$ be the observed nonnegative matrix, $\boldsymbol{F} \in \mathbb{R}_{\geq 0}^{\Phi \times K}$ and $\boldsymbol{G} \in \mathbb{R}_{\geq 0}^{K \times \Psi}$ are the nonnegative basis and activation matrices, Φ is the dimension in observation, Ψ is the number of observed data samples, and K is the number of bases. Also, $\boldsymbol{\delta}_\psi$, $\boldsymbol{f}_k$, and $\boldsymbol{g}_k$ denote the vectors in the matrices $\boldsymbol{\Delta}$, $\boldsymbol{F}$ and $\boldsymbol{G}$, respectively, namely, $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_\psi)$, $\boldsymbol{F} = (\boldsymbol{f}_1, \cdots, \boldsymbol{f}_K)$, and $\boldsymbol{G} = (\boldsymbol{g}_1, \cdots, \boldsymbol{g}_K)^{\mathrm{T}}$. The optimization method in NMF, e.g., MU rules (2.20) and (2.21), requires initial variables $\boldsymbol{F}^{(\mathrm{ini})}$ and $\boldsymbol{G}^{(\mathrm{ini})}$. Then, a neighborhood local minimum can be obtained as a solution of the cost function (2.18). Figure 5.1 shows an example of full-supervised music source separation using FSNMF, where the NMF variables are initialized using pseudorandom values obtained from various pseudorandom seeds (hereafter, referred to as *Rand1* to *Rand10*). We can confirm that the separation performance strongly depends on the NMF initialization.

"Good" initial values for NMF are defined as follows [223]: (i) those that lead to rapid minimization of the divergence and fast convergence; (ii) those that lead to low overall divergence at the point of convergence. In this chapter, I concentrate on the both of these objectives.

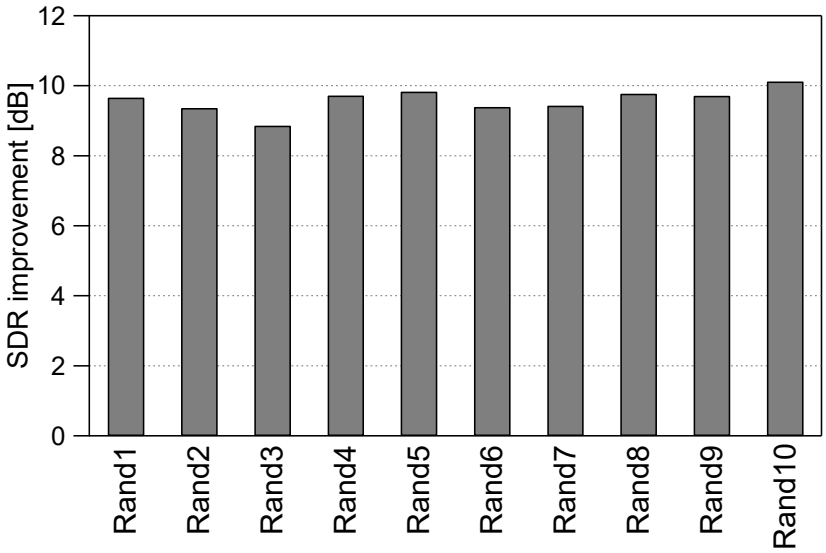

Figure 5.1: Example of SDR improvements of music source separation, where FSNMF initialized by random values with various pseudorandom seeds is performed.

5.2.1 Initialization with Random Values

The simplest initialization method is randomization, namely, we prepare $(\boldsymbol{F}^{(\mathrm{ini})}, \boldsymbol{G}^{(\mathrm{ini})})$ by producing pseudorandom values. This method leads various results depending on the random seed. Therefore, the best result should be adopted via many trials with different seeds. In [224, 225], a genetic algorithm was utilized to find good initial values. In addition, several initialization methods based on clustering of the data matrix $\boldsymbol{\Delta}$ have been proposed [226, 227, 228]. The methods in [226, 227, 228] utilized the result of clustering, i.e., centroid vectors, to define $\boldsymbol{F}^{(\mathrm{ini})}$, but they required initial centroid vectors, which were usually determined as random values.

5.2.2 Initialization without Random Values

The authors in [229] proposed the use of subtractive clustering, which does not require initial centroids, namely, a unique result can be obtained for NMF decomposition. However, the subtractive clustering includes two hyperparame-

ters that must be heuristically tuned by the users. As initialization methods that do not require both random values and hyperparameter tuning, PCA and singular value decomposition (SVD) have been utilized [230, 223]. In the former method [230], orthogonal bases and their weights obtained by applying PCA to $\boldsymbol{\Delta}$ are assigned to $(\boldsymbol{F}^{(\mathrm{ini})}, \boldsymbol{G}^{(\mathrm{ini})})$, where the negative entries are replaced by their absolute values. In the SVD-based method [223], nonnegative double SVD (NNDSVD) is applied to $\boldsymbol{\Delta}$. The initial values $(\boldsymbol{F}^{(\mathrm{ini})}, \boldsymbol{G}^{(\mathrm{ini})})$ are set to the nonnegative left and right singular vectors obtained via NNDSVD. These two methods can provide a unique decomposition.

5.3 Efficient NMF Initialization Based on ICA

5.3.1 Motivation and Strategy

The initialization methods using PCA or SVD are based on the orthogonality between the bases representing the data matrix $\boldsymbol{\Delta}$. However, it has been shown that the optimal NMF bases are along the edges of a *convex polyhedral cone*, which is defined by the observed points in $\boldsymbol{\Delta}$, in an Φ-dimensional space [231, 232]. Figure 5.2 shows the various NMF bases when $\Phi = K = 2$. The optimal bases are satisfactory for representing all the data points, whereas the close bases cannot represent them because of the nonnegative constraint of the activations. The orthogonal bases are excessive for representing the data points and have a risk to represent even a meaningless area. Therefore, PCA and SVD may not be the best methods for the initialization in NMF.

In this chapter, I propose the utilization of bases and independent sources estimated by ICA for $\boldsymbol{F}^{(\mathrm{ini})}$ and $\boldsymbol{G}^{(\mathrm{ini})}$, respectively. ICA can estimate non-orthogonal bases $\boldsymbol{a}_k$ that provide a mixing matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_K)$ for the independent sources as $\boldsymbol{AS}$, where $\boldsymbol{a}_k$ is the $K \times 1$ kth ICA basis and $\boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K)^{\mathrm{T}}$, $\boldsymbol{s}_k$ is the $\Psi \times 1$ kth source signal. Thus, ICA can estimate bases so that the sources are independent of each other, and such bases tend to be dissimilar but they are not orthogonal. In addition, the estimated sources $\boldsymbol{s}_k$ tend to be sparse if we assume a super-Gaussian distribution as a source distribution in ICA. When the coefficients are sparse, their bases will be along the edges of

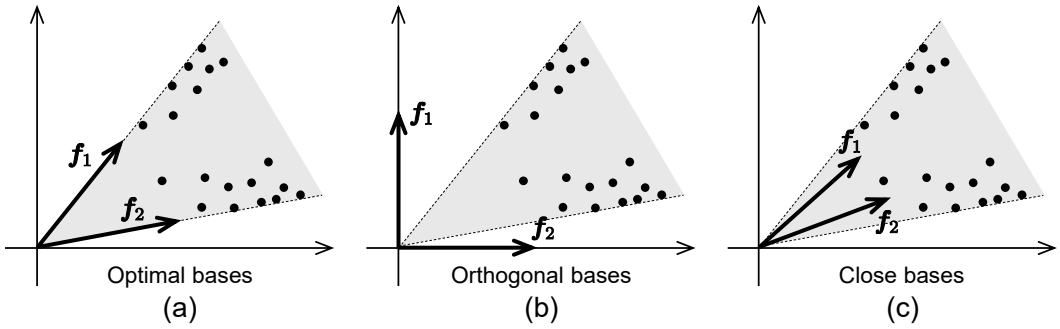

Figure 5.2: Geometry of (a) optimal, (b), orthogonal, and (c) close bases, where black dots indicate observed data points in positive orthant, gray area indicates cone defined by data points, broken lines indicate edges of cone, $\boldsymbol{f}_k$ denotes kth NMF basis, $\Phi = K = 2$, and $\Psi = 10$.

the cone as shown in Fig. 5.2 (a). Therefore, by using the independent sources and their bases for the initial values in NMF, the optimization may avoid local minima. In fact, an initialization method for probabilistic latent component analysis (PLCA) [233] based on ICA has been proposed [234], where PLCA is inherently identical to KL-divergence-based NMF. However, the method in [234] did not use the ICA bases $\boldsymbol{a}_k$ but the demixing filters $\boldsymbol{w}_k$, which are the inverse of the ICA bases, $\boldsymbol{W} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K)^{\mathrm{T}} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_K)^{-1}$, and provide the estimated sources $\boldsymbol{y}_k$. Also, the authors in [234] did not discuss how to treat the nonnegative entries in $\boldsymbol{w}_k$ and $\boldsymbol{y}_k$. Moreover, there was no comparison with other initializations such as the PCA-based method and NNDSVD. To take the nonnegativity into account, I propose the employment of ICA for the initialization in NMF. Also, the proposed method performs PCA before ICA as a preprocess for simulating the dimensionality reduction in NMF. To take the nonnegativity in NMF into account, I here propose two types of initialization algorithms: (i) applying NICA [220, 221, 222] to the observed data matrix $\boldsymbol{\Delta}$; (ii) applying simple ICA with zero-mean Laplace prior to the differential of observed data matrix, $\boldsymbol{\Delta\Theta}$, and applying nonnegative projection in each update of ICA, where $\boldsymbol{\Theta}$ is a differential matrix that takes difference between the data

point and its neighbor in each dimension ϕ, namely, $\delta_{\phi\psi} - \delta_{\phi(\psi+1)}$, as

$$\boldsymbol{\Theta} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}. \tag{5.1}$$

5.3.2 Combination of PCA and ICA

The dimensionality reduction for arbitrary nonnegative matrix $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{\Phi \times \Psi}$ using PCA can be represented as

$$\begin{cases} \boldsymbol{\mathcal{P}}_1 \boldsymbol{X} = \boldsymbol{AS} \\ \boldsymbol{\mathcal{P}}_2 \boldsymbol{X} \approx \boldsymbol{0} \end{cases}, \tag{5.2}$$

where

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} \boldsymbol{\mathcal{P}}_1 \\ \boldsymbol{\mathcal{P}}_2 \end{pmatrix} \tag{5.3}$$

is the $\Phi \times \Phi$ transform matrix of PCA and the sizes of $\boldsymbol{\mathcal{P}}_1$ and $\boldsymbol{\mathcal{P}}_2$ are $K \times \Phi$ and $(\Phi - K) \times \Phi$, respectively. The row vectors in $\boldsymbol{\mathcal{P}}$ correspond to the eigenvectors of a variance-covariance matrix $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, and the eigenvectors are arranged in descending order from the first row to the last row on the basis of their eigenvalues. Therefore, $\boldsymbol{\mathcal{P}}_1$ includes the top K eigenvectors of $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ and $\boldsymbol{\mathcal{P}}_2$ includes the remaining eigenvectors. In addition, $\boldsymbol{0}$ is the $(\Phi - K) \times \Psi$ zero matrix. Thus, we assume that the independent sources in $\boldsymbol{S}$ are mixed via the mixing matrix $\boldsymbol{A}$ and are observed as the mixture $\boldsymbol{\mathcal{P}}_1 \boldsymbol{X}$. From the NMF side, the nonnegative activations are assumed to be independent of each other, as shown in Fig. 5.3. Note that since NICA will be applied to $\boldsymbol{\mathcal{P}}_1 \boldsymbol{X}$ (after the dimensionality reduction via PCA), the estimated ICA bases $\boldsymbol{a}_k$ are not orthogonal.

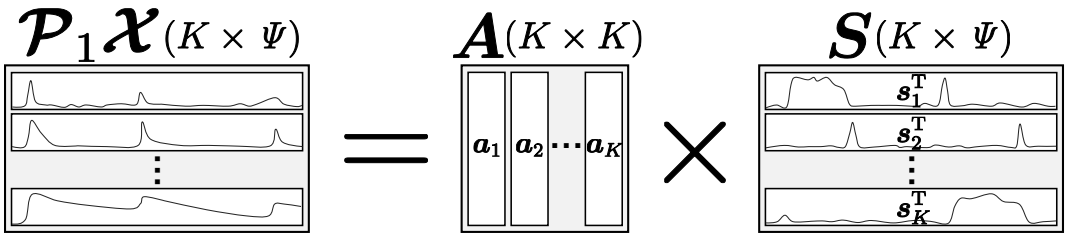

Figure 5.3: Assumption of proposed method, where nonnegative activations are assumed to be independent of each other.

5.3.3 Proposed Initialization using NICA

NICA can estimate the nonnegative independent components from an observed multichannel mixture. The essence of NICA is to find a rotation matrix $\boldsymbol{W}$ for the noncentered and whitened data so that all the estimated (separated) sources become nonnegative [220]:

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\Omega}, \tag{5.4}$$

$$\boldsymbol{\Omega} = \mathcal{W}\mathcal{P}_1\mathcal{X} = \mathcal{W}\boldsymbol{A}\boldsymbol{S}, \tag{5.5}$$

where $\mathcal{W}$ is a whitening matrix, which transforms $\mathcal{P}_1\mathcal{X}$ so that $\mathcal{P}_1\mathcal{X}(\mathcal{P}_1\mathcal{X})^{\mathrm{T}}$ becomes the identity matrix, and $\mathcal{X} = \boldsymbol{\Delta}$ in this method. Note that this whitening process does not center the data, namely, it does not remove the mean of $\mathcal{P}_1\mathcal{X}$. In addition, $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)^{\mathrm{T}}$ is a matrix that comprises of estimated sources $\boldsymbol{y}_k$, and $\boldsymbol{W}$ is a demixing matrix that rotates the whitened data $\boldsymbol{\Omega}$. If the sources $\boldsymbol{s}_k$ are truly nonnegative, we can obtain a global solution such that all the estimates $\boldsymbol{y}_k$ become nonnegative. However, in the proposed method, such a global solution probably does not exist because of the dimensionality reduction via PCA. The optimization in NICA is defined as the minimization of the total power of the residual negative estimates [220]:

$$\min_{\boldsymbol{W}} \sum_{k,\psi} \min(0, y_{k\psi})^2, \tag{5.6}$$

where $y_{k\psi}$ is the entry of $\boldsymbol{Y}$. The steepest gradient descent has been proposed for (5.6) as follows [221]:

$$\tilde{\boldsymbol{w}}_k = \boldsymbol{w}_k - 2\eta \sum_{\psi} \min(0, \omega_{k\psi})\omega_{k\psi}, \tag{5.7}$$

$$\boldsymbol{W} = \left(\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^{\mathrm{T}}\right)^{-1/2}\tilde{\boldsymbol{W}}, \tag{5.8}$$

where $\boldsymbol{w}_k$ and $\tilde{\boldsymbol{w}}_k$ are the column vectors of $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}}$, respectively, η is the stepsize parameter, $\omega_{k\psi}$ is the entry of $\boldsymbol{\Omega}$, and $\tilde{\boldsymbol{W}}$ is the matrix with $\boldsymbol{w}_k$ as its columns. Whereas optimization without a hyperparameter such as η has also been proposed as "fast NICA" [222], I use (5.7) and (5.8) in this dissertation.

The estimated sources $\boldsymbol{Y}$ can be used for the initial values of the activation matrix $\boldsymbol{G}$. Also, the basis matrix $\boldsymbol{F}$ can be calculated from the estimated demixing matrix $\boldsymbol{W}$. If we approximately assume $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{G}$, $\boldsymbol{S} = \boldsymbol{Y}$, and $\boldsymbol{A} = (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}$, the following equation can be obtained from (5.2):

$$\boldsymbol{\mathcal{P}}\boldsymbol{F}\boldsymbol{G} \approx \begin{bmatrix} (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1} \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{G}. \tag{5.9}$$

Then, the basis matrix $\boldsymbol{F}$ can be obtained as

$$\boldsymbol{F} \approx \boldsymbol{\mathcal{P}}^{-1} \begin{bmatrix} (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1} \\ \boldsymbol{0} \end{bmatrix}. \tag{5.10}$$

5.3.4 Proposed Initialization using ICA and Differential of Data Matrix

When the source $\boldsymbol{S}$ and the observed data $\boldsymbol{X}$ have both positive and negative values, the regular ICA algorithm can be used for the estimation of $\boldsymbol{W}$. Thus, I also propose a utilization of ICA with differentiated data matrix $\boldsymbol{\Delta\Theta}$. Whereas we assumed $\boldsymbol{X} = \boldsymbol{\Delta}$ and estimate $\boldsymbol{S} = \boldsymbol{G}$ in Sect. 5.3.3, in this method, $\boldsymbol{X} = \boldsymbol{\Delta\Theta}$ is assumed to estimate $\boldsymbol{S} = \boldsymbol{G\Theta}$. I here apply ICA with Laplace distribution as the super-Gaussian source distribution because the ICA cost function with Laplace distribution becomes convex with respect to $\boldsymbol{W}$, and the unique solution can be

obtained via optimization. In addition, the fast and stable optimization based on auxiliary function technique has been proposed [176]. After the estimation of $\boldsymbol{W}$, the initial activation and basis matrices can be calculated as $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{\Delta}$ (not $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{X}$) and

$$\boldsymbol{F} \approx \boldsymbol{\mathcal{P}}^{-1} \begin{bmatrix} \boldsymbol{W}^{-1} \\ \boldsymbol{0} \end{bmatrix}. \tag{5.11}$$

In this method, there is no guarantee that the basis matrix $\boldsymbol{F}$ is a nonnegative matrix. To ensure the nonnegativity, in each iteration of ICA optimization, I propose to calculate $\boldsymbol{F}$ using (5.11), update as $\boldsymbol{F} \leftarrow \max(\boldsymbol{F}, 0)$ (projected to the nonnegative values), and recalculate $\boldsymbol{W}$ from the updated $\boldsymbol{F}$.

5.3.5 Nonnegativization

Since we apply PCA for dimensionality reduction, there is no guarantee that all the entries of the obtained activation matrix $\boldsymbol{G}$ become nonnegative. In particular, the proposed method using ICA does not ensure the nonnegativity of the basis matrix $\boldsymbol{F}$. For these reasons, I apply *nonnegativization* to the obtained $\boldsymbol{F}$ and $\boldsymbol{G}$ by the proposed methods. I here perform any of the following three nonnegativizations:

Nonnegativization 1: $\boldsymbol{F}^{(\mathrm{ini})} = |\boldsymbol{F}|$, $\boldsymbol{G}^{(\mathrm{ini})} = |\boldsymbol{G}|$,

Nonnegativization 2: $\boldsymbol{F}^{(\mathrm{ini})} = |\boldsymbol{F}|$, $\boldsymbol{G}^{(\mathrm{ini})} = \alpha_G \boldsymbol{F}^{(\mathrm{ini})\mathrm{T}} \boldsymbol{\Delta}$,

Nonnegativization 3: $\boldsymbol{G}^{(\mathrm{ini})} = |\boldsymbol{G}|$, $\boldsymbol{F}^{(\mathrm{ini})} = \alpha_F \boldsymbol{\Delta} \boldsymbol{G}^{(\mathrm{ini})\mathrm{T}}$,

where α_F and α_G are coefficients for fitting the scale of $\boldsymbol{F}^{(\mathrm{ini})}\boldsymbol{G}^{(\mathrm{ini})}$ to $\boldsymbol{\Delta}$. The values of these coefficients depend on the following NMF after the proposed initialization and can easily be calculated from

$$\alpha_F = \arg\min_{\alpha} \mathcal{D}_\beta\left(\boldsymbol{\Delta} \| \alpha \boldsymbol{\Delta} \boldsymbol{G}^{(\mathrm{ini})\mathrm{T}} \boldsymbol{G}^{(\mathrm{ini})}\right), \tag{5.12}$$

$$\alpha_G = \arg\min_{\alpha} \mathcal{D}_\beta\left(\boldsymbol{\Delta} \| \alpha \boldsymbol{F}^{(\mathrm{ini})} \boldsymbol{F}^{(\mathrm{ini})\mathrm{T}} \boldsymbol{\Delta}\right). \tag{5.13}$$

Here, I describe the solutions of (5.12) and (5.13) for the cases of NMF based on EU distance (EUNMF), KLNMF, and ISNMF as follows:

$$\textbf{For EUNMF}: \ \alpha_F = \frac{\sum_{\phi,\psi} \delta_{\phi\psi} \sum_{\psi',k} \delta_{\phi\psi'} g^{(\mathrm{ini})}_{k\psi'} g^{(\mathrm{ini})}_{k\psi'}}{\sum_{\phi,\psi} \left(\sum_{\psi',k} \delta_{\phi\psi'} g^{(\mathrm{ini})}_{k\psi'} g^{(\mathrm{ini})}_{k\psi'} \right)^2}, \quad \alpha_G = \frac{\sum_{\phi,\psi} \delta_{\phi\psi} \sum_{\phi',k} f^{(\mathrm{ini})}_{\phi'k} f^{(\mathrm{ini})}_{\phi'k} \delta_{\phi'\psi}}{\sum_{\phi,\psi} \left(\sum_{\phi',k} f^{(\mathrm{ini})}_{\phi'k} f^{(\mathrm{ini})}_{\phi'k} \delta_{\phi'\psi} \right)^2},$$

$$\textbf{For KLNMF}: \ \alpha_F = \frac{\sum_{\phi,\psi} \delta_{\phi\psi}}{\sum_{\phi,\psi} \sum_{\psi',k} \delta_{\phi\psi'} g^{(\mathrm{ini})}_{k\psi'} g^{(\mathrm{ini})}_{k\psi'}}, \quad \alpha_G = \frac{\sum_{\phi,\psi} \delta_{\phi\psi}}{\sum_{\phi,\psi} \sum_{\phi',k} f^{(\mathrm{ini})}_{\phi'k} f^{(\mathrm{ini})}_{\phi'k} \delta_{\phi'\psi}},$$

$$\textbf{For ISNMF}: \ \alpha_F = \frac{1}{\Phi\Psi} \sum_{\phi,\psi} \frac{\delta_{\phi\psi}}{\sum_{\psi',k} \delta_{\phi\psi'} g^{(\mathrm{ini})}_{k\psi'} g^{(\mathrm{ini})}_{k\psi'}}, \quad \alpha_G = \frac{1}{\Phi\Psi} \sum_{\phi,\psi} \frac{\delta_{\phi\psi}}{\sum_{\phi',k} f^{(\mathrm{ini})}_{\phi'k} f^{(\mathrm{ini})}_{\phi'k} \delta_{\phi'\psi}},$$

where $f^{(\mathrm{ini})}_{\phi k}$ and $g^{(\mathrm{ini})}_{k\psi}$ are the entries of $\boldsymbol{F}^{(\mathrm{ini})}$ and $\boldsymbol{G}^{(\mathrm{ini})}$, respectively.

5.4 Experimental Comparisons

5.4.1 Performance as Initial Value for NMF

Conditions

To evaluate the performance of the proposed method, I compare the convergence speed of the cost function in NMF among five initialization methods, namely, uniform random values in the range between ε and 1, PCA-based initialization [230], NNDSVD [223], the proposed method using NICA, and the proposed method using ICA with differentiated data matrix. As the observed data $\boldsymbol{\Delta}$, I used a power spectrogram of the music signal "Actions - One Minute Smile" obtained from the MSD100 dataset, which was published by SiSEC2015 [235]. MSD100 consists of four parts for each song, namely, *vocals*, *bass*, *drums*, and *other*. I here chose only the vocals and other sources. In addition, the section of these source signals from 40 s to 100 s was extracted to obtain 60-s-long vocals and other signals, and the observed signal is a mixture of them. The power spectrogram of the observed signal was computed by an STFT with a 92-ms-long Hamming window and half-size shifting. The size of matrix $\boldsymbol{\Delta}$ was $\Phi = 2049$ and $\Psi = 1290$, and the sampling frequency was 44.1 kHz. After the initialization of $\boldsymbol{F}$ and $\boldsymbol{G}$ using the conventional or proposed method, EUNMF, KLNMF, or ISNMF was preformed, where the MU update rules (2.20) and (2.21) were used to minimize (2.18), and the number of bases was set to $K = 60$. For the random

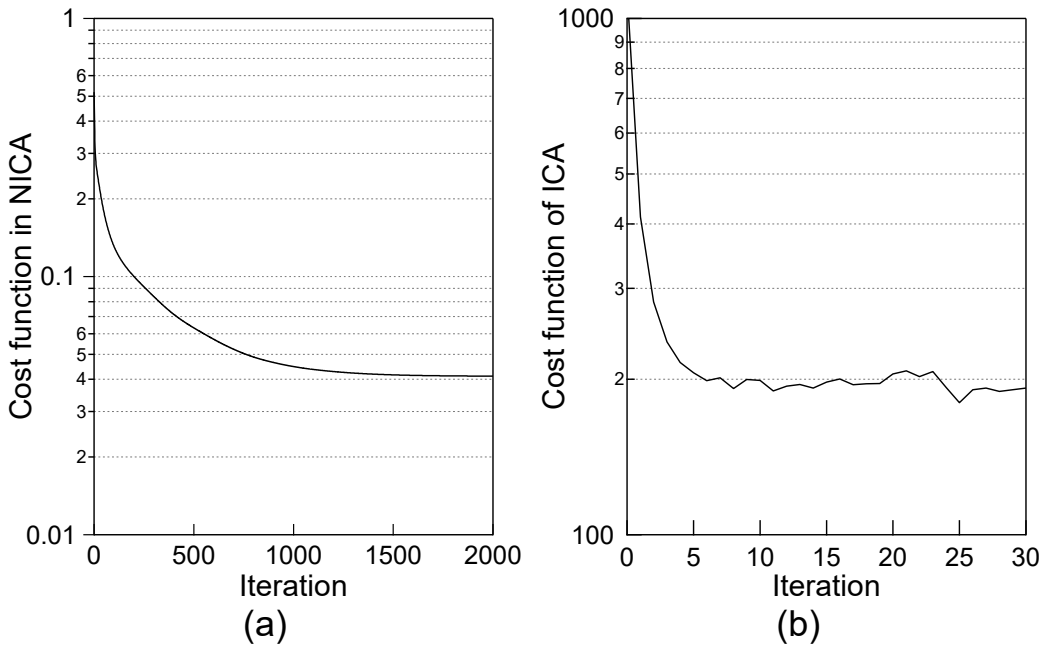

Figure 5.4: Convergence of cost function in (a) NICA and (b) ICA.

initialization, I performed 10 trials with various pseudorandom seeds (Rand1 to Rand10). Also, for the proposed methods, I compared six methods, namely, two algorithms (NICA and ICA) and three nonnegativizations (hereafter, referred to as *NICA1* to *NICA3* and *ICA1* to *ICA3*, where the number corresponds to the type of nonnegativization described in Sect. 5.3.5). The number of iterations for optimization was set to 2000 in NICA1–NICA3 and 30 in ICA1–ICA3.

Results

Figure 5.4 shows the convergence of the cost function (5.6) in NICA and ICA. Since I used the steepest gradient descent (5.7) and (5.8), more than 1000 iterations were required for NICA. This may be reduced by using fast NICA [222]. The convergence of ICA based on auxiliary function technique (Fig. 5.6 (b)) is much faster than that of NICA. However, since this algorithm includes the nonnegative projection in each iteration, the optimization does not ensure monotonic decrease. In addition, the computational complexity of ICA is large

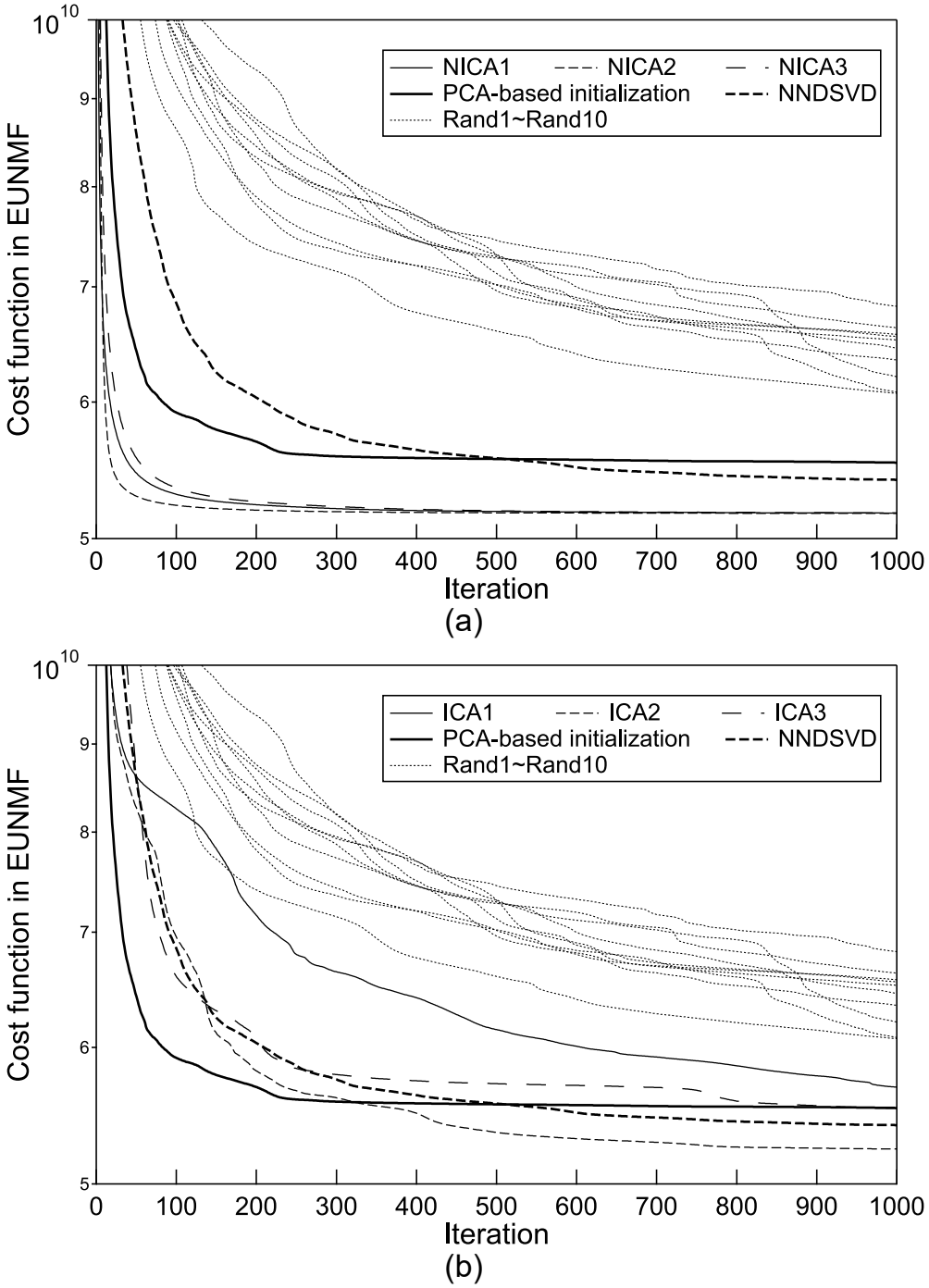

Figure 5.5: Convergences of cost function in EUNMF, where NICA1–NICA3 are depicted in (a), ICA1–ICA3 are depicted in (b), and conventional methods are depicted in both.

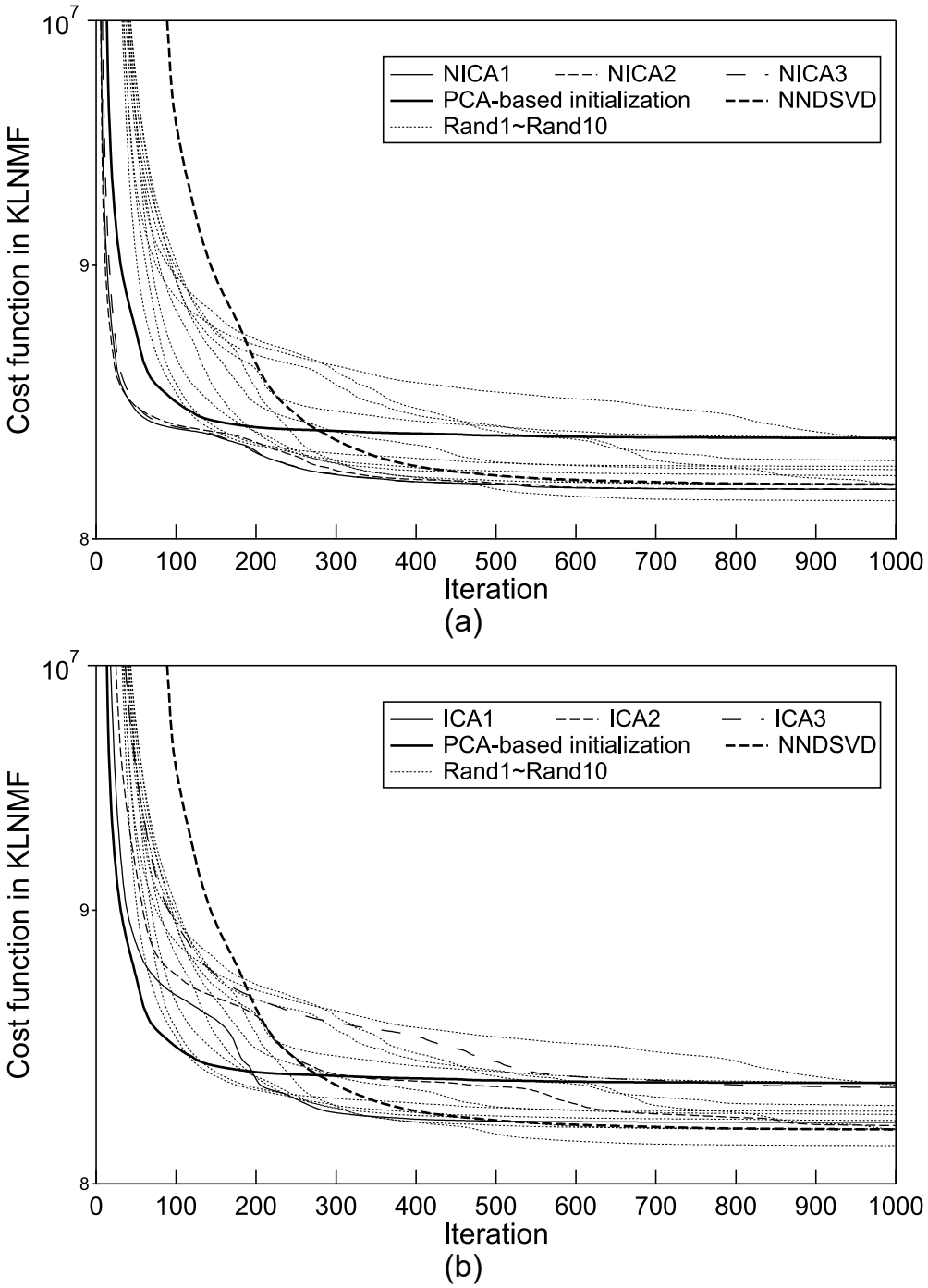

Figure 5.6: Convergences of cost function in KLNMF, where NICA1–NICA3 are depicted in (a), ICA1–ICA3 are depicted in (b), and conventional methods are depicted in both.

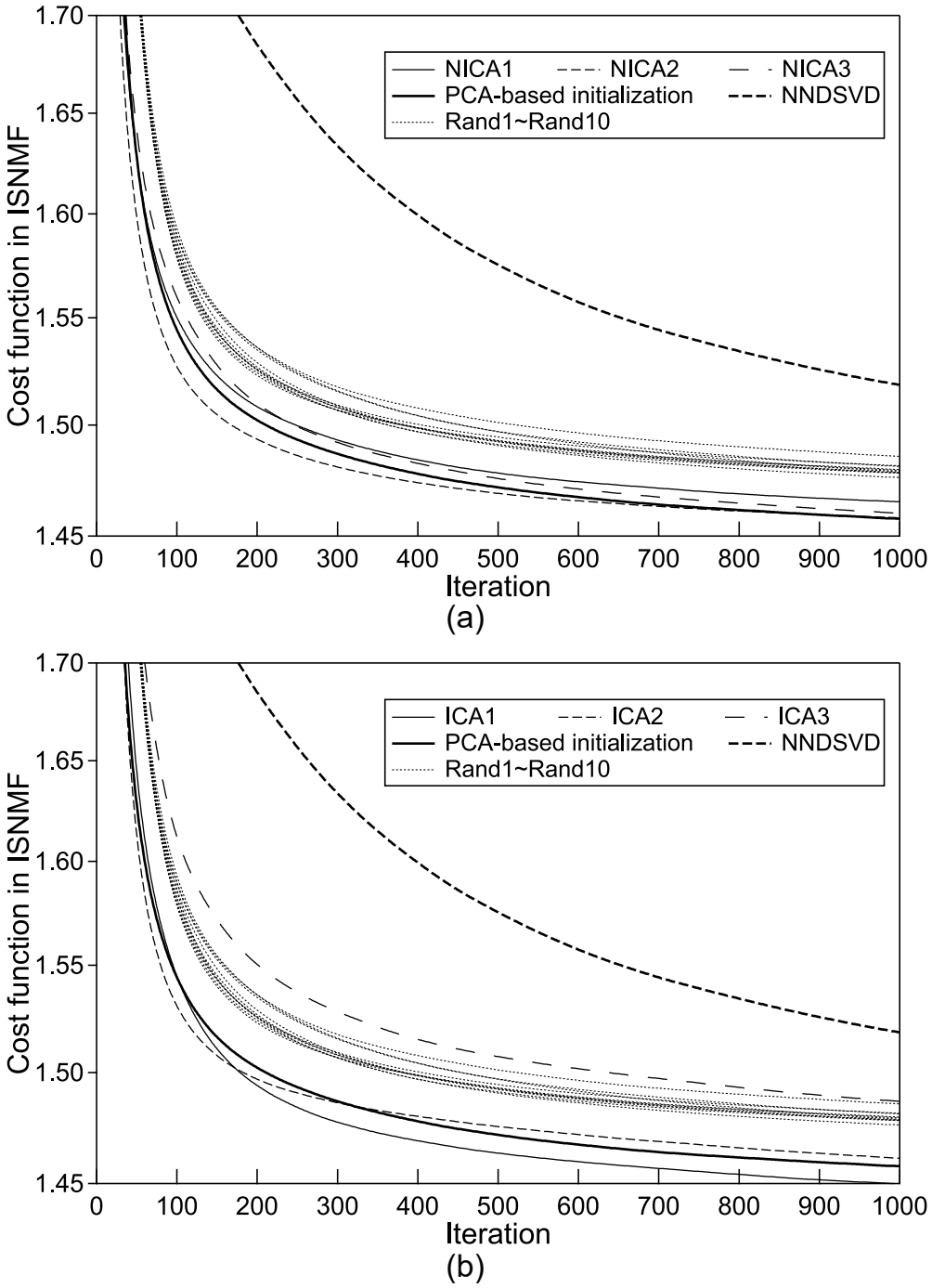

Figure 5.7: Convergences of cost function in ISNMF, where NICA1–NICA3 are depicted in (a), ICA1–ICA3 are depicted in (b), and conventional methods are depicted in both.

Table 5.1: Examples of computational time in each process (s)

Process	Algorithm	Processing time
Initialization	NICA1–NICA3 (2000 iterations)	4.36
	ICA1–ICA3 (30 iterations)	10.89
	PCA-based initialization	0.98
	NNDSVD	2.40
NMF	EUNMF (1000 iterations)	12.78
	KLNMF (1000 iterations)	48.07
	ISNMF (1000 iterations)	214.26

because it includes inverse calculation of $K \times K$ matrix $\boldsymbol{W}$. Figures 5.5–5.7 show the convergences of the cost function (2.18) in NMF, where NICA1–NICA3 are depicted in (a), ICA1–ICA3 are depicted in (b), and the conventional methods are depicted in both. From these results, we can confirm that the random initialization has a slow convergence speed and a poor local minimum at the point of convergence. PCA-based initialization provides faster and deeper minimization than random initialization, especially for EUNMF and ISNMF. NNDSVD has the best convergence except for ISNMF. The methods based on NICA outperform these conventional methods in all the cases of NMF. In particular, NICA2 provides the most stable convergence. The methods based on ICA also provide comparable performance with PCA-based initialization and NNDSVD. However, the best nonnegativization depends on the criteria in NMF cost function. Examples of the computational time for each initialization method and NMF are given in Table 5.1. Although the proposed methods have larger computational costs than the other methods, the increase is not critical compared with the case of NMF iterations.

5.4.2 Full-Supervised Audio Source Separation

Conditions

I compare the performance of audio source separation using FSNMF. In this experiment, I chose the top 15 songs in alphabetical order of a test dataset in

MSD100 and only the vocals and other parts as the source signals. Similarly to in Sect. 5.4.1, 60-s-long vocals and other signals were used. The power spectrograms of the vocals and other signals were calculated via STFT as $\boldsymbol{\Delta}_{\mathrm{V}}$ and $\boldsymbol{\Delta}_{\mathrm{O}}$, respectively. In addition, sixfold cross-validation was applied to them, namely, the training sounds of each source ($\boldsymbol{\Delta}_{\mathrm{V}}^{(\mathrm{train})}$ and $\boldsymbol{\Delta}_{\mathrm{O}}^{(\mathrm{train})}$) were obtained from five-sixths of the frames in $\boldsymbol{\Delta}_{\mathrm{V}}$ and $\boldsymbol{\Delta}_{\mathrm{O}}$, and the remaining one-sixth of the frames, $\boldsymbol{\Delta}_{\mathrm{V}}^{(\mathrm{test})}$ and $\boldsymbol{\Delta}_{\mathrm{O}}^{(\mathrm{test})}$, were used to obtain a mixture $\boldsymbol{\Delta}_{\mathrm{mix}}^{(\mathrm{test})} \approx \boldsymbol{\Delta}_{\mathrm{V}}^{(\mathrm{test})} + \boldsymbol{\Delta}_{\mathrm{O}}^{(\mathrm{test})}$ of vocals and other source signals, where the mixing was performed in the time domain. The supervised sourcewise basis matrices $\boldsymbol{F}_{\mathrm{V}}$ and $\boldsymbol{F}_{\mathrm{O}}$ were trained by the following NMF:

$$\boldsymbol{F}_{\mathrm{V}} = \underset{\boldsymbol{F},\, \boldsymbol{G}}{\arg\min}\ \mathcal{D}_{\beta}\left(\boldsymbol{\Delta}_{\mathrm{V}}^{(\mathrm{train})} \| \boldsymbol{F}\boldsymbol{G}\right), \tag{5.14}$$

$$\boldsymbol{F}_{\mathrm{O}} = \underset{\boldsymbol{F},\, \boldsymbol{G}}{\arg\min}\ \mathcal{D}_{\beta}\left(\boldsymbol{\Delta}_{\mathrm{O}}^{(\mathrm{train})} \| \boldsymbol{F}\boldsymbol{G}\right), \tag{5.15}$$

where $\boldsymbol{F}$ and $\boldsymbol{G}$ are initialized by $\boldsymbol{F}^{(\mathrm{ini})}$ and $\boldsymbol{G}^{(\mathrm{ini})}$, respectively, and the nonnegative constraint is assumed in this minimization. In the separation stage, the supervised NMF was performed with fixed bases $\boldsymbol{F}_{\mathrm{V}}$ and $\boldsymbol{F}_{\mathrm{O}}$ under the nonnegative constraint as follows:

$$\min_{\boldsymbol{G}_{\mathrm{V}},\, \boldsymbol{G}_{\mathrm{O}}}\ \mathcal{D}_{\beta}\left(\boldsymbol{\Delta}_{\mathrm{mix}}^{(\mathrm{test})} \| \boldsymbol{F}_{\mathrm{V}}\boldsymbol{G}_{\mathrm{V}} + \boldsymbol{F}_{\mathrm{O}}\boldsymbol{G}_{\mathrm{O}}\right), \tag{5.16}$$

where the initial values of $\boldsymbol{G}_{\mathrm{V}}$ and $\boldsymbol{G}_{\mathrm{O}}$ were set to $\boldsymbol{G}_{\mathrm{V}} = \alpha_G \boldsymbol{F}_{\mathrm{V}}^{\mathrm{T}} \boldsymbol{\Delta}_{\mathrm{mix}}^{(\mathrm{test})}$ and $\boldsymbol{G}_{\mathrm{O}} = \alpha_G \boldsymbol{F}_{\mathrm{O}}^{\mathrm{T}} \boldsymbol{\Delta}_{\mathrm{mix}}^{(\mathrm{test})}$, respectively. Therefore, the separation performance only depends on the initial values $\boldsymbol{F}^{(\mathrm{ini})}$ and $\boldsymbol{G}^{(\mathrm{ini})}$ used in (5.14) and (5.15). I used IS-NMF for both the training and separation stages. The number of iterations in both stages was set to 200, and the numbers of bases in $\boldsymbol{F}_{\mathrm{V}}$ and $\boldsymbol{F}_{\mathrm{O}}$ were set to 50. As a performance measure of the source separation, I used the improvement of SDR. The other experimental settings were the same as those in Sect. 5.4.1.

Results

Figure 5.8 shows the average SDR improvements of the 15 songs. The proposed methods achieve better separation performance than random initialization, as

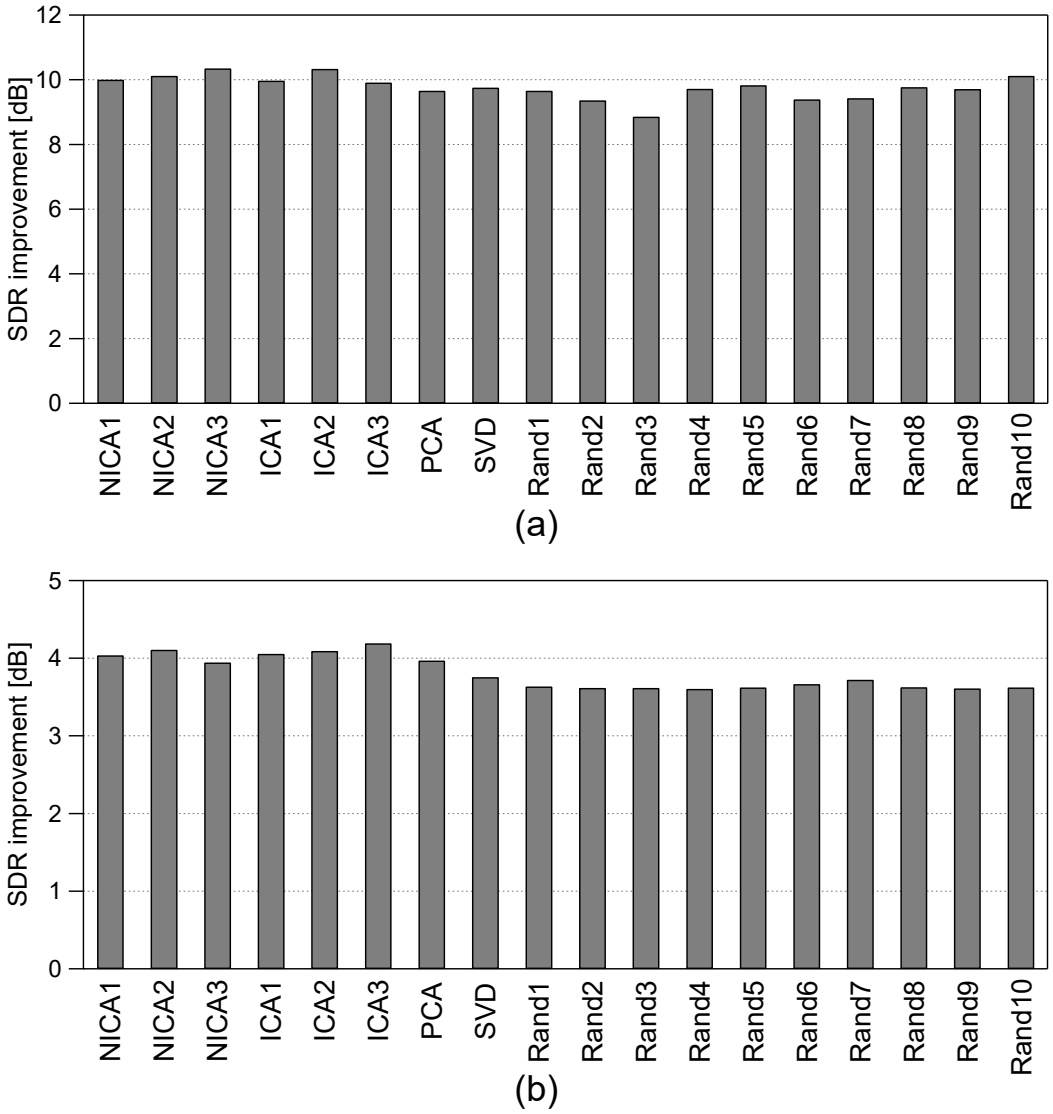

Figure 5.8: SDR improvement of supervised NMF for (a) vocals and (b) other.

particularly clearly shown in Fig. 5.8 (b). In supervised NMF, it is important to train the appropriate bases that represent only the corresponding source and do not represent interfering components. This result suggests that the ICA bases are preferable to the orthogonal bases, which can cover a wider convex polyhedral cone as shown in Fig. 5.2 (b).

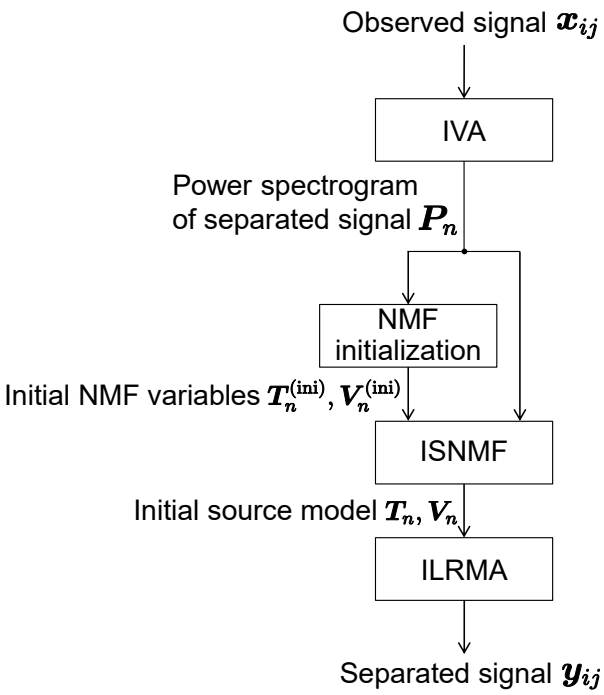

Figure 5.9: Process flow of BSS based on ILRMA with NMF initialization method.

5.5 Application to ILRMA and Discriminative SSNMF

The separation performance of NMF-based source separation always depends on the initial values of basis and activation matrices. Although the deep minimization of NMF cost function does not guarantee good separation performance, faster and stable optimization is preferable for many NMF-based algorithms. In this section, I investigate the availability of the proposed initializations for BSS based on ILRMA proposed in Chap. 3 and discriminative SSNMF proposed in Chap. 4.

5.5.1 BSS Based on ILRMA with Various Initialization for NMF

Figure 5.9 shows a process flow of initialization in ILRMA. Since the task is a blind situation, I first apply IVA to an observed multichannel signal $\boldsymbol{x}_{ij}$ to obtain

Table 5.2: Averaged SDR improvements over various speech signals and sources with same recording conditions, where these scores are obtained by ILRMA w/o partitioning function with various NMF initializations

Conditions (rev. time and mic. spacing)	NICA1	NICA2	NICA3	ICA1	ICA2	ICA3	PCA-based initialization	NNDSVD	Random initialization
130 ms & 1 m	5.13	4.13	5.18	12.40	7.07	**12.63**	10.27	10.21	11.91
130 ms & 5 cm	5.52	5.67	5.69	6.17	6.08	4.67	6.54	5.79	**8.97**
250 ms & 1 m	2.71	2.06	3.24	4.81	3.37	5.76	3.10	2.26	**7.34**
250 ms & 5 cm	5.89	5.40	5.61	5.74	5.85	4.73	5.75	5.04	**6.43**

a tentative separated signal $\boldsymbol{y}_{ij}$ and its power spectrogram $\boldsymbol{P}_n$. It is used for determining the initial basis and activation matrices $\boldsymbol{T}_n^{(\mathrm{ini})}$ and $\boldsymbol{V}_n^{(\mathrm{ini})}$ for each source. Then, the initial source model can be obtained by simple ISNMF. When I apply ILRMA, the demixing matrix $\boldsymbol{W}_i$ is reset to the identity matrix to avoid the poor local solution.

I conducted the same experiment in Sect. 3.6.3. Tables 5.2 and 5.3 show the result of average SDR improvement, where only ILRMA w/o partitioning function is performed and compared with various NMF initialization methods. The result of ILRMA using random initialization is the same as those in Tables 3.6 and 3.7, namely, those are the average scores of 10 trials with different pseudorandom seeds, and the other scores are the results of only the single trial. From these results, we can confirm that the initialization of source model in ILRMA does not lead better separation performance than the random source model for both speech and music signals. The reason may be that the estimates of IVA is one of the local minimum solution for ILRMA, and the pretrained source model does not drastically change from its initial values in some signals.

5.5.2 Discriminative SSNMF with Various Initialization for NMF

I here conducted the same experiment in Sect. 4.5.2 with proposed discriminative SSNMF, where the various NMF initialization methods are applied to (4.15) in the training stage. The other NMF variables $\boldsymbol{T}$, $\boldsymbol{V}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ used in (4.16) and (4.17) are initialized by random values, but the same random values are used for

Table 5.3: Averaged SDR improvements over various music signals and sources with same impulse response, where these scores are obtained by ILRMA w/o partitioning function with various NMF initializations

Impulse response	NICA1	NICA2	NICA3	ICA1	ICA2	ICA3	PCA-based initialization	NNDSVD	Random initialization
E2A	8.28	8.66	8.07	9.12	8.92	8.68	9.02	8.46	**14.41**
JR2	3.52	3.87	3.57	4.04	4.07	4.49	4.13	3.98	**9.06**

all the methods.

Figures 5.10–5.15 show the average scores of each data shown in Table 4.2, where the I showed only the best nonnegativization in NICA1–NICA3 and ICA1–ICA3 in each result for readability, but the other results also achieve comparable improvements with those. From these results, we can confirm that the proposed initialization clearly improves the separation performance of SSNMF, and these facts fit to the results in Fig. 5.8. Thus, we can guess that the good initialization for NMF may be effective particularly for the supervised source separation using NMF, such as FSNMF and SSNMF.

5.6 Summary

In this chapter, I addressed an efficient initialization method for NMF and proposed the utilization of ICA bases and estimated independent sources as the initial values of the basis and activation matrices, respectively. From an experimental comparison, some of the proposed method provides faster and deeper convergence of the NMF cost function than the conventional methods. Also, in supervised audio source separation, the proposed method achieves better performance than that obtained by random initialization.

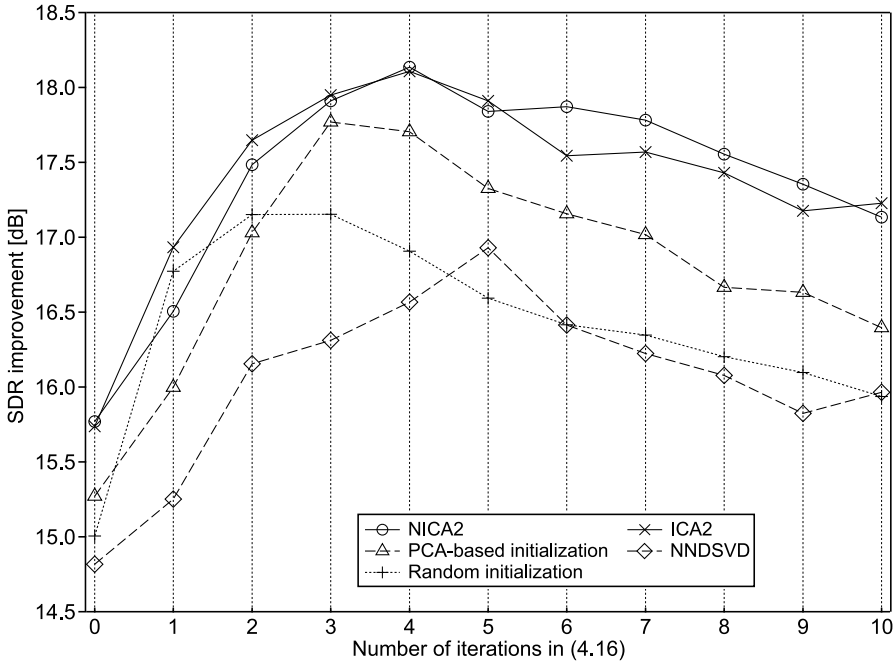

Figure 5.10: Average SDR improvement of ID1 for each number of iterations in (4.16) with various NMF initializations.

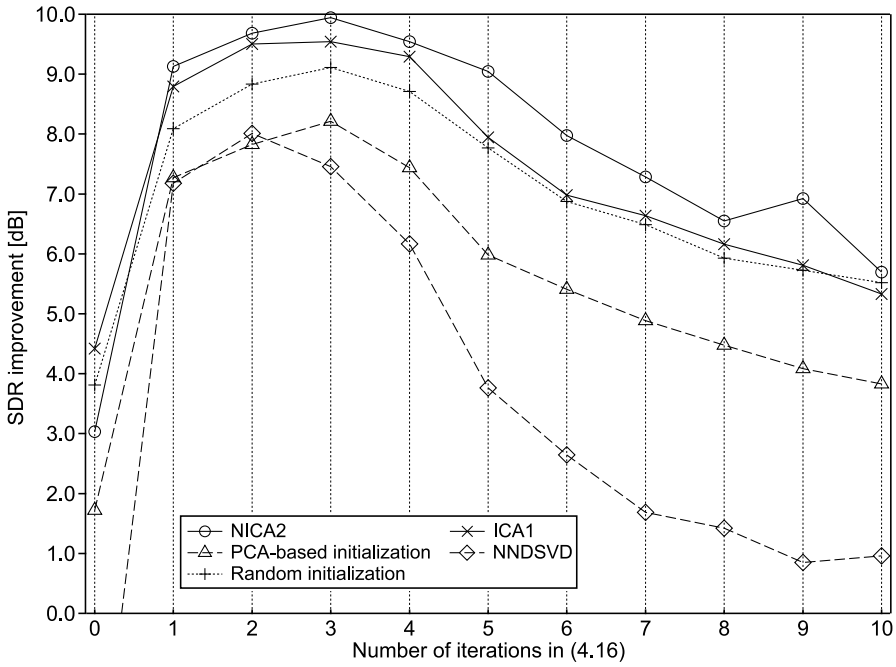

Figure 5.11: Average SDR improvement of ID2 for each number of iterations in (4.16) with various NMF initializations.

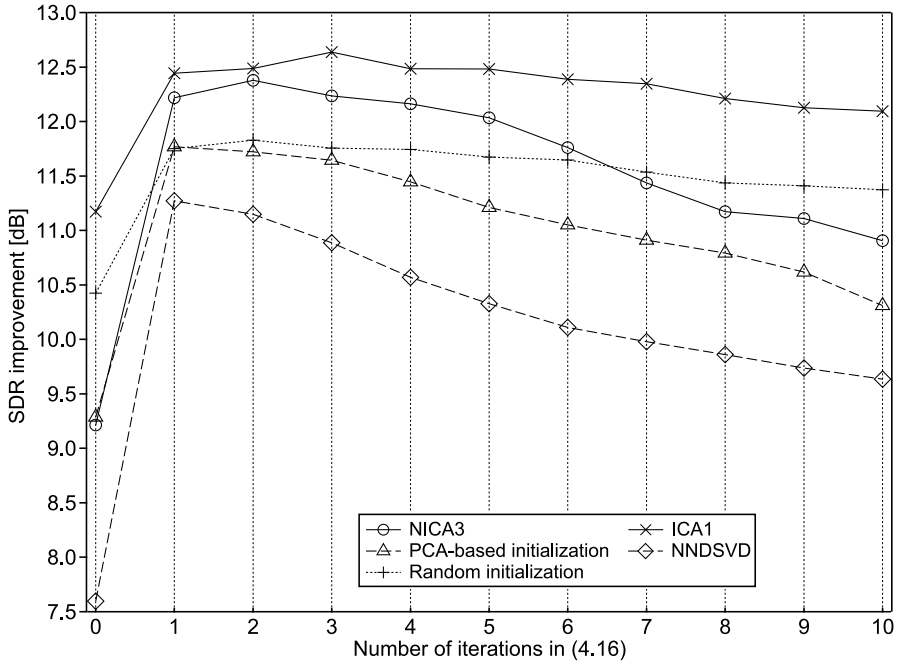

Figure 5.12: Average SDR improvement of ID3 for each number of iterations in (4.16) with various NMF initializations.

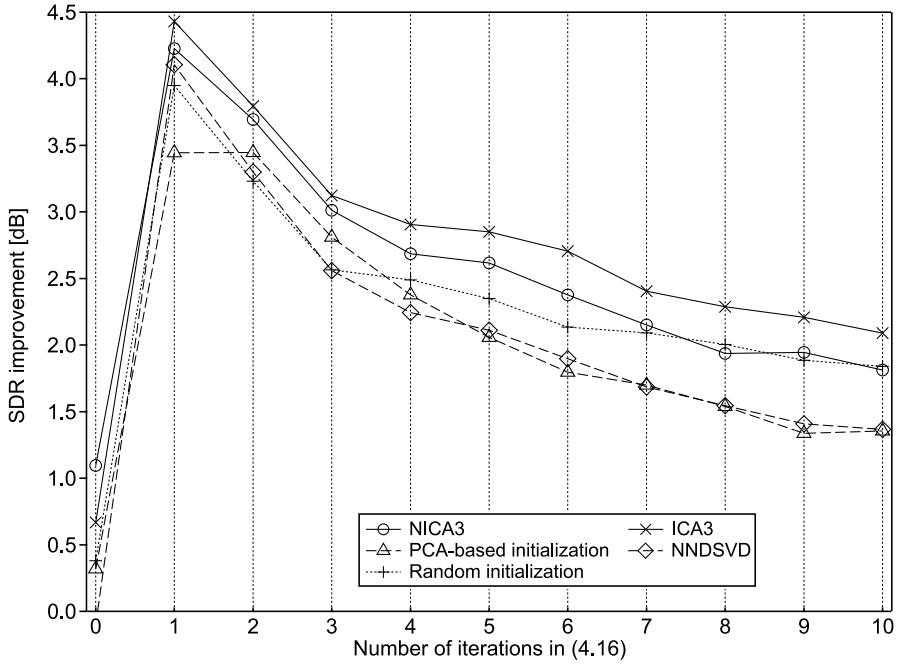

Figure 5.13: Average SDR improvement of ID4 for each number of iterations in (4.16) with various NMF initializations.

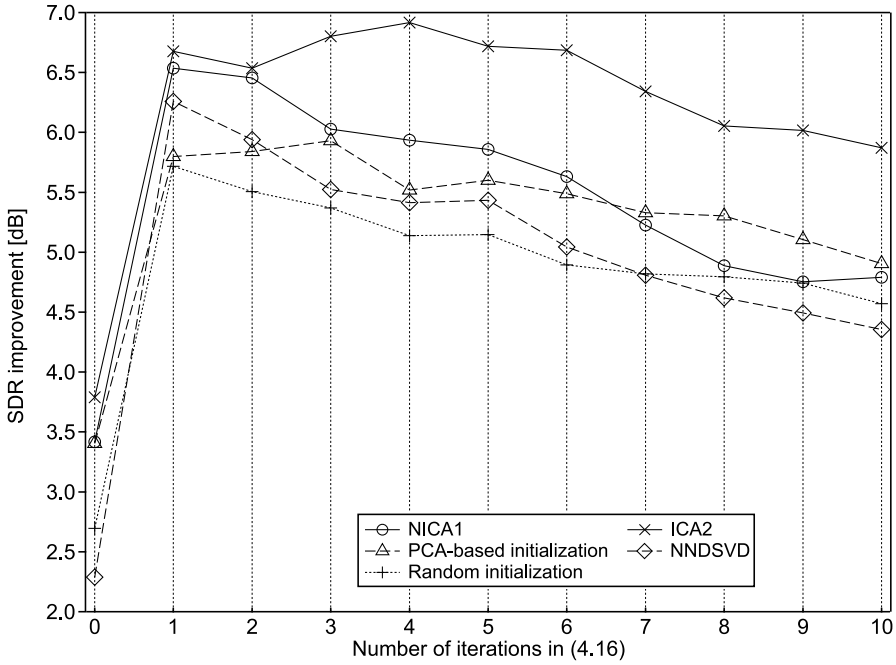

Figure 5.14: Average SDR improvement of ID5 for each number of iterations in (4.16) with various NMF initializations.

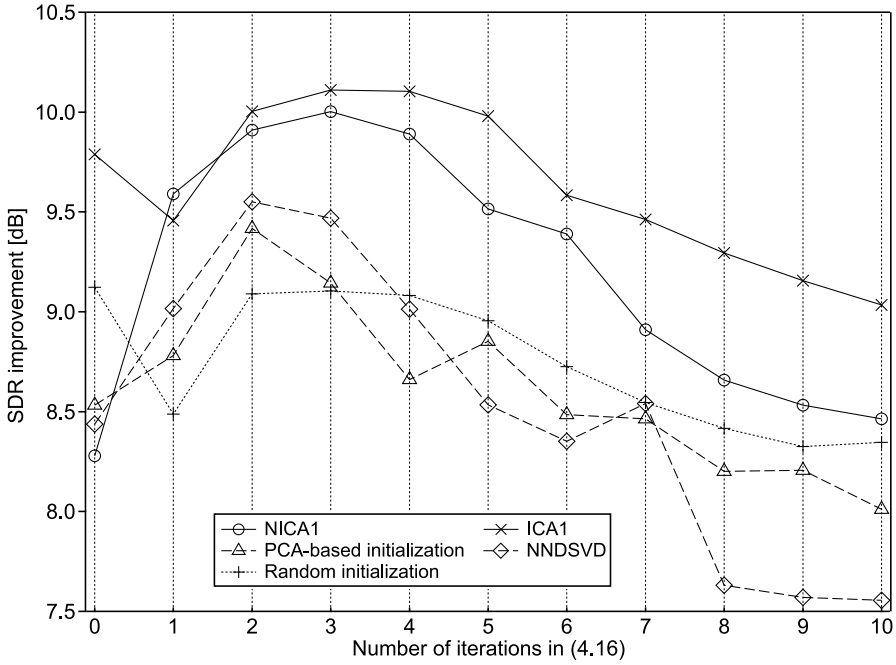

Figure 5.15: Average SDR improvement of ID6 for each number of iterations in (4.16) with various NMF initializations.

6 Conclusion

6.1 Summary of Dissertation

In this dissertation, I addressed a problem of music source separation, which can be applied to many valuable systems. This problem generally includes many situations, and I mainly dealt with following two main topics:

- determined (and overdetermined) BSS
- single-channel semi-supervised source separation

For both of them, I proposed new effective optimization algorithms based on NMF, which simultaneously achieve satisfactory separation performance and practical efficiency.

In Chap. 3, I proposed a new determined BSS algorithm called ILRMA. This method can be considered as a natural extension of traditional BSS algorithms, FDICA and IVA. In FDICA, the source model was a scalar random variable,

which obeys a non-Gaussian distribution. This scalar source model is extended to a multivariate vector variable as IVA, which enables a permutation-problem-free frequency domain separation. As a further extension of the source model, ILRMA employs NMF decomposition for capturing the low-rank time-frequency structure of the sources. Owing to the low-rankness of the music spectrogram, ILRMA can effectively model its structure and accurately estimate spatial demixing filters. Also, I revealed the intriguing relationships between IVA, MNMF, and ILRMA; MNMF with rank-1 spatial model is identical to ILRMA, and ILRMA with the single NMF basis for each source is essentially the same model as IVA. The optimization of ILRMA is based on the auxiliary function technique, resulting in a faster and more stable convergence than the conventional MNMF. The advantage of its separation performance was experimentally confirmed in both speech and music signals.

In Chap. 4, I proposed a new algorithm for discriminative training of NMF bases. This algorithm employs two types of supervised bases for one target source, which are called reconstructive and discriminative bases. The reconstructive bases include all the frequency components of the target source, and the discriminative bases consist of only the unique components to maximize the ability of discrimination from the other non-target sources. Whereas the discriminative bases training comes down to a bilevel optimization problem, I proposed a simple optimization algorithm for obtaining both reconstructive and discriminative bases instead. The efficacy of the proposed discriminative SSNMF was validated via semi-supervised music source separation task.

In Chap. 5, a general problem in NMF optimization was dealt. Since NMF is not a convex optimization problem, all the performance of NMF-based application always depend on the initial values of basis and activation matrices. As a remedy of this problem, I proposed a new efficient initialization method based on statistical independence. In this method, the estimates of ICA are utilized for the initial values of NMF variables, which are non-orthogonal bases and their coefficients. The proposed initialization achieved faster and deeper minimization than the conventional orthogonality-based initializations. In addition, the availability of the proposed initialization for determined BSS, single-channel FSNMF, and discriminative SSNMF was experimentally

investigated. The result showed that the proposed method can lead better separation for a supervised NMF approach.

6.2 Future Works

The following points still remain to be investigated or clarified.

- Since ILRMA has an NMF source model, we can easily employ some prior knowledges of each source for the optimization. The simple extension is a supervised ILRMA, which utilizes pretrained NMF bases of each source. As further supervision, external information such as music score or an annotation by users can be a good candidate for improving source model in ILRMA. However, to utilize such external information, the scale ambiguity among frequency bins in ILRMA must be solved in advance. This is because the scale ambiguity has a risk to distort the pretrained NMF bases or the given structure by the users. Some ICA optimizations without the scale ambiguity were proposed, for example, ICA with minimal distortion principle [236, 237], single-input and multiple-output ICA [238, 239, 240, 241], and multiple-input and single-output ICA [242]. These algorithms may be unified with ILRMA.

- Whereas a spectrogram of instrumental source has a low-rank time-frequency structure, a vocal spectrogram may not have the low-rank property as shown in Fig. 2.7 owing to its variety of pitches and phonemes. In the experiments presented in Sect. 3.6, ILRMA can achieve accurate source separation even for the mixture with a vocal source. This might be a collateral effect of an accurate modeling of the other instrumental source, namely, the better vocal separation might be achieved as a side effect of the accurate separation of remaining instrumental sources. Indeed, the speech separation based on ILRMA with a large number of bases could not give the highest performance of separation (see Fig. 3.14). For vocal and speech signals, another property may be useful to capture their time-frequency structures. In the literature [243, 244], robust PCA [245, 246] is used for extracting a vocal source from music signals, where robust PCA can

decompose an observed matrix into a low-rank structure and a sparse component, and the vocal source tends to be estimated as the sparse component. If we introduce such sparse source model to ILRMA for representing vocals, the music source separation would be improved.

- The cost function in ILRMA (3.46) is based on IS divergence, and the MU rules (3.54) and (3.55) are the same as those of simple ISNMF. The other popular criteria, KL divergence and EU distance, may also be used for the cost function in ILRMA. The source models corresponding these criteria (Poisson distribution for KL divergence and Gaussian distribution for EU distance) do not have a reproductive property. Thus, unlike ISNMF, the additivity of power spectrograms in expectation sense does not hold in KLNMF and EUNMF. As another source model with the reproductive property, NMF based on Cauchy distribution was proposed as Cauchy NMF [247], and it is reported that it gives slightly better separation performance than ISNMF. This source model can be used for an extension of ILRMA.

- In Chap. 4, I have dealt with only SSNMF with a new discriminative training method. However, the main idea in the proposed method can also be used for full-supervised situation. The comparison with the conventional discriminative NMF and the proposed algorithm in the full-supervised situation need to be investigated.

- The proposed algorithm of discriminative bases training is an approximation of the solution obtained by original bilevel optimization. This fact causes the decrease of SDR improvement at the converged point of algorithm as shown in Fig. 4.5. To avoid this problem, another approximative optimization or a stopping criteria must be considered for a practical use.

- The proposed algorithm of discriminative SSNMF can be combined with a penalized SSNMF [98], which forces the non-target bases to be different from the target bases.

Bibliography

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[3] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, 2006, pp. 601–608.

[4] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: an extension of ICA to multivariate components," in *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, 2006, pp. 165–172.

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[6] 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.

[7] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proceedings of IEEE International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 414–421.

[8] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Supervised non-Euclidean sparse NMF via bilevel optimization with applications to speech enhancement," in *Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 11–15.

[9] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Proceedings of Interspeech*, 2014, pp. 865–869.

[10] H. Chung, E. Plourde, and B. Champagne, "Discriminative training of NMF model based on class probabilities for speech enhancement," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 502–506, 2016.

[11] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

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

[13] 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.

[14] H. Kameoka, "Non-negative matrix factorization and its variants for audio signal processing," in *Applied Matrix and Tensor Variate Data Analysis*, T. Sakata, Ed. Springer Japan, 2016, ch. 2, pp. 21–50.

[15] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1998, pp. 1941–1944.

[16] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

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

[18] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[19] P. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer-Verlag London, 2000.

[20] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[21] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[22] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation and Modulation Theory*. Wiley-Interscience, 2002.

[23] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

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

[25] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Prentice Hall, 1989.

[26] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[27] B. Yang, "A study of inverse short-time Fourier transform," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 3541–3544.

[28] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[29] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, no. 4, pp. 112–114, 1997.

[30] T.-W. Lee, *Independent Component Analysis: Theory and Applications*. Springer US, 1998.

[31] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. Wiley-Interscience, 2000.

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

[33] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Wiley, 2002.

[34] H. Sawada, R. Mukai, S. Araki and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*. Springer Berlin Heidelberg, 2005, ch. 13, pp. 299–327.

[35] S. Makino, H. Sawada and S. Araki, "Frequency-domain blind source separation," in *Blind Speech Separation*. Springer Netherlands, 2007, ch. 2, pp. 47–78.

[36] T.-W. Lee, J. B. Anthony, and R. H. Lambert, "Blind separation of delayed and convolved sources," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 1997, pp. 758–764.

[37] K. Torkkola, "Blind separation for audio signals – are we there yet?" in *Proceedings of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA)*, 1999, pp. 11–15.

[38] K. Torkkola, "Blind separation of delayed and convolved sources," in *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*, S. Haykin, Ed. Wiley-Interscience, 2000, ch. 8, pp. 321–375.

[39] R. H. Lambert and C. L. Nikias, "Blind deconvolution of multipath mixtures," in *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*, S. Haykin, Ed. Wiley-Interscience, 2000, ch. 9, pp. 377–436.

[40] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume II: Blind Deconvolution)*. Wiley-Interscience, 2000.

[41] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Convolutive blind source separation for more than two sources in the frequency domain," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. III–885–III–888.

[42] H. Buchner, R. Aichner, and W. Kellerman, "A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2005.

[43] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[44] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[45] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[46] T. J. Shan and T. Kailath, "Adaptive beamforming for coherent signals and interference," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 3, pp. 527–536, 1985.

[47] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, pp. 1–10, 2003.

[48] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[49] O. Hoshuyama and A. Sugiyama, "Robust adaptive beamforming," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag Berlin Heidelberg, 2001, ch. 5, pp. 87–109.

[50] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 313–324, 2003.

[51] S. A. Vorobyov, A. B. Gershman, Z.-Q. Luo, and N. Ma, "Adaptive beamforming with joint robustness against mismatched signal steering vector and interference nonstationarity," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 108–111, 2004.

[52] J. Li and P. Stoica, *Robust adaptive beamforming*. John Wiley & Sons, 2005, vol. 88.

[53] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Letters*, vol. 6, no. 4, pp. 87–90, 1999.

[54] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

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

[56] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proceedings of International Computer Music Conference (ICMC)*, 2003, pp. 231–234.

[57] F. J. Theis, E. W. Lang, and C. G. Puntonet, "A geometric algorithm for overcomplete linear ICA," *Neurocomputing*, vol. 56, pp. 381–398, 2004.

[58] A. Jourjine, S. Rickard, and Ö. Yılmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 2000, pp. 2985–2988.

[59] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[60] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[61] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 436–443.

[62] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5210–5214.

[63] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each

frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.

[64] N. Roman, D.L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[65] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 147–150.

[66] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[67] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted non-negative matrix factorisation for sound source separation," in *Proceedings of IEEE Workshop on Statistical Signal Processing (SSP)*, 2005, pp. 1132–1137.

[68] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006, pp. 5479–5482.

[69] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2006, pp. 700–707.

[70] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[71] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 121–124.

[72] O. Dikmen and A. T. Cemgil, "Unsupervised single-channel source separation using Bayesian NMF," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 93–96.

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

[74] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 140–148.

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

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

[77] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proceedings of Irish Signals and Systems Conference (ISSC)*, 2005, pp. 8–12.

[78] R. M. Parry and I. A. Essa, "Estimating the spatial position of spectral components in audio," in *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, 2006, pp. 666–673.

[79] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2009, pp. 15–19.

[80] H. Kameoka, T. Yoshioka, M. Hamamura, J. L. Roux, and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 245–253.

[81] S. Arberet, A. Ozerov, N. Q. K. 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 *Proceedings of Information Sciences Signal Processing and their Applications (ISSPA)*, 2010, pp. 1–4.

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

[83] Y. Mitsufuji and A. Roebel, "Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 71–75.

[84] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "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.

[85] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *Proceedings of The International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 583–588.

[86] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 53–56.

[87] D. FitzGerald, "Vocal separation using nearest neighbours and median filtering," in *IET Irish Signals and Systems Conference (ISSC)*, 2012, pp. 1–5.

[88] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 1, pp. 73–84, 2013.

[89] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitch-based methods for background and melody separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1884–1893, 2014.

[90] Z. Rafii, A. Liutkus, and B. Pardo, "A simple user interface system for recovering patterns repeating in time and frequency in mixtures of sounds," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 271–275.

[91] A. Liutkus, Z. Rafii, B. Pardo, D. FitzGerald, and L. Daudet, "Kernel spectrogram models for source separation," in *Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 6–10.

[92] A. Liutkus, D. FitzGerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, 2014.

[93] D. FitzGerald, A. Liukus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/percussive separation using kernel additive modelling," in *Proceedings of IET Irish Signals & Systems Conference and China-Ireland International Conference on Information and Communications Technologies (ISSC/CIICT)*, 2013, pp. 35–40.

[94] A. Liutkus, D. FitzGerald, and Z. Rafii, "Scalable audio separation with light kernel additive modelling," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 76–80.

[95] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[96] P. D. O'grady and B. A. Pearlmutter, "Discovering convolutive speech phones using sparseness and non-negativity," in *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 520–527.

[97] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[98] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E97-A, no. 5, pp. 1113–1118, 2014.

[99] 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.

[100] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 45–48.

[101] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 129–132.

[102] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 888–891.

[103] N. Q. K. Duong, A. Ozerov, and L. Chevallier, "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization," in *Proceedings of IEEE International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, 2014, pp. 220–224.

[104] T. Ono, N. Ono, and S. Sagayama, "User-guided independent vector analysis with source activity tuning," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2417–2420.

[105] M. Fakhry and F. Nesta, "Underdetermined source detection and separation using a normalized multichannel spatial dictionary," in *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[106] F. Nesta and M. Fakhry, "Unsupervised spatial dictionary learning for sparse underdetermined multichannel source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 86–90.

[107] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[108] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7092–7096.

[109] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proceedings of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 577–581.

[110] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014, pp. 250–254.

[111] P.-S. Huang, M. Kim, M. H.-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proceedings of The International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 477–482.

[112] P.-S. Huang, M. Kim, M. H.-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[113] P.-S. Huang, M. Kim, M. H.-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[114] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 116–120.

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

[116] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training,"

IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 92–101, 2015.

[117] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4390–4394.

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

[119] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5542–5546.

[120] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Proceedings of Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 172–176.

[121] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 86, no. 3, pp. 590–596, 2003.

[122] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012, pp. 414–422.

[123] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[124] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE*

Transactions on Audio, Speech, and Language Processing, vol. 16, no. 4, pp. 766–778, 2008.

[125] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura–saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[126] M. Nakano, J. L. Roux, H. Kameoka, Y. Kitano, N. Ono, and S. Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 149–156.

[127] M. Nakano, J. L. Roux, H. Kameoka, T. Nakamura, N. Ono, S. Sagayama, "Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 325–328.

[128] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 5365–5368.

[129] K. Ochiai, H. Kameoka, and S. Sagayama, "Explicit beat structure modeling for non-negative matrix factorization-based multipitch analysis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 133–136.

[130] D. Kitamura, H. Saruwatari, Y. Iwao, K. Shikano, K. Kondo, and Y. Takahashi, "Superresolution-based stereo signal separation via supervised nonnegative matrix factorization," in *Proceedings of IEEE International Conference on Digital Signal Processing (DSP)*, 2013.

[131] D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, "Music signal separation by supervised nonnegative matrix factorization

with basis deformation," in *Proceedings of IEEE International Conference on Digital Signal Processing (DSP)*, 2013.

[132] D. Kitamura, H. Saruwatari, S. Nakamura, Y. Takahashi, K. Kondo, and H. Kameoka, "Divergence optimization in nonnegative matrix factorization with spectrogram restoration for multichannel signal separation," in *Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 92–96.

[133] H. Nakajima, D. Kitamura, N. Takamune, S. Koyama, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Music signal separation using supervised nmf with all-pole-model-based discriminative basis deformation," in *Proceedings of The European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1143–1147.

[134] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," in *Proceedings of Interspeech*, 2013, pp. 808–812.

[135] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3749–3753.

[136] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 66–70.

[137] K. Kwon, J. W. Shin, and N. S. Kim, "Discriminative training of NMF model based on class probabilities for speech enhancement," *IEICE Transactions on Information and Systems*, vol. E98-D, no. 11, pp. 2017–2020, 2015.

[138] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 267–273.

[139] M. W. Berry and M. Browne, "Email surveillance using non-negative matrix factorization," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 249–264, 2005.

[140] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, vol. 1, 2005, pp. 370–377.

[141] D. Soukup and I. Bajla, "Robust object recognition under partial occlusions using NMF," *Computational Intelligence and Neuroscience*, vol. 2008, 2008.

[142] M. M. Kalayeh, H. Idrees, and M. Shah, "NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 184–191.

[143] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 19–60, 2010.

[144] R. Sandler and M. Lindenbaum, "Nonnegative matrix factorization with earth mover's distance metric for image analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1590–1602, 2011.

[145] V. Monga and M. K. Mihçak, "Robust and secure image hashing via non-negative matrix factorizations." *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3-1, pp. 376–390, 2007.

[146] W. Lu, W. Sun, and H. Lu, "Robust watermarking based on DWT and nonnegative matrix factorization," *Computers & Electrical Engineering*, vol. 35, no. 1, pp. 183–188, 2009.

[147] Z. Chen, A. Cichocki, and T. M. Rutkowski, "Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer disease," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5, 2006, pp. 893–896.

[148] H. Lee, A. Cichocki, and S. Choi, "Kernel nonnegative matrix factorization for spectral EEG feature extraction," *Neurocomputing*, vol. 72, no. 13, pp. 3182–3190, 2009.

[149] C. Damon, A. Liutkus, A. Gramfort, and S. Essid, "Non-negative matrix factorization for single-channel EEG artifact rejection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 1177–1181.

[150] C. Damon, A. Liutkus, A. Gramfort, and S. Essid, "Non-negative tensor factorization for single-channel EEG artifact rejection," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[151] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[152] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005.

[153] D. Greene, G. Cagney, N. Krogan, and P. Cunningham, "Ensemble non-negative matrix factorization methods for clustering protein–protein interactions," *Bioinformatics*, vol. 24, no. 15, pp. 1722–1728, 2008.

[154] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Proceedings of IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, 2007, pp. 431–436.

[155] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proceedings of Interspeech*, 2008, pp. 411–414.

[156] J. L. Roux, H. Kameoka, N. Ono, A. D. Cheveigne, and S. Sagayama, "Computational auditory induction as a missing-data model-fitting

problem with Bregman divergence," *Speech Communication*, vol. 53, no. 5, pp. 658–676, 2011.

[157] K. Y. Yılmaz, A. T. Cemgil, and U. Şimşekli, "Generalised coupled tensor factorisation," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2151–2159.

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

[159] S. A. Abdallah and M. D. Plumbley, "Polyphonic music transcription by non-negative sparse coding of power spectra," in *Proceedings of The International Society for Music Information Retrieval Conference (ISMIR)*, 2004, pp. 318–325.

[160] E. Vincent, N. Bertin, and R. Badeau, "Two nonnegative matrix factorization methods for polyphonic pitch transcription," in *Proceedings of Music Information Retrieval Evaluation eXchange (MIREX)*, 2007.

[161] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 109–112.

[162] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 538–549, 2010.

[163] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[164] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, pp. 146–158, 1975.

[165] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proceedings of International Congress on Acoustics (ICA)*, 1968, pp. C–17–C–20.

[166] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," Institute of Statistical Mathematics, Tech. Rep., 2001.

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

[168] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[169] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[170] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 60–77, 2000.

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

[172] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[173] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[174] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[175] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 8, 1996, pp. 757–763.

[176] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-gaussian sources," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 165–172.

[177] T. Adali, H. Ki, and J.-F. Cardoso, "Complex ICA using nonlinear functions," *IEEE Transactions on Signal Processing*, vol. 56, no. 9, pp. 4536–4544, 2008.

[178] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2000, pp. 3140–3143.

[179] 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.

[180] H. Sawada, R. Mukai, S. Araki, and S. Makino,, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[181] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, 2007, pp. 3247–3250.

[182] S. Kotz, T. J. Kozubowski, and K. Podgórski, "Symmetric multivariate laplace distribution," in *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Birkhäuser Basel, 2001, ch. 5, pp. 231–238.

[183] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate Laplace distribution," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 300–303, 2006.

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

[185] N. Ono, "Fast stereo independent vector analysis and its implementation on mobile phone," in *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[186] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012.

[187] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1293–1302, 1993.

[188] A. T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill, "Prior structures for time-frequency energy distributions," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 151–154.

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

[190] K. Hild, H. T. Attias, and S. Nagarajan, "An expectation-maximization method for spatio-temporal blind source separation using an AR-MOG source model," *IEEE Transactions on Neural Networks*, vol. 19, no. 3, pp. 508–519, 2008.

[191] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "An expectation-maximization method for spatio-temporal blind source separation using an AR-MOG source model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1512–1527, 2008.

[192] A. R. López, N. Ono, U. Remes, K. Palomäki, and M. Kurimo, "Designing multichannel source separation based on single-channel source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 469–473.

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

[194] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[195] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 129–132.

[196] N. Q. K. 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.

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

[198] W. James and C. Stein, "Estimation with quadratic loss," in *Proceedings of Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1961, pp. 361–379.

[199] B. Kulis, M. Sustik, and I. Dhillon, "Learning low-rank kernel matrices," in *Proceedings of International Conference on Machine Learning (ICML)*, 2006, pp. 505–512.

[200] Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musical-noise analysis in methods of integrating microphone array and spectral subtraction based on higher-order statistics," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, 2010, article ID 431347.

[201] 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.

[202] 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 *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, 2000, pp. 965–968.

[203] P. D. O'Grady and B. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proceedings of Independent Component Analysis and Blind Signal Separation (ICA)*, 2004, pp. 430–436.

[204] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proceedings F - Radar and Signal Processing*, vol. 140, no. 6, pp. 362–370, 1993.

[205] D. B. Ward, R. A. Kennedy, and R. C. Williamson, "Constant directivity beamforming," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer-Verlag Berlin Heidelberg, 2001, ch. 1, pp. 3–17.

[206] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for

convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[207] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," in *Proceedings of The European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1271–1275.

[208] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 45–48.

[209] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe-cracking," in *Proceedings of the Symposium on Time Series Analysis*, M. Rosenblatt, Ed. Wiley, New York, 1963, ch. 15, pp. 209–243.

[210] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proceedings of the IEEE*, vol. 65, no. 10, pp. 1428–1443, 1977.

[211] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[212] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, vol. 116, pp. 374–388, 1976.

[213] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[214] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, no. Nov, pp. 1457–1469, 2004.

[215] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1705–1713.

[216] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4206–4209.

[217] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proceedings of Interspeech*, 2010, pp. 717–720.

[218] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing," in *Proceedings of IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2013, pp. 392–397.

[219] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Annals of Operations Research*, vol. 153, no. 1, pp. 235–256, 2007.

[220] M. D. Plumbley, "Algorithms for nonnegative independent component analysis," *IEEE Transactions on Neural Networks*, vol. 14, no. 3, pp. 534–543, 2003.

[221] E. Oja and M. D. Plumbley, "Blind separation of positive sources by globally convergent gradient search," *Neural Computation*, vol. 16, no. 9, pp. 1811–1825, 2004.

[222] Z. Yuan and E. Oja, "A FastICA algorithm for non-negative independent component analysis," in *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2004, pp. 1–8.

[223] C. Boutsidis and E. Gallopoulos, "SVD based initialization: a head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008.

[224] K. Stadlthanner, F. J. Theis, E. W. Lang, A. M. Tomé, C. G. Puntonet, P. G. Vilda, T. Langmann, and G. Schmitz, "Sparse nonnegative matrix factorization applied to microarray data sets," in *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2006, pp. 254–2618.

[225] A. Janecek and Y. Tan, "Using population based algorithms for initializing nonnegative matrix factorization," in *Proceedings of International Conference in Swarm Intelligence (ICSI)*, 2011, pp. 307–316.

[226] Y. Xue, C. S. Tong, Y. Chen, and W. S. Chen, "Clustering-based initialization for non-negative matrix factorization," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 525–536, 2008.

[227] Z. Zheng, J. Yang, and Y. Zhu, "Initialization enhancer for non-negative matrix factorization," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 1, pp. 101–110, 2007.

[228] M. Rezaei, R. Boostani, and M. Rezaei, "An efficient initialization method for nonnegative matrix factorization," *Journal of Applied Sciences*, vol. 11, no. 2, pp. 354–359, 2011.

[229] G. Casalino, N. D. Buono, and C. Mencar, "Subtractive clustering for seeding non-negative matrix factorizations," *Journal of Applied Sciences*, vol. 257, pp. 369–387, 2014.

[230] L. Zhao, G. Zhuang, and X. Xu, "Facial expression recognition based on PCA and NMF," in *Proceedings of World Congress on Intelligent Control and Automation (WCICA)*, 2008, pp. 6826–6829.

[231] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2003.

[232] C. Bauckhage, "A purely geometric approach to non-negative matrix factorization," in *Proceedings of LWA*, 2014, pp. 125–136.

[233] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," *Advances in Models for Acoustic Processing, NIPS*, vol. 148, 2006.

[234] Y. Kawaguchi, M. Togami, H. Nagano, Y. Hashimoto, M. Sugiyama, and Y. Takada, "Ica-based acceleration of probabilistic latent component analysis for mass spectrometry-based explosives detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 2795–2799.

[235] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2015, pp. 387–395.

[236] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proceedings of International Symposium on ICA and BSS*, 2001, pp. 722–727.

[237] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proceedings of the 41st SICE Annual Conference SICE 2002*, vol. 4, 2002, pp. 2138–2143.

[238] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "High-fidelity blind source separation of acoustic signals using SIMO-model-based ICA with information-geometric learning," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2003, pp. 251–254.

[239] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind separation of binaural sound mixtures using SIMO-model-based independent component analysis," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. IV, 2004, pp. 113–116.

[240] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "High-fidelity blind separation of acoustic signals using SIMO-model-based independent component analysis," *IEICE Transactions on Fundamentals of Electronics,*

Communications and Computer Sciences, vol. E87-A, no. 8, pp. 2063–2072, 2004.

[241] T. Takatani, S. Ukai, T. Nishikawa, H. Saruwatari, and K. Shikano, "A selfgenerator method for initial filters of SIMO-ICA applied to blind separation of binaural sound mixtures," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A, no. 7, pp. 1673–1682, 2005.

[242] K. Matsuoka and S. Nakashima, "Overdetermined blind separation of acoustic signals based on MISO-constrained frequency-domain ICA," in *Proceedings of International Congress on Acoustics (ICA)*, 2004, pp. IV–3143–IV–3146.

[243] P. S. Huang, S. D. Chen, P. Smaragdis, and M. H.-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 57–60.

[244] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent F0 estimation and source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 574–578.

[245] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 2080–2088.

[246] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2010, pp. 2496–2504.

[247] A. Liutkus, D. FitzGerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015.

A

Derivation of Shape Parameter for Artificial Random Spectrogram with Constant Kurtosis

To produce an artificial random spectrogram $\boldsymbol{F}\boldsymbol{G}$ with constant kurtosis, we derive the optimal shape parameter κ for each value of $\mathcal{R}$. Hereafter, we denote an $I \times J$ matrix whose elements are $f_{ir}g_{rj}$ as $\boldsymbol{F}_r\boldsymbol{G}_r$, namely, $\boldsymbol{F}\boldsymbol{G} = \sum_r \boldsymbol{F}_r\boldsymbol{G}_r$. Also, we denote a pth-order moment and pth-order cumulant of $\boldsymbol{F}_r\boldsymbol{G}_r$ as μ_{pr} and c_{pr} and those of $\boldsymbol{F}\boldsymbol{G}$ as μ'_p and c'_p, respectively. When $\mathcal{R}$ increases beyond one, the matrix $\boldsymbol{F}\boldsymbol{G}$ becomes a linear combination expressed as $\sum_{r=1}^{\mathcal{R}} \boldsymbol{F}_r\boldsymbol{G}_r$. Therefore, the kurtosis of $\boldsymbol{F}\boldsymbol{G}$ can be derived via the moment-cumulant transform [200]. Since f_{ir} and g_{rj} are generated from i.i.d. gamma distributions, μ_{pr} is equal to the

product of pth-order moments of $\boldsymbol{F}_r$ and $\boldsymbol{G}_r$ as follows:

$$\mu_{pr} = \theta^{2p} \prod_{q=0}^{p-1} (\kappa + q)^2. \tag{A.1}$$

By the moment-cumulant transform, c_{pr} from $p = 1$ to $p = 4$ can be represented as

$$c_{1r} = \mu_{1r}, \tag{A.2}$$
$$c_{2r} = \mu_{2r} - \mu_{1r}^2, \tag{A.3}$$
$$c_{3r} = \mu_{3r} - 3\mu_{1r}\mu_{2r} + 2\mu_{1r}^3, \tag{A.4}$$
$$c_{4r} = \mu_{4r} - 4\mu_{1r}\mu_{3r} - 3\mu_{2r}^2 + 12\mu_{1r}^2\mu_{2r} - 6\mu_{1r}^4. \tag{A.5}$$

Since a cumulant satisfies additivity for the variables, the cumulant of $\boldsymbol{FG}$ is easily derived as follows:

$$c'_p = \sum_{r=1}^{R} c_{pr} = Rc_{pr}. \tag{A.6}$$

The moments of $\boldsymbol{FG}$ for $p = 2$ and $p = 4$ are given by the moment-cumulant transform as

$$\mu'_1 = c'_1, \tag{A.7}$$
$$\mu'_2 = c'_2 + c'^2_1, \tag{A.8}$$
$$\mu'_3 = c'_3 + 3c'_1c'_2 + c'^3_1, \tag{A.9}$$
$$\mu'_4 = c'_4 + 3c'^2_2 + 4c'_1c'_3 + 6c'^2_1c'_2 + c'^4_1. \tag{A.10}$$

Finally, the kurtosis of $\boldsymbol{FG}$ can be derived as

$$\text{kurtosis}(\boldsymbol{FG}) = \frac{\mu'_4}{\mu'^2_2} = \frac{\zeta(\kappa, \mathcal{R})}{\xi(\kappa, \mathcal{R})}. \tag{A.11}$$

Therefore, by solving (3.92), we can obtain the shape parameter κ so that the kurtosis of $\boldsymbol{FG}$ has the same value (kurt) for any value of $\mathcal{R}$.

Publication List

Book Chapters

[B.1] **Daichi Kitamura**, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," *Audio Source Separation*, Shoji Makino, Eds. Springer, 31 pages (to appear).

[B.2] **Daichi Kitamura**, "Q11 ビームフォーミングって何ですか？," *Acousticpedia for Beginners*, Acoustical Society of Japan, Eds. Corona Publishing, pp. 44–47, March 2017 (in Japanese).

Journal Papers

[J.1] **Daichi Kitamura**, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, September, 2016.

[J.2] Yoshiaki Bando, Hiroshi Saruwatari, Nobutaka Ono, Shoji Makino, Katsutoshi Itoyama, **Daichi Kitamura**, Masaru Ishimura, Moe Takakusaki, Narumi Mae, Kouei Yamaoka, Yutaro Matsui, Yuichi Ambe, Masashi Konyo, Satoshi Tadokoro, Kazuyoshi Yoshii, and Hiroshi G. Okuno, "Low-latency and high-quality two-stage human-voice-enhancement system for

a hose-shaped rescue robot," *Journal of Robotics and Mechatronics*, vol. 27, no. 1, pp.198–212, February 2017.

Peer-Reviewed International Conference Proceedings

[C.1] **Daichi Kitamura**, Nobutaka Ono, and Hiroshi Saruwatari, "Experimental analysis of optimal window length for independent low-rank matrix analysis," in *The 2017 European Signal Processing Conference (EUSIPCO)*, Kos, Greece, September 2017 (invited special session, submitted).

[C.2] **Daichi Kitamura** and Nobutaka Ono, "Efficient initialization for nonnegative matrix factorization based on nonnegative independent component analysis," in *Proceedings of The 15th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, September 2016.

[C.3] **Daichi Kitamura**, Nobutaka Ono, Hiroshi Saruwatari, Yu Takahashi, and Kazunobu Kondo, "Discriminative and reconstructive basis training for audio source separation with semi-supervised nonnegative matrix factorization," in *Proceedings of The 15th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, September 2016.

[C.4] **Daichi Kitamura**, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," in *Proceedings of The 2015 European Signal Processing Conference (EUSIPCO)*, pp. 1271–1275, Nice, France, September 2015 (invited special session).

[C.5] **Daichi Kitamura**, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 276–280, Brisbane, Australia, April 2015.

[C.6] Yoshiki Mitsui, **Daichi Kitamura**, Shinnosuke Takamichi, Nobutaka Ono, and Hiroshi Saruwatari, "Blind source separation based on independent

low-rank matrix analysis with sparse regularization for time-series activity," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, New Orleans, U.S.A., March, 2017.

[C.7] Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, **Daichi Kitamura**, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave, "The 2016 signal separation evaluation campaign," in *Proceedings of 13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 323–332, Grenoble, France, February, 2017.

[C.8] Narumi Mae, Masaru Ishimura, **Daichi Kitamura**, Nobutaka Ono, Takeshi Yamada, Shoji Makino, and Hiroshi Saruwatari, "Ego noise reduction for hose-shape rescue robot combining independent low-rank matrix analysis and multichannel noise cancellation," in *Proceedings of 13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 141–151, Grenoble, France, February, 2017.

[C.9] Narumi Mae, **Daichi Kitamura**, Masaru Ishimura, Takeshi Yamada, and Shoji Makino, "Ego noise reduction for hose-shaped rescue robot combining independent low-rank matrix analysis and noise cancellation," in *Proceedings of Asia-Pacific Signal and Information Process Processing Association Annual Summit and Conference (APSIPA ASC)*, Jeju, Korea, December 2016

[C.10] Hiroaki Nakajima, **Daichi Kitamura**, Norihiro Takamune, Shoichi Koyama, Hiroshi Saruwatari, Yu Takahashi, and Kazunobu Kondo, "Audio signal separation using supervised NMF with time-variant all-pole-model-based basis deformation," in *Proceedings of Asia-Pacific Signal and Information Process Processing Association Annual Summit and Conference (APSIPA ASC)*, Jeju, Korea, December 2016

[C.11] Moe Takakusaki, **Daichi Kitamura**, Nobutaka Ono, Takeshi Yamada, Shoji Makino, and Hiroshi Saruwatari, "Ego-noise reduction for a hose-shaped rescue robot using determined rank-1 multichannel nonnegative matrix factorization," in *Proceedings of The 15th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, September 2016.

[C.12] Hiroaki Nakajima, **Daichi Kitamura**, Norihiro Takamune, Shoichi Koyama, Hiroshi Saruwatari, Nobutaka Ono, Yu Takahashi, and Kazunobu Kondo, "Music signal separation using supervised NMF with all-pole-model-based discriminative basis deformation," in *Proceedings of The 2016 European Signal Processing Conference (EUSIPCO)*, pp.1143–1147, Budapest, Hungary, August 2016.

[C.13] Nobutaka Ono, Zafar Rafii, **Daichi Kitamura**, Nobutaka Ito, and Antoine Liutkus, "The 2015 signal separation evaluation campaign," in *Proceedings of 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Liberec, Czech, August 2015.

Non-Reviewed Domestic Workshop Proceedings and Technical Reports

[D.1] **Daichi Kitamura**, Nobutaka Ono, Hiroshi Saruwatari, Yu Takahashi, and Kazunobu Kondo, "Effective basis learning for sound source separation by semi-supervised nonnegative matrix factorization," *IEICE Technical Report*, EA2015-130, vol. 115, no. 521, pp. 355–360, Oita, March 2016 (in Japanese).

[D.2] **Daichi Kitamura** and Nobutaka Ono, "Statistical-independence-based effective initialization for nonnegative matrix factorization," *Proceedings of 2016 Spring Meeting of Acoustical Society of Japan (ASJ)*, 3-3-5, pp. 619–622, Kanagawa, March 2016 (in Japanese).

[D.3] **Daichi Kitamura**, Hiroshi Saruwatari, Nobutaka Ono, Hiroshi Sawada, and Hirokazu Kameoka, "Study on source and spatial models for BSS with rank-1 spatial approximation," *Proceedings of 2015 Autumn Meeting of Acoustical Society of Japan (ASJ)*, 3-6-10, pp. 583–586, Fukushima, September 2015 (in Japanese).

[D.4] **Daichi Kitamura**, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Relaxation of rank-1 spatial model in overdeter-

mined BSS," *Proceedings of 2015 Spring Meeting of Acoustical Society of Japan (ASJ)*, 3-10-11, pp. 629–632, Tokyo, March 2015 (in Japanese).

[D.5] **Daichi Kitamura**, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Efficient multichannel nonnegative matrix factorization with rank-1 spatial model," *Proceedings of 2014 Autumn Meeting of Acoustical Society of Japan (ASJ)*, 2-1-11, pp. 579–582, Hokkaido, September 2014 (in Japanese).

[D.6] **Daichi Kitamura**, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Performance evaluation of source separation in multichannel nonnegative matrix factorization with rank-1 spatial model," *Proceedings of 107th IPSJ Special Interest Group on Music and Computer (IPSJ-SIGMUS)*, vol. 2015-MUS-107, no. 31, Tokyo, May 2015 (in Japanese).

[D.7] Yoshiki Mitsui, Satoshi Mizoguchi, Hiroshi Saruwatari, **Daichi Kitamura**, Nobutaka Ono, Masaru Ishimura, Narumi Mae, Moe Takakusaki, Yutaro Matsui, Kouei Yamaoka, and Shoji Makino, "Development of blind source separation system for flexible hose-shaped robot using independent low-rank matrix analysis and statistical speech enhancement," *Proceedings of 2017 Spring Meeting of Acoustical Society of Japan (ASJ)*, 1-P-3, pp. 517–518, Kanagawa, March 2017 (in Japanese).

[D.8] Yoshiki Mitsui, **Daichi Kitamura**, Shinnosuke Takamichi, Nobutaka Ono, and Hiroshi Saruwatari, "Study on efficient solver for independent low-rank matrix analysis with sparse time-series-activity regularization," *Proceedings of 2016 Autumn Meeting of Acoustical Society of Japan (ASJ)*, 1-7-3, pp. 325–328, Toyama, September 2016 (in Japanese).

[D.9] Yoshiki Mitsui, **Daichi Kitamura**, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Blind source separation using independent low-rank matrix analysis with sparse regularization for time-series activations," *IEICE Technical Report*, EA2016-72, vol. 116, no. 420, pp. 25–30, Kyoto, January 2017 (in Japanese).

[D.10] Moe Takakusaki, **Daichi Kitamura**, Nobutaka Ono, Takeshi Yamada, Shoji Makino, and Hiroshi Saruwatari, "Noise reduction for a hose-shaped rescue robot using rank-1 multichannel NMF," *Proceedings of The 2016 JSME Conference on Robotics and Mechatronics (ROBOMECH 2016)*, no. 16-2, 1A2-10a3, Kanagawa, June 2016 (in Japanese).

[D.11] Moe Takakusaki, **Daichi Kitamura**, Nobutaka Ono, Takeshi Yamada, Shoji Makino, and Hiroshi Saruwatari, "Noise reduction using rank-1 multichannel NMF for a hose-shaped rescue robot," *Proceedings of 2016 IEICE General Conference*, A-5-1, Fukuoka, March 2016 (in Japanese).

[D.12] Kazuma Takata, **Daichi Kitamura**, Hiroaki Nakajima, Shoichi Koyama, Hiroshi Saruwatari, Nobutaka Ono, and Shoji Makino, "Source separation with supervised multichannel NMF and statistical speech enhancement method for hose-shaped rescue robot," *Proceedings of 2016 Spring Meeting of Acoustical Society of Japan (ASJ)*, 3-3-2, pp. 609-612, Kanagawa, March 2016 (in Japanese).

[D.13] Hiroaki Nakajima, **Daichi Kitamura**, Norihiro Takamune, Shoichi Koyama, Hiroshi Saruwatari, Nobutaka Ono, Yu Takahashi, and Kazunobu Kondo, "Music signal separation using supervised NMF with time-valiant all-pole-model-based basis deformation," *Proceedings of 2016 Spring Meeting of Acoustical Society of Japan (ASJ)*, 3-3-11, pp. 635-638, Kanagawa, March 2016 (in Japanese).

[D.14] Hiroaki Nakajima, **Daichi Kitamura**, Norihiro Takamune, Shoich Koyama, Hiroshi Saruwatari, Nobutaka Ono, Yu Takahashi, and Kazunobu Kondo, "Analysis on degree of freedom for supervised NMF with all-pole-model-based basis deformation," *IEICE Technical Report*, EA2015-42, vol. 115, no. 359, pp. 13–18, Kanazawa, December 2015 (in Japanese).

Awards Received

1. The 7th Ikushi Prize from Japan Society for the Promotion of Science (JSPS), March 2017.
2. The Telecom System Technology Student Award from The Telecommunication Advancement Foundation (TAF), March 2016.
3. Student Conference Paper Award from IEEE Signal Processing Society (SPS) Japan Chapter, November 2015.
4. 2015 The 1st Best Student Award from National Institute of Informatics (NII), September 2015.
5. The 37th Awaya Prize Young Researcher Award from The Acoustical Society of Japan (ASJ), March 2015.
6. 2013 Technical Group on Signal Processing Young Researcher's Award from The Institute of Electronics, Information and Communication Engineers (IEICE), November 2014.