



卒業研究論文

論文題目

一般化ガウス分布に基づく非負値行列因子分解を用いた音楽音源分離

提出年月日	平成31年 2月 27日
学 科	電 気 情 報 工 学 科
氏 名	田 中 勇 気 印
指導教員(主査)	北 村 大 地 印
副 査	重 田 和 弘 印
学 科 長	辻 正 敏 印

香川高等専門学校

Music Source Separation Using Nonnegative Matrix Factorization Based on Complex Generalized Gaussian Distribution

Yuki Tanaka

Department of Electrical and Computer Engineering
National Institute of Technology, Kagawa College

Abstract

In this thesis, I aim to improve the performance of monaural audio source separation using simple sample signals of each source as supervised information. In monaural audio source separation, nonnegative matrix factorization (NMF), which is a mathematical algorithm, is a popular method and has been used for many years. Recently, a new NMF algorithm based on the complex generalized Gaussian distribution was proposed (GGD-NMF). However, the application of GGD-NMF to the supervised audio source separation has never been reported, and its efficacy is not clarified. In this thesis, I compare the performance of GGD-NMF-based supervised audio source separation and that of a conventional method and reveal the potential of GGD-NMF. The experimental results show that GGD-NMF can outperform the conventional method in many cases although its performance strongly depends on parameters. Also, I propose a new cosine-similarity-based criterion for predicting separation performance, which enables us to determine the optimal parameters to obtain better separation results.

Key Words: audio source separation, nonnegative matrix factorization, complex generalized Gaussian distribution, supervised method

(和訳)

本論文では、音源の短いサンプル信号のような簡易的な教師情報を活用するモノラル信号の音源分離技術を対象として、性能の向上を目指す。モノラル信号の音源分離では、非負値行列因子分解 (NMF) と呼ばれる数理アルゴリズムに基づく手法が有名であり、長く研究されている。近年、新しい生成モデルに基づく NMF として、複素一般化ガウス分布を仮定したアルゴリズム (GGD-NMF) が提案されているが、本手法の教師あり音源分離への適用例は未だ報告されておらず、調査が不十分である。そこで本論文では、GGD-NMF を用いた教師あり音源分離の性能を従来手法と比較し、その有効性を実験的に確認する。実験結果より、多くの混合信号において、GGD-NMF による音源分離が既存手法の性能を上回ることが明らかとなった。ただし、音源分離にとって最適な GGD-NMF のパラメータは、混合されている音源の種類や組み合わせに強く依存することが新たに示された。この問題を解決するために、音源分離性能を予測するための基準として、 $\cos$ 類似度を用いることを新たに提案し、この基準が音源分離性能と相関関係を持つことを実験的に示す。

目次

1. はじめに	2
2. NMF を用いた音源分離	5
2.1 短時間フーリエ変換	5
2.2 NMF	6
2.3 教師あり NMF	7
2.4 類似度関数と生成モデル	8
3. 複素 GGD に基づく NMF	14
3.1 複素 GGD	14
3.2 GGD-NMF	15
3.3 類似度関数と反復更新式	16
4. 音楽信号の分離実験	18
4.1 実験条件	18
4.2 実験結果	20
5. GGD-NMF の最適パラメータ推定法の検討	30
5.1 音源分離性能の予測	30
5.2 音源分離性能と相関を持つ類似度の提案	30
5.3 相関関係の実験的調査	31
6. まとめ	33
謝辞	34
参考文献	34

1. はじめに

音源分離とは、音響メディアを対象とした信号処理技術の一つである。音声・雑音・歌声・楽器音・機械音等の混合音を、混合前の個々の音源に分離する技術である。これは、人間の持つカクテルパーティ効果を機械で実現することに対応する。カクテルパーティ効果とは、図 1 に示すように、様々な音が混在する中で、特定の音に注意を向けることでその音だけを聞き取ることができる人間の能力のことである。音源分離は、音声認識（図 2）、音の検知・追跡、自動採譜（図 3）、音拡張現実感等、あらゆる応用技術の最初に必要な信号処理であり、性能向上には必要不可欠である。

本論文では、モノラル信号における音楽音源分離を対象とする。モノラル信号は最も基本的な音響信号であり、音色に関する情報しか得られない。すなわち、ステレオ等の多チャンネル信号におけるチャンネル間音量比やチャンネル間位相差等の情報が得られないため、音源分離においては最も困難な問題となる。一方で、本論文で示すモノラル音源分離技術の解析内容は、容易に多チャンネル音源分離技術にも応用可能である。

モノラル信号における音源分離技術では、非負値行列因子分解（nonnegative matrix factorization: NMF）[1, 2]と呼ばれる数値アルゴリズムを活用した手法が高精度であり、盛んに研究されている[3–7]。NMF を用いた音源分離は、混合されている音源の教師情報（各音源の音色サンプルや楽譜等、ユーザが事前に用意できる情報）を用いない「教師なし音源分離」[3–5]と、積極的に援用する「教師あり音源分離」[6, 7]に大別される（図 4）。前者は教師情報を必要としないためあらゆる音響信号に適用できるが、高精度な音源分離を実現することが非常に困難である。例えば、音楽信号の音源分離では音質が劣化してしまい、芸術性が保てなくなってしまう。本論文では教師あり音源分離を対象とする。さらにその中でも「音サンプルを用いるもの」と、「音以外（カメラや楽譜等）を用いるもの」があり、一般的に前者の方が高音質であるため、こちらを対象とする。音サンプルを用いる場合、必要最小限のデータを取り扱う「省サンプル教師あり」と、大量のデータを取り扱う「ビッグデータ教師あり」がある。後者は、近年注目を浴びている深層学習等を用いるものが考えられる[8, 9]が、例えば、各楽器音やボーカルの単独の録音データをユーザが大量に集めることは現実的ではない。以上の理由により、本論文では省サンプル教師あり音源分離を研究対象とする。

本論文では、現時点で調査が不十分である複素一般化ガウス分布（generalized Gaussian distribution: GGD）に基づく NMF（GGD-NMF）[10, 11]と既存の一般的な手法である Itakura-Saito NMF（IS-NMF）[12]を比較することで、GGD-NMF の有効性を実験的に確認し、GGD-NMF による適した音源の特徴を解明する。

本論文の構成は次の通りである。第 2 章では、NMF の理論について述べ、音源分離への適用方法を示す。また、既存の一般的な手法である IS-NMF の確率的な解釈について述べる。第 3 章では、まず複素 GGD について詳細を説明し、次にこれを NMF の生成モデルに。

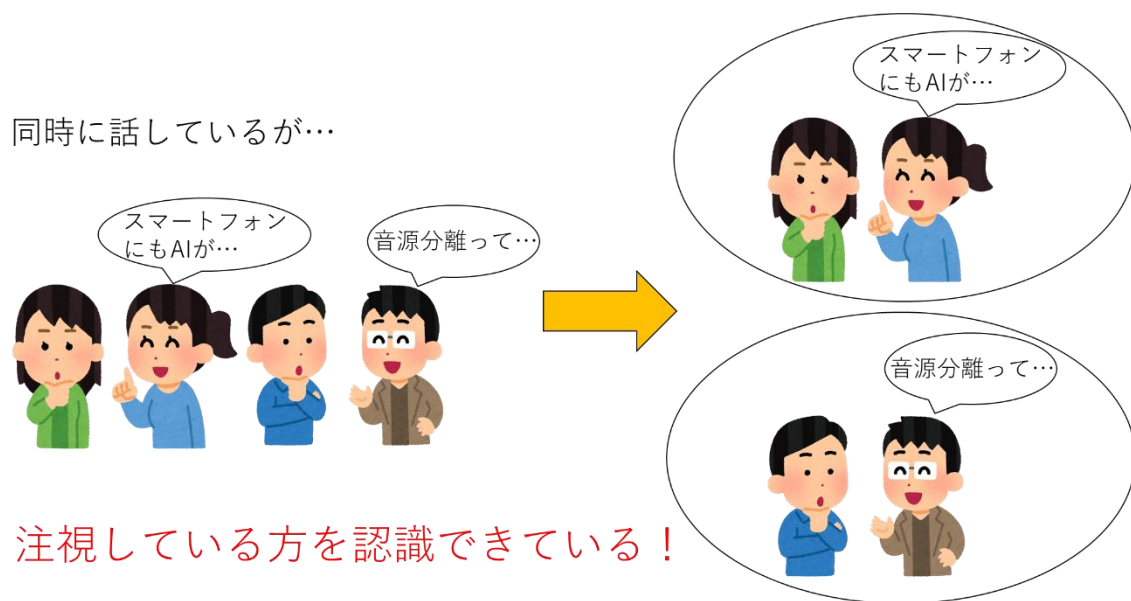


図 1 カクテルパーティ効果



図 2 AI スピーカー

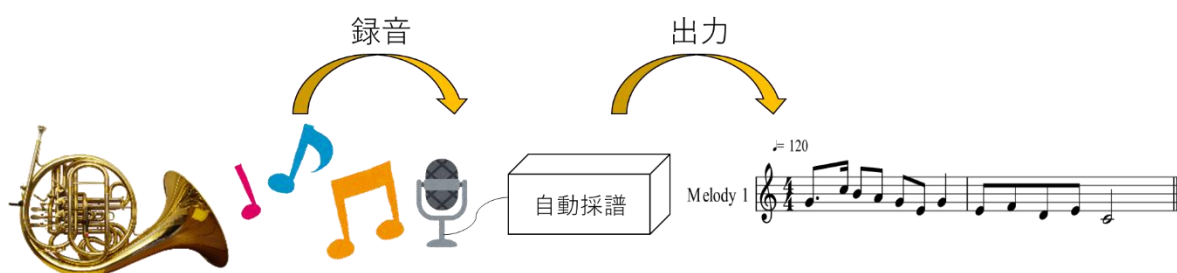


図 3 自動採譜

導入した GGD-NMF について述べる．さらに，GGD-NMF における類似度関数について解析する．第 4 章では，同条件下で GGD-NMF と既存手法である IS-NMF の音源分離精度を比較し，IS-NMF を上回るような複素 GGD の形状母数及び尺度母数を特定する．第 5 章では，教師情報の $\cos$ 距離から音源分離にとって最適な形状母数及び尺度母数を事前に特定できるか否かを検証し，その結果を述べる．第 6 章にて，本論文のまとめを述べる

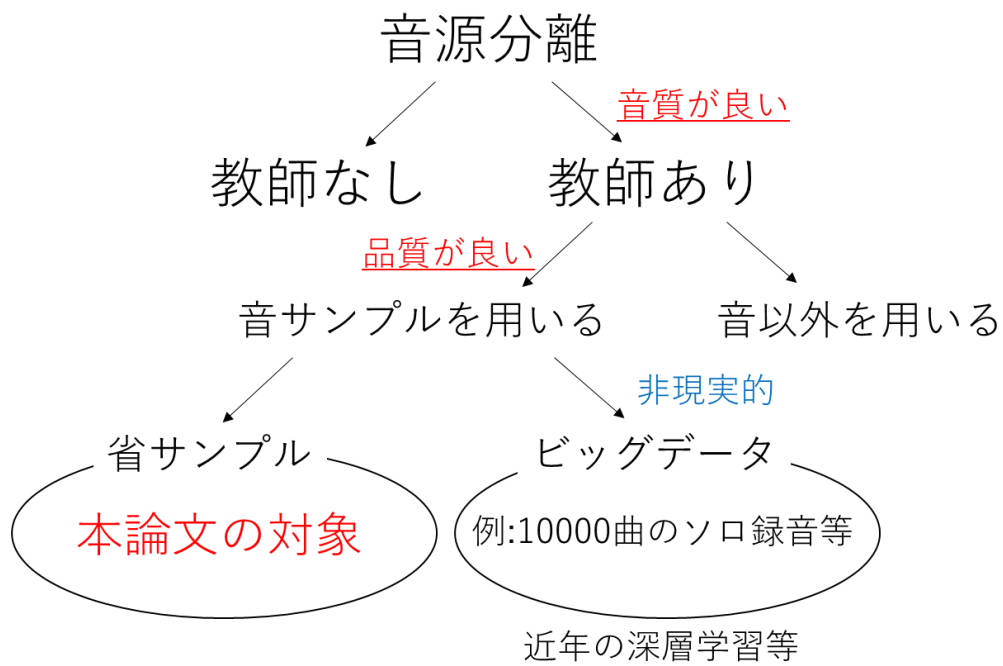


図 4 音源分離の分類

2. NMF を用いた音源分離

本章では、NMF の音源分離における基礎的な技術と既存の NMF のアルゴリズムについて述べる。また、NMF における類似度関数と生成モデルとしての解釈についても説明する。

2.1 短時間フーリエ変換

短時間フーリエ変換（short-time Fourier transform: STFT）とは、一次元（時間）の信号を二次元（時間と周波数）の信号に変換することであり、その処理内容は図 5 に示す通りである。観測した時間波形を短い時間波形に分割し、窓関数を乗じて離散フーリエ変換を適用する。この処理を時間に関してシフトさせていくことで、時間と周波数の二次元信号を得ることができる。これをスペクトログラムと呼び、複素数要素を持つ行列 $\mathbf{X}$ で表現する。音響信号処理においては、複素スペクトログラム $\mathbf{X}$ ではなく、複素数の絶対値の指数乗をとった「非負スペクトログラム $|\mathbf{X}|^p$ 」を対象とすることが多い。ここで、行列に対して絶対値記号とドット付きの指数を組み合わせた記号 ($|\cdot|^p$) を、行列の要素毎の絶対値の指数乗と定義する。従って、複素スペクトログラム $\mathbf{X}$ の振幅スペクトログラムは $|\mathbf{X}|^1$ 、パワースペクトログラムは $|\mathbf{X}|^2$ と表せる。ここで、要素毎に絶対値をとると、位相の情報が失われてしまうが、人間の聴覚はスペクトログラムの位相の変化をほとんど認識することができないことが一般に知られている。本論文においても、非負スペクトログラムのみを対象とし、位相については取り扱わない。

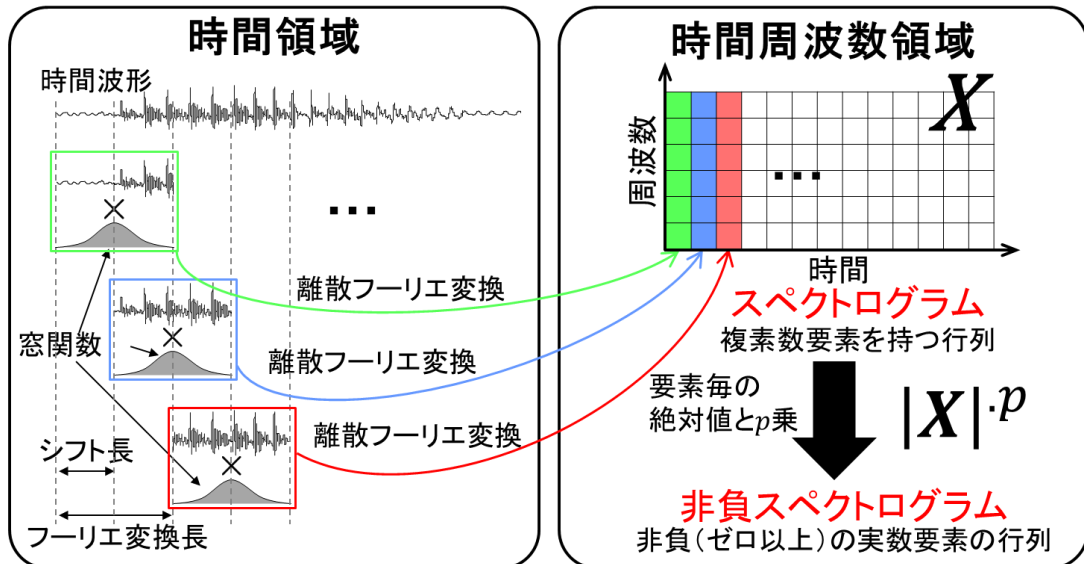


図 5 STFT の過程

2.2 NMF

NMF とは、非負制約付きの任意基底数による低ランク近似手法である。実世界には個数、順序といった非負値で表されるデータが多く、そのような非負値のデータを加法的な構成成分に分解することを目的として用いられている。本実験では、2つの異なる音響信号が混在した混合信号から各楽器の音響信号に分解するために用いている。

NMF は次式に示す行列の因子分解を行う。

$$\mathbf{D} \approx \hat{\mathbf{D}} = \mathbf{W}\mathbf{H} \quad (1)$$

ここで、 $\mathbf{D}$ は観測された $I \times J$ の非負行列（要素が全て零以上の行列）、 $\mathbf{W}$ は $I \times K$ の非負行列変数、 $\mathbf{H}$ は $K \times J$ の非負行列変数である。特に $\mathbf{W}$ と $\mathbf{H}$ は、それぞれ基底行列及びアクティベーション行列を表す。式(1)を要素毎に表すと、

$$d_{ij} \approx \hat{d}_{ij} = \sum_{k=1}^K w_{ik} h_{kj} \quad (2)$$

となる。ここで、 d_{ij} 、 $\hat{d}_{ij}$ 、 w_{ik} 、及び h_{kj} は、それぞれ行列 $\mathbf{D}$ 、 $\hat{\mathbf{D}}$ 、 $\mathbf{W}$ 、及び $\mathbf{H}$ の非負の要素（スカラー）である。また、添え字の i 、 j 、及び k はそれぞれ $\mathbf{D}$ の行、列、及び $\mathbf{W}$ の列番号である。 $\mathbf{W}$ の列ベクトル（以後、基底ベクトルと呼ぶ）の数 K は通常 I や J よりも小さな値に設定される。すなわち、NMF は観測された非負行列 $\mathbf{D}$ を、より低ランク（ランク K ）な行列 $\hat{\mathbf{D}}$ で近似するアルゴリズムであり、非負行列 $\mathbf{W}$ 及び $\mathbf{H}$ を推定することで、 K 個の基底ベクトルとそのアクティベーション（係数）を得ることができる。

NMF を音響信号に適用した例を図6に載せる。観測行列 $\mathbf{D}$ を、図5における非負スペクトログラム $|\mathbf{X}|^p$ としたとき、NMF を用いることで基底行列 $\mathbf{W}$ とアクティベーション行列 $\mathbf{H}$ の行列積で $\mathbf{D}$ を近似することができる。この時、基底行列 $\mathbf{W}$ は頻出するスペクトルパターン（音色のパーツ）を列ベクトルとして K 個含むような行列となり、図6では、 $K=2$ 本の基底ベクトルがスペクトルパターンとして得られている。さらに、アクティベーション行列 $\mathbf{H}$ は各基底ベクトルの時間的な強度変化（その音色の長さや音量）を行ベクトルとして K 個含むような行列となる。従って、図6では、 $K=2$ として、たった2本の基底ベクトルで観測行列 $\mathbf{D}$ を表そうとしている。このように、楽器音信号の非負スペクトログラムは、おおよそ音階の数の音色成分（スペクトルパターン）で表現できるため、 K は $\mathbf{D}$ に含まれる音色や音階の数である程度近似できる。このような特性は、数学的には「行列 $\mathbf{D}$ のランクが低い」と表現される。

このようにして得られた基底ベクトルやアクティベーションは、音源分離、自動採譜、雑音抑圧、残響抑圧等、様々な音アプリケーションに応用することができる。本論文では、次節で説明する教師あり NMF に基づく音源分離を取り扱う。

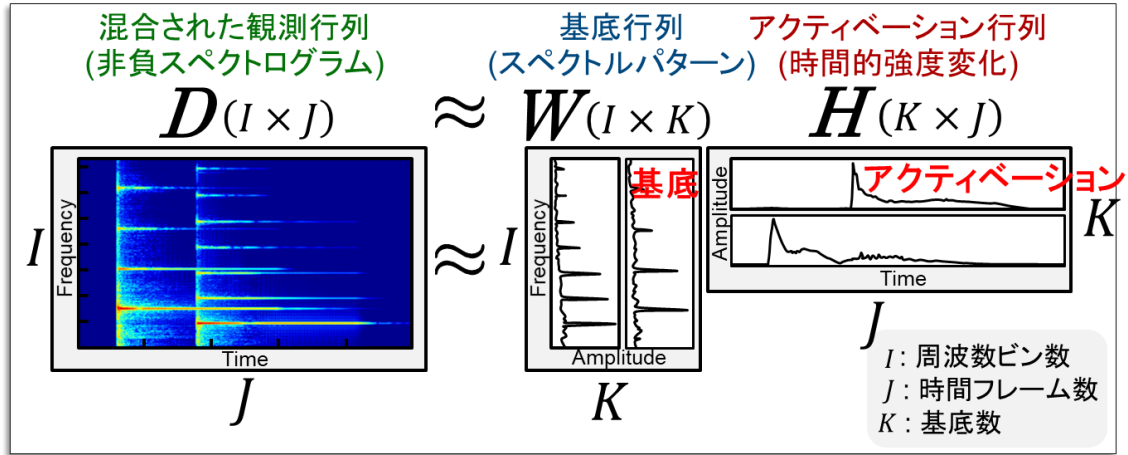


図 6 NMF の例 ($K = 2$)

2.3 教師あり NMF

NMF を用いた音源分離において、各音源のサンプル信号を教師情報として用いる教師あり NMF は比較的高性能な音源分離が達成できることから、よく用いられる。教師あり NMF は、「学習ステージ」と「分離ステージ」の二段階の処理に分けられる。説明の簡略化のために、2 つの音源の分離問題を考える。

教師あり NMF のアルゴリズムを図 7 に示す。今、各音源のサンプル信号が教師情報として与えられている。1 番目の音源のサンプル信号の複素スペクトログラムを $\mathbf{X}_1$ 、2 番目の音源のサンプル信号の複素スペクトログラムを $\mathbf{X}_2$ とし、さらにそれらの非負スペクトログラムを $\mathbf{D}_1 = |\mathbf{X}_1|^p$ 及び $\mathbf{D}_2 = |\mathbf{X}_2|^p$ とおく。はじめに学習ステージで $\mathbf{D}_1$ 及び $\mathbf{D}_2$ を NMF で分解する。

$$\mathbf{D}_1 \approx \mathbf{W}_1 \mathbf{H}_1 \quad (3)$$

$$\mathbf{D}_2 \approx \mathbf{W}_2 \mathbf{H}_2 \quad (4)$$

ここで、基底行列 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ のサイズはそれぞれ $I \times K_1$ 及び $I \times K_2$ であり、 $\mathbf{H}_1$ 及び $\mathbf{H}_2$ のサイズはそれぞれ $K_1 \times J$ 及び $K_2 \times J$ である。この結果得られる行列 $\mathbf{W}_1 \mathbf{H}_1$ 及び $\mathbf{W}_2 \mathbf{H}_2$ のうち、基底行列 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ は、図 6 で示したように音源 1 と音源 2 のスペクトルパターンを含んでいる。サンプル信号が図 7 のように音階を並べた信号であれば、各音階の基底ベクトルが得られるため、 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ はそれぞれ音源 1 と音源 2 の音色の辞書のような行列となる。

次に分離ステージでは、音源 1 と音源 2 が混合した信号の複素スペクトログラムを $\mathbf{M}$ とすると、その非負スペクトログラム $\mathbf{Y} = |\mathbf{M}|^p$ を NMF で次式のように分解する。

$$\mathbf{Y} \approx \mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2 \quad (5)$$

ここで、 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ は、いずれも学習ステージで得た基底行列であり、 $\mathbf{G}_1$ 及び $\mathbf{G}_2$ は新たに推定するアクティベーション行列である。従って、理想的には、 $\mathbf{W}_1 \mathbf{G}_1$ は音源 1 の非負スペクトログラム、 $\mathbf{W}_2 \mathbf{G}_2$ は音源 2 の非負スペクトログラムをそれぞれ表すため、2 つの音源を分離することができる。 $\mathbf{W}_1 \mathbf{G}_1$ 及び $\mathbf{W}_2 \mathbf{G}_2$ を逆 STFT すれば、音源 1 及び音源 2 の推定された

分離信号を得ることができる．ただし，NMF の分解では，複素スペクトログラムの位相は無視しているため，逆 STFT する前に，推定した $W_1 G_1$ 及び $W_2 G_2$ に何らかの形で位相を付与する必要がある．本論文では，次式の Wiener filter を適用することで，音源 1 の複素スペクトログラム A と音源 2 の複素スペクトログラム B をそれぞれ得る．

$$A = \frac{(W_1 G_1)^{\frac{2}{p}}}{(W_1 G_1)^{\frac{2}{p}} + (W_2 G_2)^{\frac{2}{p}}} \times M \quad (6)$$

$$B = \frac{(W_2 G_2)^{\frac{2}{p}}}{(W_1 G_1)^{\frac{2}{p}} + (W_2 G_2)^{\frac{2}{p}}} \times M \quad (7)$$

ここで，演算子 $\times$ 及び行列同士の分数はそれぞれ要素毎の乗算及び要素毎の除算を表し，行列に対するドット付きの指数乗 $(\cdot)^p$ は行列の要素毎の指数乗を表す．

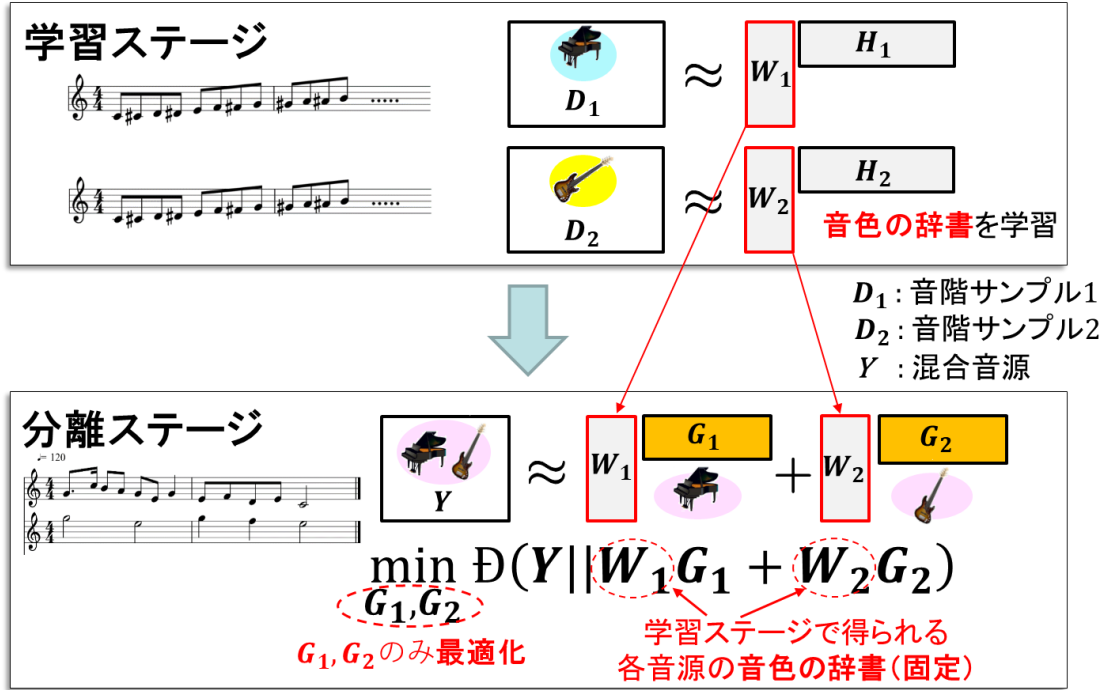


図 7 教師あり NMF による音源分離

2.4 類似度関数と生成モデル

NMF における W 及び H の推定は，観測行列 D とモデル WH 間の類似度を非負値制約の下で最小化する問題として定式化される．

$$\min_{W, H} D(D || WH) \quad \text{s.t. } w_{ik}, h_{kj} \geq 0 \quad \forall i, j, k \quad (8)$$

$D(\cdot || \cdot)$ は 2 つの入力変数間の類似度を測る関数である．特に，次式に示す二乗ユークリッド距離 (squared Euclidean distance: Eu 距離)，一般化 Kullback Leibler ダイバージェンス

(generalized Kullback Leibler divergence: KL ダイバージェンス), 及び板倉斎藤擬距離 (Itakura-Saito divergence: IS ダイバージェンス) の3つが音源分離等で頻繁に用いられている.

$$\begin{aligned} D_{\text{Eu}}(\mathbf{D}||\mathbf{WH}) &= \|\mathbf{D} - \mathbf{WH}\|_{\text{Fr}}^2 \\ &= \sum_{i=1}^I \sum_{j=1}^J \left(d_{ij} - \sum_{k=1}^K w_{ik} h_{kj} \right)^2 \end{aligned} \quad (9)$$

$$D_{\text{KL}}(\mathbf{D}||\mathbf{WH}) = \sum_{i=1}^I \sum_{j=1}^J \left(d_{ij} \log \frac{d_{ij}}{\sum_{k=1}^K w_{ik} h_{kj}} - d_{ij} + \sum_{k=1}^K w_{ik} h_{kj} \right) \quad (10)$$

$$D_{\text{IS}}(\mathbf{D}||\mathbf{WH}) = \sum_{i=1}^I \sum_{j=1}^J \left(\frac{d_{ij}}{\sum_{k=1}^K w_{ik} h_{kj}} - \log \frac{d_{ij}}{\sum_{k=1}^K w_{ik} h_{kj}} - 1 \right) \quad (11)$$

式(8)の $\mathbf{W}$ 及び $\mathbf{H}$ の解は閉形式では求まらない. その代わり, $\mathbf{W}$ と $\mathbf{H}$ に何らかの初期値を与えて繰り返し計算することで, 式(8)中の類似度関数の値を小さくするような $\mathbf{W}$ と $\mathbf{H}$ を求めることができる. 例えば, 式(9)の Eu 距離を類似度関数に用いた場合, 次式のような反復更新式が導出できる[2].

$$w_{ik} \leftarrow w_{ik} \cdot \frac{\sum_{j=1}^J h_{kj} d_{ij}}{\sum_{j=1}^J h_{kj} (\sum_{k'=1}^K w_{ik'} h_{k'j})} \quad (12)$$

$$h_{kj} \leftarrow h_{kj} \cdot \frac{\sum_{i=1}^I w_{ik} d_{ij}}{\sum_{i=1}^I w_{ik} (\sum_{k'=1}^K w_{ik'} h_{k'j})} \quad (13)$$

これらは, 行列形式で次式のように表すこともできる.

$$\mathbf{W} \leftarrow \mathbf{W} \times \frac{\mathbf{D}\mathbf{H}^T}{(\mathbf{W}\mathbf{H})\mathbf{H}^T} \quad (14)$$

$$\mathbf{H} \leftarrow \mathbf{H} \times \frac{\mathbf{W}^T \mathbf{D}}{\mathbf{W}^T (\mathbf{W}\mathbf{H})} \quad (15)$$

また, 式(10)の KL ダイバージェンスを類似度関数に用いた場合, 反復更新式は

$$w_{ik} \leftarrow w_{ik} \cdot \frac{\sum_{j=1}^J h_{kj} \frac{d_{ij}}{\sum_{k'=1}^K w_{ik'} h_{k'j}}}{\sum_{j=1}^J h_{kj}} \quad (16)$$

$$h_{kj} \leftarrow h_{kj} \cdot \frac{\sum_{i=1}^I w_{ik} \frac{d_{ij}}{\sum_{k'=1}^K w_{ik'} h_{k'j}}}{\sum_{i=1}^I w_{ik}} \quad (17)$$

となる[2]. 行列形式で表すと, 次のようになる.

$$\mathbf{W} \leftarrow \mathbf{W} \times \frac{\left[\frac{\mathbf{D}}{(\mathbf{W}\mathbf{H})} \right] \mathbf{H}^T}{\mathbf{E}\mathbf{H}^T} \quad (18)$$

$$\mathbf{H} \leftarrow \mathbf{H} \times \frac{\mathbf{W}^T \left[\frac{\mathbf{D}}{(\mathbf{W}\mathbf{H})} \right]}{\mathbf{W}^T \mathbf{E}} \quad (19)$$

ここで、 $\mathbf{E}$ は要素が全て1のサイズ $I \times J$ の行列である。さらに、式(11)の IS ダイバージェンスを類似度関数に用いた場合、反復更新式は

$$w_{ik} \leftarrow w_{ik} \cdot \sqrt{\frac{\sum_{j=1}^J h_{kj} \frac{d_{ij}}{(\sum_{k'=1}^K w_{ik'} h_{k'j})^2}}{\sum_{j=1}^J \frac{h_{kj}}{\sum_{k'=1}^K w_{ik'} h_{k'j}}}} \quad (20)$$

$$h_{kj} \leftarrow h_{kj} \cdot \sqrt{\frac{\sum_{i=1}^I w_{ik} \frac{d_{ij}}{(\sum_{k'=1}^K w_{ik'} h_{k'j})^2}}{\sum_{i=1}^I \frac{w_{ik}}{\sum_{k'=1}^K w_{ik'} h_{k'j}}}} \quad (21)$$

となる[13]。行列形式で表すと、次のようになる。

$$\mathbf{W} \leftarrow \mathbf{W} \times \left(\frac{\left[\frac{\mathbf{D}}{(\mathbf{W}\mathbf{H})^2} \right] \mathbf{H}^T}{\left[\frac{\mathbf{E}}{(\mathbf{W}\mathbf{H})} \right] \mathbf{H}^T} \right)^{\frac{1}{2}} \quad (22)$$

$$\mathbf{H} \leftarrow \mathbf{H} \times \left(\frac{\mathbf{W}^T \left[\frac{\mathbf{D}}{(\mathbf{W}\mathbf{H})^2} \right]}{\mathbf{W}^T \left[\frac{\mathbf{E}}{(\mathbf{W}\mathbf{H})} \right]} \right)^{\frac{1}{2}} \quad (23)$$

いずれの反復更新式も、現在の変数値に何らかの係数を乗じたものを新たな変数値とするような計算になっており、このような反復更新式は乗算型反復更新式と呼ばれる。

教師あり NMF の分離ステージでは、学習ステージで得られる基底行列 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ が与えられたもとで、次の最小化問題を解く必要がある。

$$\min_{\mathbf{G}_1, \mathbf{G}_2} D(\mathbf{Y} || \mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2) \quad (24)$$

すなわち、各音源の音色の辞書 ($\mathbf{W}_1$ 及び $\mathbf{W}_2$) を固定した状態で、各音源のアクティベーション ($\mathbf{G}_1$ 及び $\mathbf{G}_2$) を求める問題である。このときの反復更新式は、類似度関数が Eu 距離の場合

$$\mathbf{G}_1 \leftarrow \mathbf{G}_1 \times \frac{\mathbf{W}_1^T \mathbf{D}}{\mathbf{W}_1^T (\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)} \quad (25)$$

$$\mathbf{G}_2 \leftarrow \mathbf{G}_2 \times \frac{\mathbf{W}_2^T \mathbf{D}}{\mathbf{W}_2^T (\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)} \quad (26)$$

となる。また、KL ダイバージェンスの場合

$$\mathbf{G}_1 \leftarrow \mathbf{G}_1 \cdot \frac{\mathbf{W}_1^T \left[\frac{\mathbf{D}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)} \right]}{\mathbf{W}_1^T \mathbf{E}} \quad (27)$$

$$\mathbf{G}_2 \leftarrow \mathbf{G}_2 \cdot \frac{\mathbf{W}_2^T \left[\frac{\mathbf{D}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)} \right]}{\mathbf{W}_2^T \mathbf{E}} \quad (28)$$

となる．さらに，IS ダイバージェンスの場合

$$\mathbf{G}_1 \leftarrow \mathbf{G}_1 \cdot \frac{\mathbf{W}_1^T \left[\frac{\mathbf{D}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)^2} \right]}{\mathbf{W}_1^T \left[\frac{\mathbf{E}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)} \right]} \quad (29)$$

$$\mathbf{G}_2 \leftarrow \mathbf{G}_2 \cdot \frac{\mathbf{W}_2^T \left[\frac{\mathbf{D}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)^2} \right]}{\mathbf{W}_2^T \left[\frac{\mathbf{E}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)} \right]} \quad (30)$$

となる．

NMF における類似度の選択は，観測行列の生成モデルを選択することと等価である．例えば，Eu 距離に基づく NMF（以後，Eu-NMF と呼ぶ）は，図 8 に示すように，非負観測行列 $\mathbf{D}$ の生成モデルをガウス分布と仮定した際の平均の最尤推定と等価であり，KL ダイバージェンスに基づく NMF（以後，KL-NMF と呼ぶ）は $\mathbf{D}$ に対するポアソン分布，IS-NMF は $\mathbf{D}$ に対する指数分布を仮定した母数の最尤推定とそれぞれ等価である．また，IS-NMF については，以下で説明するように，非負観測行列 $\mathbf{D}$ を得る前の複素観測行列 $\mathbf{X}$ の生成モデルも与えることが知られている [12]．

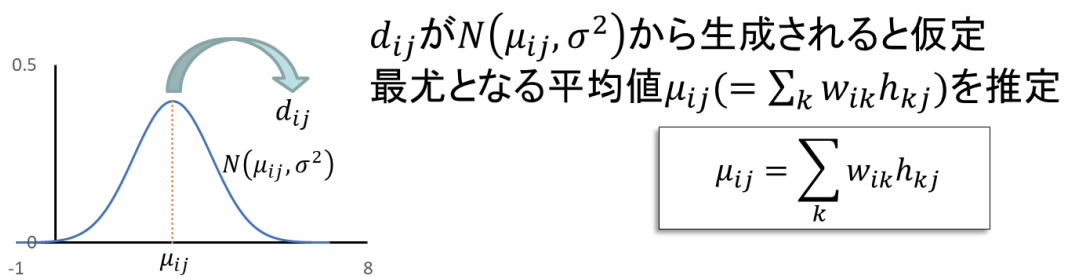


図 8 ガウス分布と仮定した際の平均の最尤推定

今，複素スペクトログラム $\mathbf{X}$ 中の各要素 x_{ij} が次のように K 個の複素成分に分解できると仮定する．

$$x_{ij} = x_{ij,1} + x_{ij,2} + \cdots + x_{ij,k} + \cdots + x_{ij,K} \quad (31)$$

さらに，これらの複素成分は，次式のような分布から生成されていると仮定する．

$$x_{ij,k} \sim \text{Nc}(0, \sigma_{ij,k}^2)$$

$$= \frac{1}{\pi \sigma_{ij,k}^2} \exp\left(-\frac{|x_{ij,k}|^2}{\sigma_{ij,k}^2}\right) \quad (32)$$

$$\equiv p(x_{ij,k})$$

ここで, $Nc(0, \sigma^2)$ は, 平均 0, 分散 σ^2 のパラメータを持つ原点对称な複素ガウス分布である. $Nc(0,1)$ の例を図 9 に示す. この時, 複素ガウス分布の再生性より複素スペクトル x_{ij} も次式の複素ガウス分布に従う.

$$x_{ij} \sim Nc\left(0, \sum_{k=1}^K \sigma_{ij,k}^2\right) \quad (33)$$

$$\equiv p(x_{ij})$$

すなわち, 複素成分 $x_{ij,k}$ を足したものである複素スペクトル x_{ij} の従う分布 (生成モデル, 式 (33)) は, $x_{ij,k}$ の生成モデル (式 (32)) の分散 $\sigma_{ij,k}^2$ を足した値を分散とする複素ガウス分布になる. 従って, 複素数の和を分散の和 (非負実数値の和) で表現していることになる. 一般に, 音の混合現象は時間波形の和であり, これは複素スペクトログラムの和と等価である. NMF による音の分解は図 6 に示すように, K 個の非負のパーツ (スペクトルパターンとアクティベーション) の和で表現されるため, 物理的には間違ったモデル化である. しかしながら, 式 (31) 及び式 (32) を仮定することで, 複素数の和 (音の混合現象) を非負のパーツ (基底ベクトルとアクティベーション) の和で表現することが可能となり, 音響信号に対して NMF を適用することの妥当性が統計的に保証される.

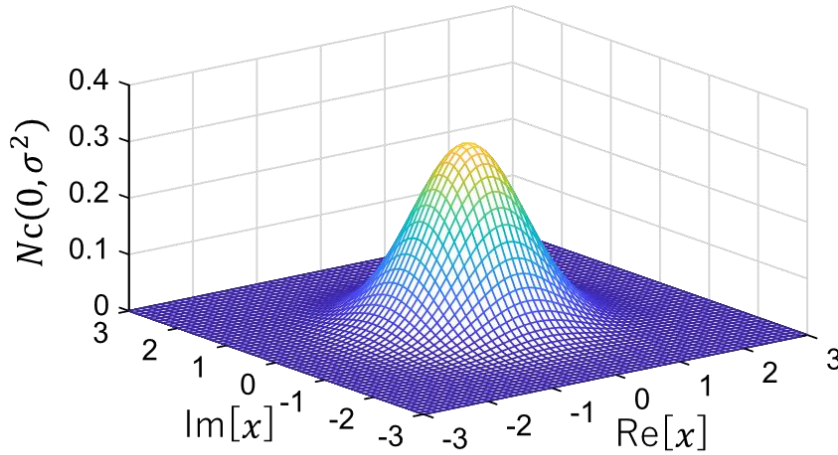


図 9 複素ガウス分布

式 (33) の分散 $\sigma_{ij,k}^2$ の最尤推定問題を考える. 式 (32) の生成モデルが i, j , 及び k に関して独立であると仮定すると, 複素スペクトログラム $\mathbf{X}$ の生成モデルは次式となる.

$$p(\mathbf{X}) = p(x_{11}, x_{12}, \dots, x_{1J}, x_{21}, \dots, x_{IJ})$$

$$\begin{aligned}
&= \prod_{i=1}^I \prod_{j=1}^J p(x_{ij}) \\
&= \prod_{i=1}^I \prod_{j=1}^J \text{Nc} \left(0, \sum_{k=1}^K \sigma_{ij,k}^2 \right)
\end{aligned} \tag{34}$$

さらに、 $p(\mathbf{X})$ の尤度関数 $L(\mathbf{X})$ は次式となる。

$$\begin{aligned}
L(\mathbf{X}) &= \prod_{i=1}^I \prod_{j=1}^J \text{Nc} \left(0, \sum_{k=1}^K \sigma_{ij,k}^2 \right) \\
&= \prod_{i=1}^I \prod_{j=1}^J \frac{1}{\pi \sum_{k=1}^K \sigma_{ij,k}^2} \exp \left(-\frac{|x_{ij}|^2}{\sum_{k=1}^K \sigma_{ij,k}^2} \right)
\end{aligned} \tag{35}$$

また、負対数尤度関数は

$$-\log L(\mathbf{X}) = \sum_{i=1}^I \sum_{j=1}^J \left[\log \pi + \log \sum_{k=1}^K \sigma_{ij,k}^2 + \frac{|x_{ij}|^2}{\sum_{k=1}^K \sigma_{ij,k}^2} \right] \tag{36}$$

となる。従って、分散 $\sigma_{ij,k}^2$ の最尤値を得るには、式(36)を最小化する $\sigma_{ij,k}^2$ を求めれば良い。

ここで、IS ダイバージェンスの最小化問題である式(8)及び式(11)を見直すと、 $d_{ij} = |x_{ij}|^2$ かつ $\sigma_{ij,k}^2 = w_{ik} h_{kj}$ の時、式(36)の $\sigma_{ij,k}^2$ に関する最小化問題は等価となる（式(11)と式(36)は定数項の違いを除いて一致する）。

以上より、IS-NMF をパワースペクトログラム $\mathbf{D} = |\mathbf{X}|^2$ に対して適用することは、式(31)及び式(32)の生成モデルを仮定したときの分散の最尤推定に等しいため、音響信号に NMF を適用することの正当性が統計的に保証される。

3. 複素 GGD に基づく NMF

近年提案された GGD-NMF は, IS-NMF で仮定されていた複素ガウス分布を複素 GGD へと拡張した新しい NMF であり, 音源分離へ適用した際の性能については詳細な調査がなされていない. 本章では, GGD-NMF について, その特徴やアルゴリズムについて述べる.

3.1 複素 GGD

複素 GGD とは, 原点对称な複素ガウス分布を拡張したものであり, 確率密度関数は次式で定義される.

$$\begin{aligned} x &\sim \text{GGDc}(0, \sigma, \rho) \\ &= \frac{\rho^{1-\frac{2}{\rho}}}{2^{1-\frac{2}{\rho}} \pi \sigma^2 \Gamma\left(\frac{2}{\rho}\right)} \exp \left[-\frac{2}{\rho} \left(\frac{|x|}{\sigma} \right)^\rho \right] \end{aligned} \quad (37)$$

ここで, $\rho > 0$ 及び $\sigma > 0$ はそれぞれ分布の形状母数及び尺度母数であり, $\Gamma(\cdot)$ はガンマ関数を表す. 図 10 は, 形状母数 ρ を変化させたときの複素 GGD の分布である. 形状母数 $\rho = 2$ のとき, 式(37)は IS-NMF で仮定されていた原点对称な複素ガウス分布に一致する. 複素 GGD は $0 < \rho < 2$ で優ガウスな分布となり, ρ の値が小さいほどヘビーテイルな分布になる. また, $\rho > 2$ で劣ガウスな分布となり, ρ の値が大きいくほど一様分布に近づく.

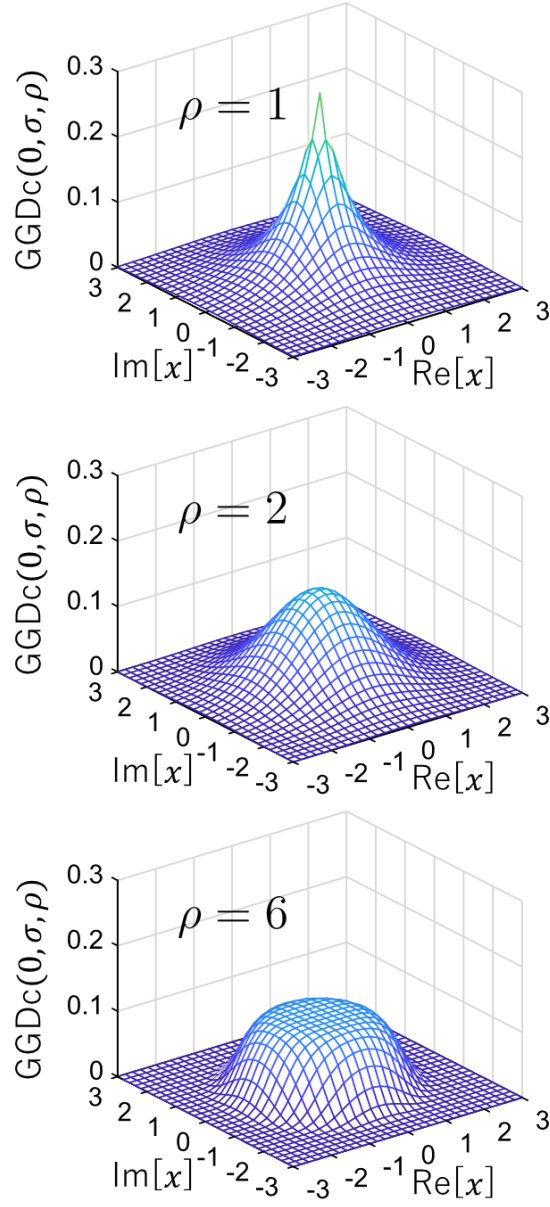


図 10 複素 GGD の概形

3.2 GGD-NMF

GGD-NMF では、式(37)を複素スペクトル x_{ij} の生成モデルと仮定する.

$$\begin{aligned}
 x_{ij} &\sim \text{GGDc}(0, \sigma_{ij}, \rho) \\
 &= \frac{\rho^{1-\frac{2}{\rho}}}{2^{1-\frac{2}{\rho}} \pi \sigma_{ij}^2 \Gamma(\frac{2}{\rho})} \exp \left[-\frac{2}{\rho} \left(\frac{|x_{ij}|}{\sigma_{ij}} \right)^\rho \right] \\
 \sigma_{ij}^p &= \sum_{k=1}^K w_{ik} h_{kj}
 \end{aligned} \tag{38}$$

ここで、 p はNMFによる低ランク近似の領域に対応するドメインパラメータ ($\mathbf{D} = |\mathbf{X}|^p$ における p)である。IS-NMFの場合と同様に、各時間周波数スロット (i 及び j)に独立な複素GGDを生成モデルとして仮定する。尺度母数 σ_{ij} は時間周波数毎に変動し、形状母数 ρ は全時間周波数で同じ値とする。さらに、NMFで近似表現する非負値のドメインを可変にするために、ドメインパラメータ p を導入する。 $p = 1$ の時は振幅スペクトログラム $|\mathbf{X}|^1$ に対して、また $p = 2$ の時はパワースペクトログラム $|\mathbf{X}|^2$ に対してそれぞれNMFを適用することになる。

3.3 類似度関数と反復更新式

以上のような生成モデルを仮定した場合の σ_{ij}^p の最尤推定問題は、次式のような類似度関数を最小化する w_{ik} 及び h_{kj} を推定することと等価であることが示されている[10]。

$$D_{\text{GGD}}(\mathbf{D}||\mathbf{WH}) = \sum_{i=1}^I \sum_{j=1}^J \left[\frac{2}{\rho} \left(\left[\frac{d_{ij}^{\frac{1}{p}}}{(\sum_{k=1}^K w_{ik} h_{kj})^{\frac{1}{p}}} \right]^{\rho} - \log \left[\frac{d_{ij}^{\frac{1}{p}}}{(\sum_{k=1}^K w_{ik} h_{kj})^{\frac{1}{p}}} \right] - 1 \right) \right] \quad (39)$$

式(39)は、式(11)のISダイバージェンスを形状母数 ρ に関して一般化した類似度関数とみなすことができ、 $\rho = p = 2$ のとき、式(11)と一致する。

式(39)を最小化する w_{ik} と h_{kj} は次の反復更新式で求めることができる[10, 11]。

$$w_{ik} \leftarrow w_{ik} \frac{\left[\sum_{j=1}^J \frac{z_{ij}}{(\sum_{k'=1}^K w_{ik'} h_{k'j})^2} h_{kj} \right]^{\frac{p}{\rho+p}}}{\sum_{j=1}^J \frac{1}{\sum_{k'=1}^K w_{ik'} h_{k'j}} h_{kj}} \quad (40)$$

$$h_{kj} \leftarrow h_{kj} \frac{\left[\sum_{i=1}^I \frac{z_{ij}}{(\sum_{k'=1}^K w_{ik'} h_{k'j})^2} w_{ik} \right]^{\frac{p}{\rho+p}}}{\sum_{i=1}^I \frac{1}{\sum_{k'=1}^K w_{ik'} h_{k'j}} w_{ik}} \quad (41)$$

$$z_{ij} = d_{ij}^{\frac{\rho}{p}} \left(\sum_{k=1}^K w_{ik} h_{kj} \right)^{1-\frac{\rho}{p}} \quad (42)$$

これらを行列形式に表すと、次のようになる。

$$\mathbf{W} \leftarrow \mathbf{W} \times \left(\frac{\left[\frac{\mathbf{Z}}{(\mathbf{WH})^2} \right] \mathbf{H}^T}{\left[\frac{\mathbf{E}}{(\mathbf{WH})} \right] \mathbf{H}^T} \right)^{\frac{p}{\rho+p}} \quad (43)$$

$$\mathbf{H} \leftarrow \mathbf{H} \times \left(\frac{\mathbf{W}^T \left[\frac{\mathbf{Z}}{(\mathbf{WH})^2} \right]}{\mathbf{W}^T \left[\frac{\mathbf{E}}{(\mathbf{WH})} \right]} \right)^{\frac{p}{\rho+p}} \quad (44)$$

$$\mathbf{Z} = \mathbf{D}^{\frac{\rho}{p}} \times (\mathbf{W}\mathbf{H})^{1-\frac{\rho}{p}} \quad (45)$$

また, 分離ステージでは, 学習ステージで得られる基底行列 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ が与えられたもとで, 以下の反復更新式を計算することでアクティベーション行列 $\mathbf{G}_1$ 及び $\mathbf{G}_2$ を得ることができる.

$$\mathbf{G}_1 \leftarrow \mathbf{G}_1 \times \left(\frac{\mathbf{W}_1^T \left[\frac{\mathbf{Z}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)^2} \right]}{\mathbf{W}_1^T \left[\frac{\mathbf{E}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)} \right]} \right)^{\frac{p}{\rho+p}} \quad (46)$$

$$\mathbf{G}_2 \leftarrow \mathbf{G}_2 \times \left(\frac{\mathbf{W}_2^T \left[\frac{\mathbf{Z}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)^2} \right]}{\mathbf{W}_2^T \left[\frac{\mathbf{E}}{(\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)} \right]} \right)^{\frac{p}{\rho+p}} \quad (47)$$

$$\mathbf{Z} = \mathbf{Y}^{\frac{\rho}{p}} \times (\mathbf{W}_1 \mathbf{G}_1 + \mathbf{W}_2 \mathbf{G}_2)^{1-\frac{\rho}{p}} \quad (48)$$

4. 音楽信号の分離実験

本章では、3章で述べた GGD-NMF を教師あり音源分離に適用した際の分離性能について、実験を通して調査する。実験では、音楽信号の音源分離を対象とする。

4.1 実験条件

本実験では、図 11 に示す音楽信号を音源分離の対象とした。ここで、各楽器名の略称は表 1 の通りである。この音楽信号は 4 つのパートからなり、各パートは 2 種類の楽器のうちいずれかで演奏される。合計 8 つの楽器音信号から 2 つを選んで混合したものを混合信号とし、その複素スペクトログラムを $\mathbf{M}$ とした。ただし、同一パートの 2 楽器を選ぶことはない（複数楽器のユニゾン演奏の混合信号は対象外とする）。また、各楽器音のサンプル信号（教師情報）は図 12 に示す信号とした。これらのサンプル信号は、いずれのパートも、図 11 の旋律の音域を含む 24 音（2 オクターヴ）の上昇音階信号である。これらのサンプル信号の複素スペクトログラムを $\mathbf{X}_1$ 及び $\mathbf{X}_2$ とし、式(3)及び式(4)の学習ステージを通して基底行列 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ を得た。次に、混合信号の複素スペクトログラム $\mathbf{M}$ と $\mathbf{W}_1$ 及び $\mathbf{W}_2$ を用いて式(5)の分離ステージを適用し、アクティベーション行列 $\mathbf{G}_1$ 及び $\mathbf{G}_2$ を推定した。最後に、式(6)及び式(7)の Wiener filter を適用し、逆 STFT を施すことで、両音源の推定分離信号を得た。音源分離性能の指標には、信号対歪み比（source-to-distortion ratio: SDR）[14]を用いた。これは、音源分離度合いと分離音の品質（歪みの少なさ）の両方を加味した総合的な客観指標である。今、複数の音源が混合しているとき、全楽器の混合音を $X(t)$ 、推定したい音源（目的音）を $S(t)$ 、推定したい音源以外の音源（妨害音）を $I(t)$ とすると、次式で表される。

$$X(t) = S(t) + I(t) \quad (49)$$

ここで、 t は時間サンプルのインデックスを表す。従って、 $X(t)$ 、 $S(t)$ 、及び $I(t)$ は離散時間信号である。この混合音源 $X(t)$ に対して音源分離を適用して得られる目的音の推定信号を $\hat{S}(t)$ とすると

$$\hat{S}(t) = e_s(t) + e_i(t) + e_A(t) \quad (50)$$

と仮定できる。ここで、 $e_s(t)$ 、 $e_i(t)$ 、及び $e_A(t)$ はそれぞれ目的音成分、残留した妨害音成分、及び音源分離によって生じた人工的な歪み成分（ $S(t)$ にも $I(t)$ にも属さない成分）を表す。SDR とは、 $e_s(t)$ 、 $e_i(t)$ 、及び $e_A(t)$ で求められ、次式で表される。

$$\text{SDR} = 10 \log_{10} \frac{\sum_{t=1}^T |e_s(t)|^2}{\sum_{t=1}^T |e_i(t) + e_A(t)|^2} \quad (51)$$

高い SDR 値を達成するためには、妨害音の残留成分 $e_i(t)$ と人工的な歪み成分 $e_A(t)$ が少なく、目的音成分 $e_s(t)$ が高精度に推定されている必要がある。その他の実験条件は表 2 の通りである。



図 11 実験に用いた音楽信号の楽譜

表 1 楽器名の略称

略称	楽器名
Hr.	ホルン
Tp.	トランペット
Cl.	クラリネット
Vn.	ヴァイオリン
Hp.	ハープシコード
Pf.	ピアノ
Tb.	トロンボーン
Vc.	チェロ



図 12 実験に用いたサンプル信号の楽譜

表 2 その他の実験条件

基底ベクトル数 K	30 [本]
反復更新式の計算回数	200 [回]
サンプリング周波数	44100 [Hz]
STFT の条件	92.9 [ms] のハミング窓を窓長の 1/2 でシフト
各変数行列の初期値	0~1 の一様乱数値
試行回数	5 [回]

4.2 実験結果

図 13-16 に、Cl.と Vc.の混合信号、Tp.と Vn.の混合信号、Hr.と Tb.の混合信号、Cl.と Tb.の混合信号の結果について、ドメインパラメータが $p = 0.4$, $p = 1.0$, 及び $p = 2.0$ のときの SDR 値を各形状母数 ρ についてそれぞれまとめている。ここで、 $p = \rho = 2$ のときが、既存手法である IS-NMF の結果であり、それ以外の値が GGD-NMF の結果を示している。また、各 SDR 値は各変数行列の乱数初期値を変えて、5 回試行した結果の平均値を示している。

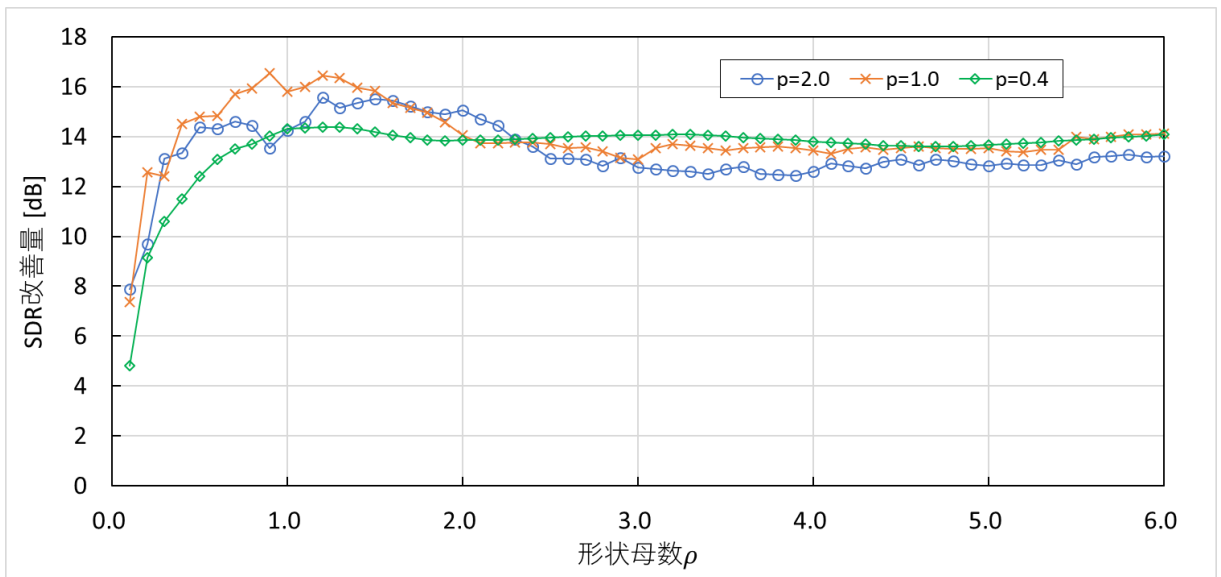


図 13 Cl.と Vc.の混合信号に対する音源分離性能

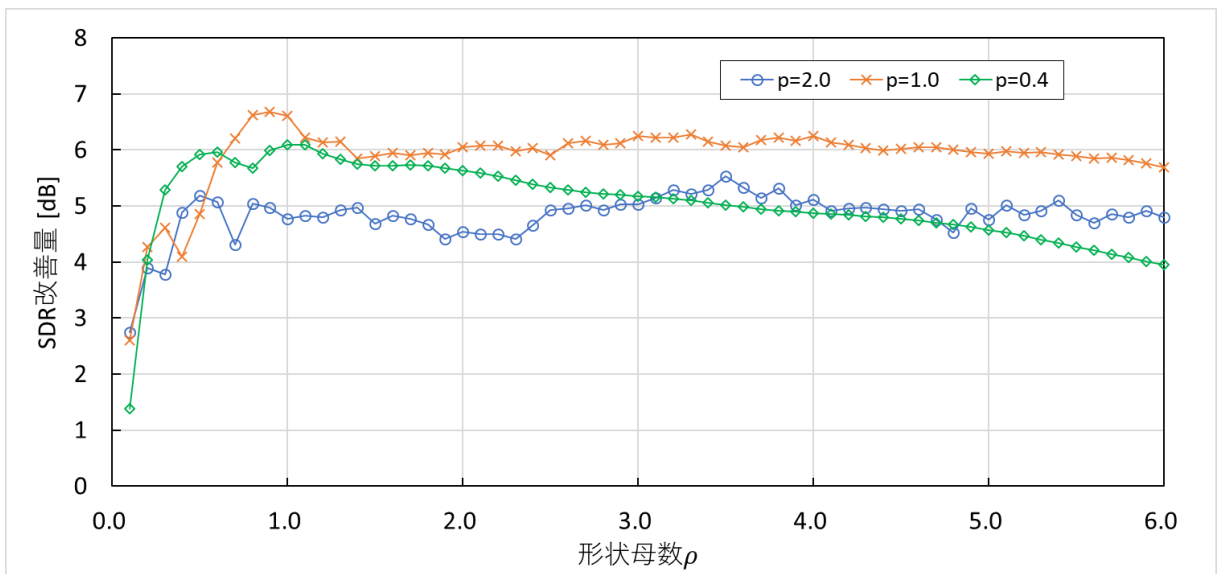


図 14 Tp.と Vn.の混合信号に対する音源分離性能

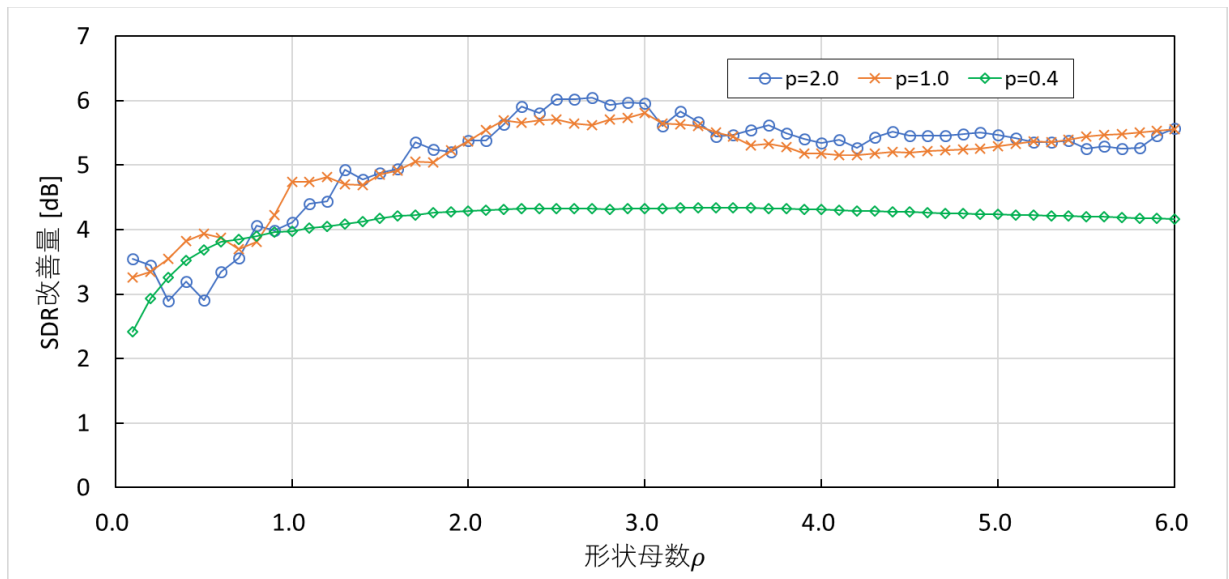


図 15 Hr.と Tb.の混合信号に対する音源分離性能

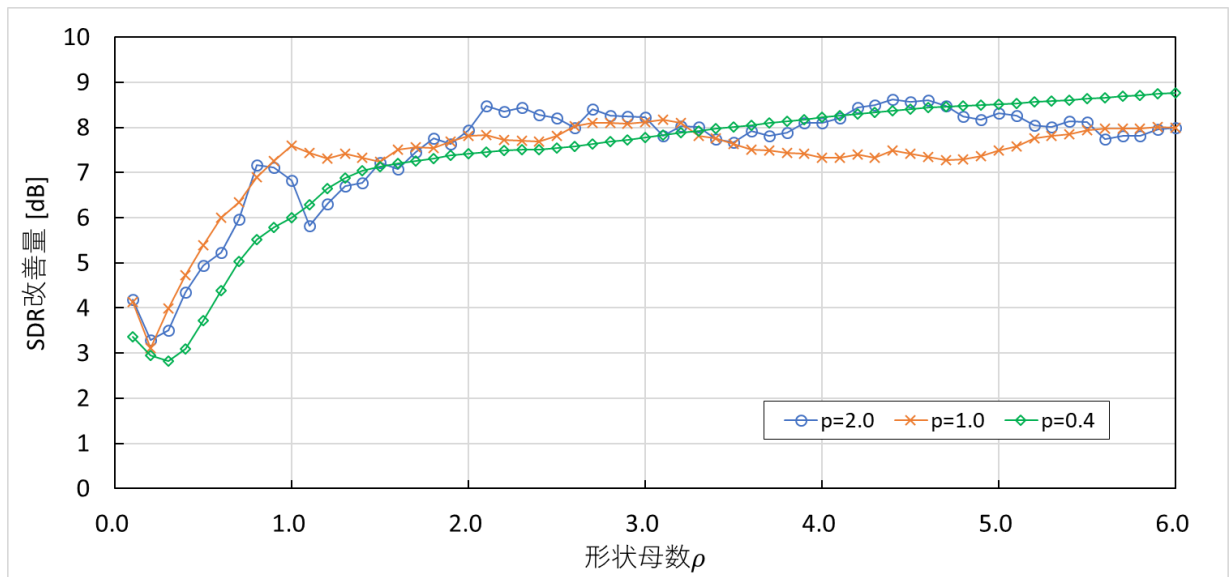


図 16 Cl.と Tb.の混合信号に対する音源分離性能

図 13 及び図 14 を見ると、形状母数 $\rho < 2$ のときに分離性能が向上している例がある．すなわち、生成モデルで優ガウスな分布を仮定することで、高精度な分離が達成できることがわかる．一方で、図 15 及び図 16 では、 $\rho > 2$ のときに分離性能が向上する傾向にあり、劣ガウスな生成モデルがより適切であることがわかる．その他の実験結果については、図 17-40 にヒートマップとして示す．各楽器の組み合わせ毎にばらつきはあるが、全体を通して、 $p = \rho = 2$ よりもヒートマップが明るい (SDR 値が高い) パラメータがあるのが見て取れる．

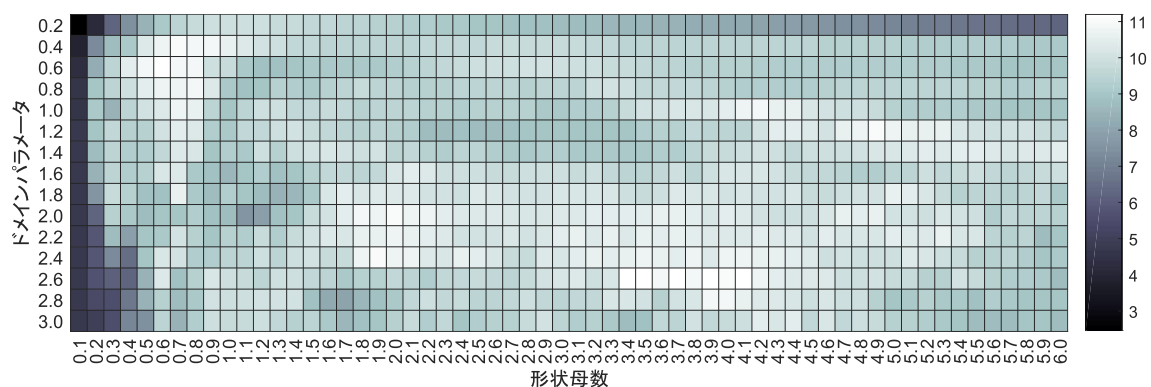


図 17 Cl.と Hr.の混合信号の実験結果 (SDR 値のヒートマップ)

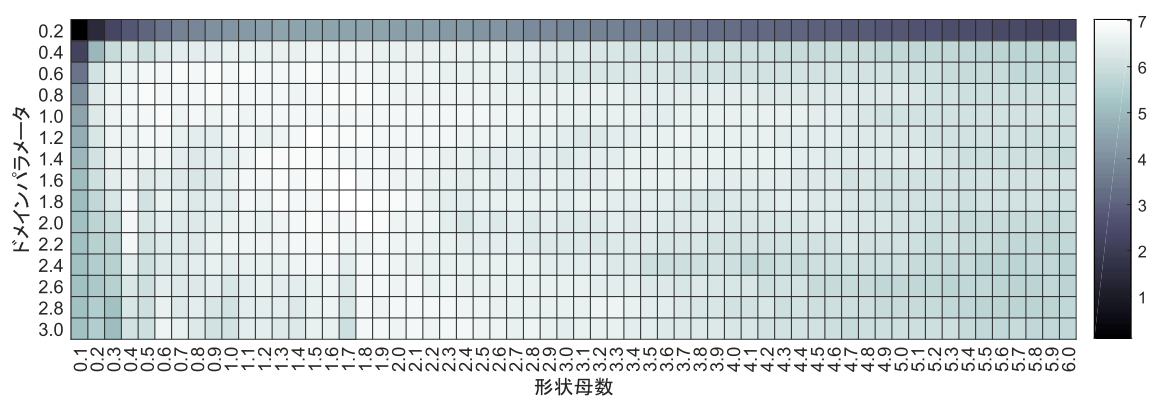


図 18 Hr.と Hp.の混合信号の実験結果 (SDR 値のヒートマップ)

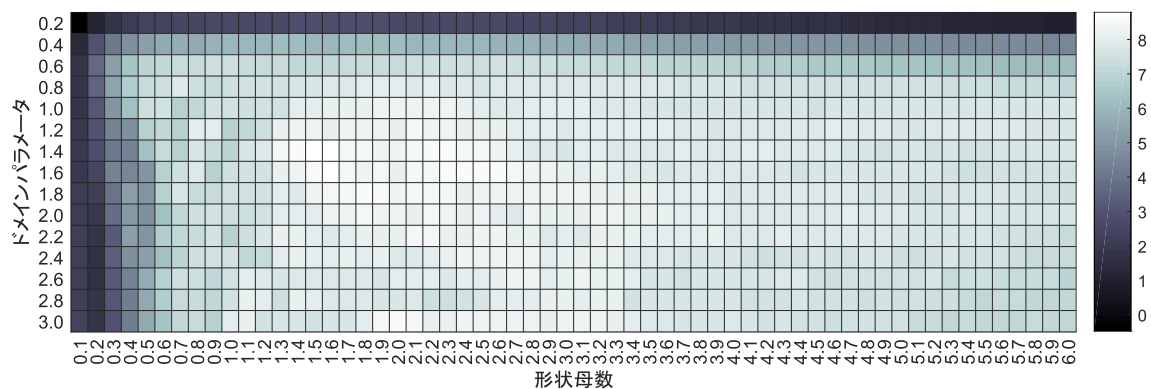


図 19 Hr.と Pf.の混合信号の実験結果 (SDR 値のヒートマップ)

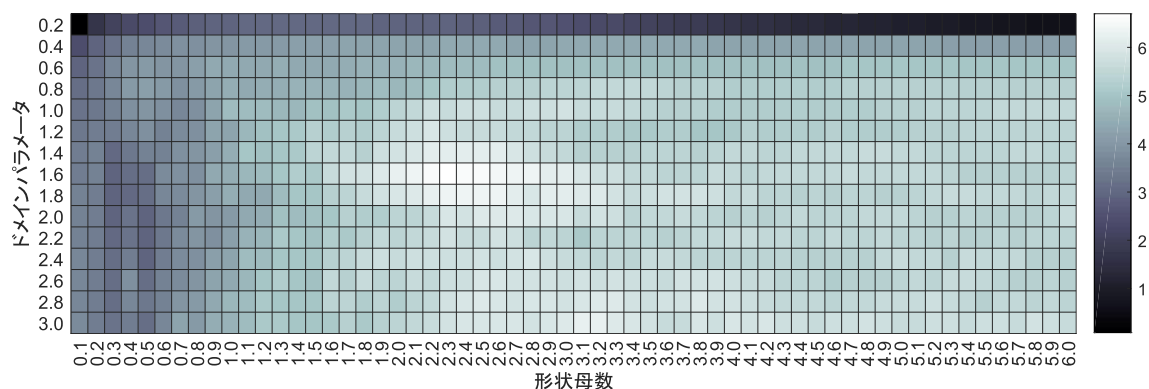


図 20 Hr.と Tb.の混合信号の実験結果 (SDR 値のヒートマップ)

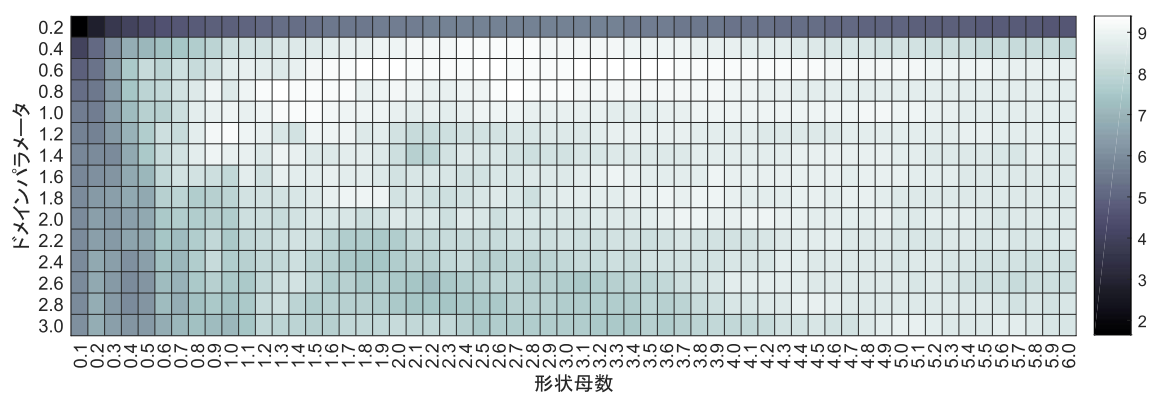


図 21 Hr.と Vc.の混合信号の実験結果 (SDR 値のヒートマップ)

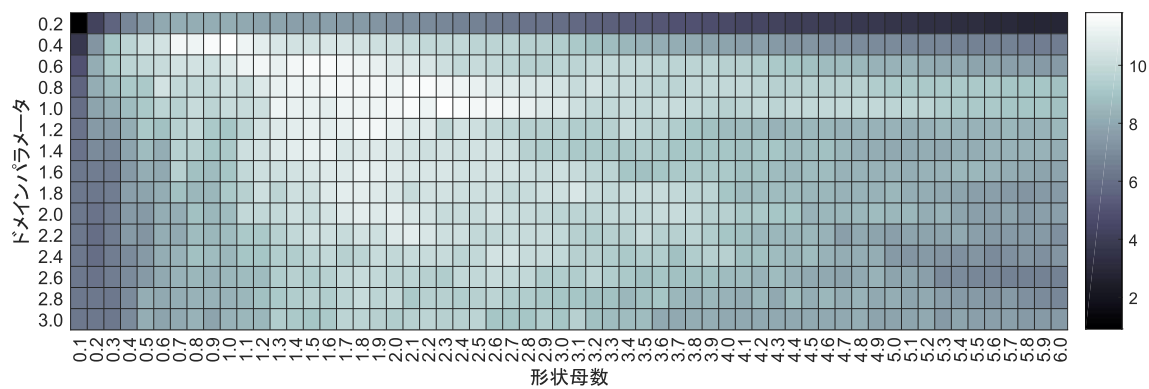


図 22 Hr.と Vn.の混合信号の実験結果 (SDR 値のヒートマップ)

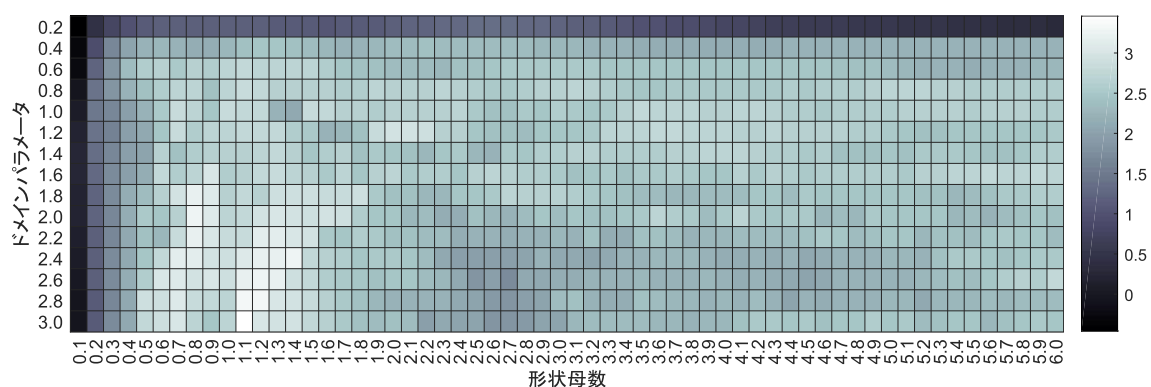


図 23 Tp.と Cl.の混合信号の実験結果 (SDR 値のヒートマップ)

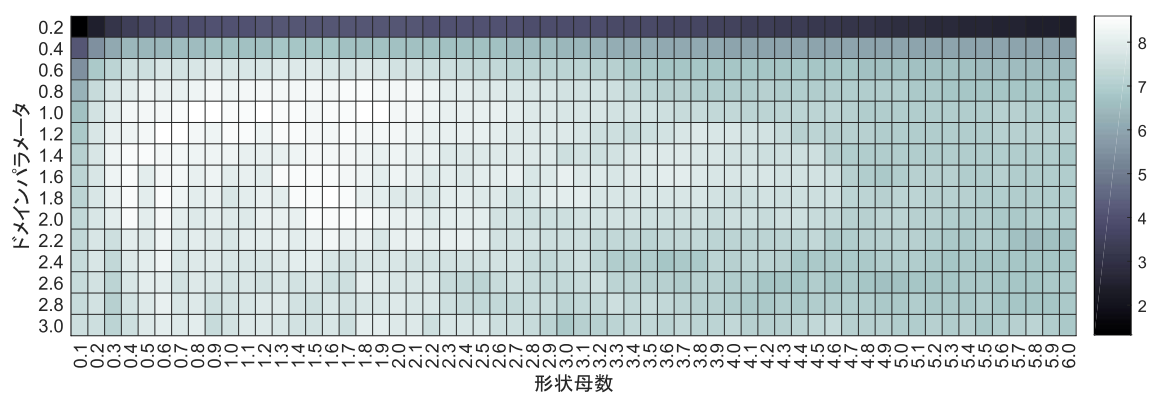


図 24 Tp.と Hp.の混合信号の実験結果 (SDR 値のヒートマップ)

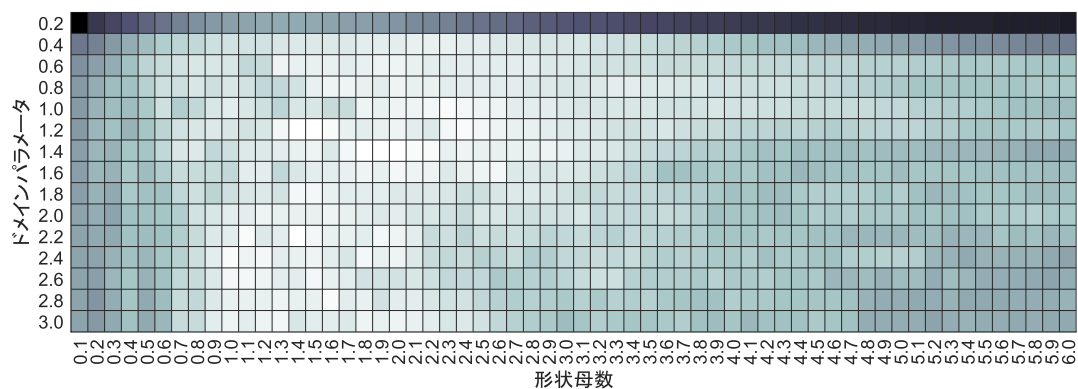


図 25 Tp.と Pf.の混合信号の実験結果 (SDR 値のヒートマップ)

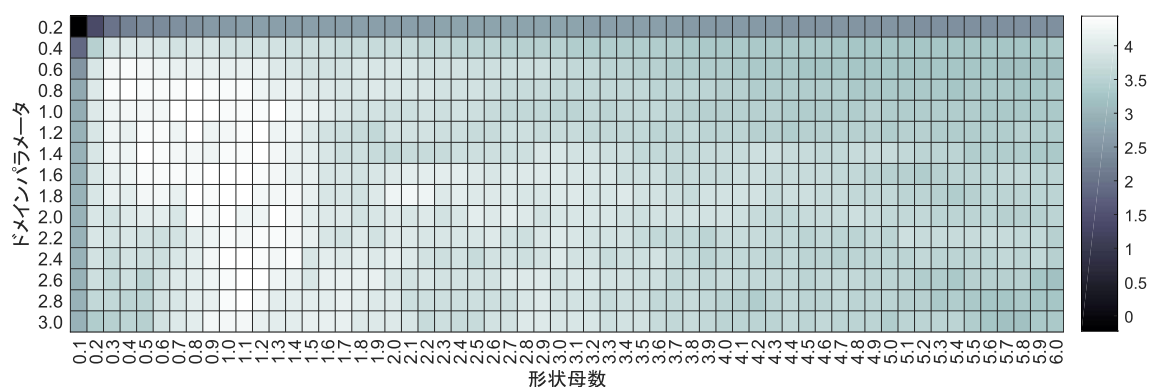


図 26 Tp.と Tb.の混合信号の実験結果 (SDR 値のヒートマップ)

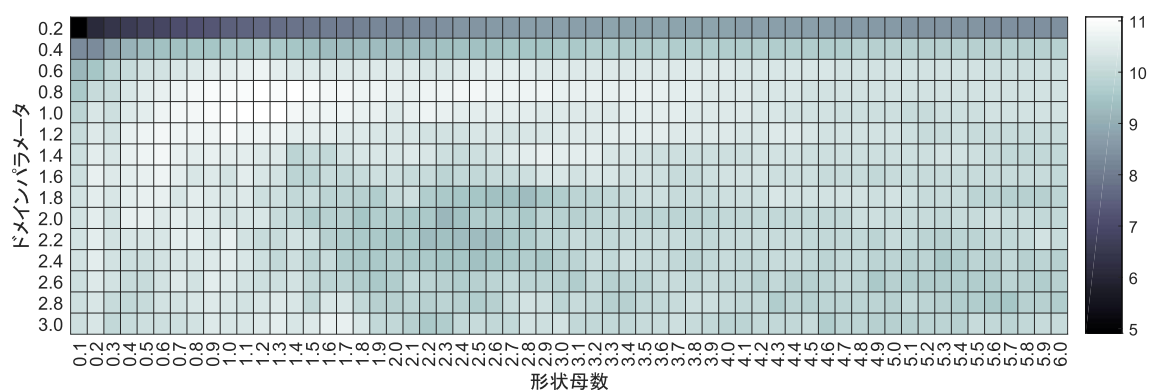


図 27 Tp.と Vc.の混合信号の実験結果 (SDR 値のヒートマップ)

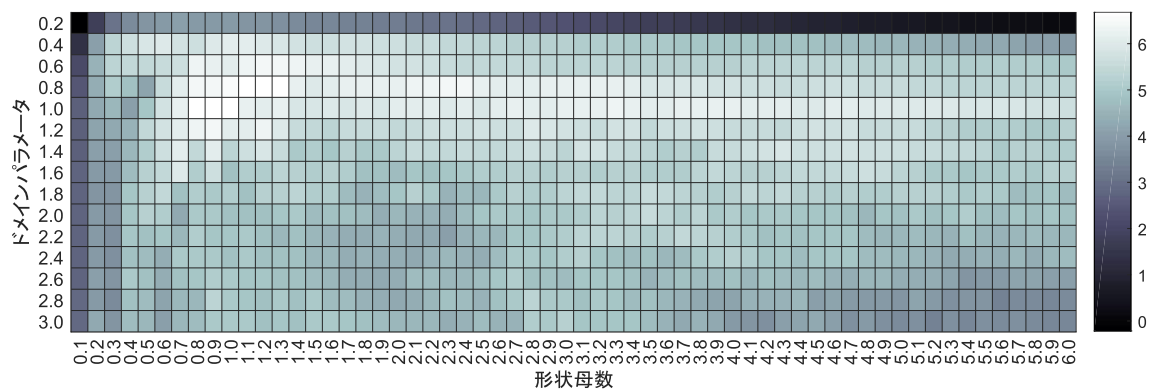


図 28 Tp.と Vn.の混合信号の実験結果 (SDR 値のヒートマップ)

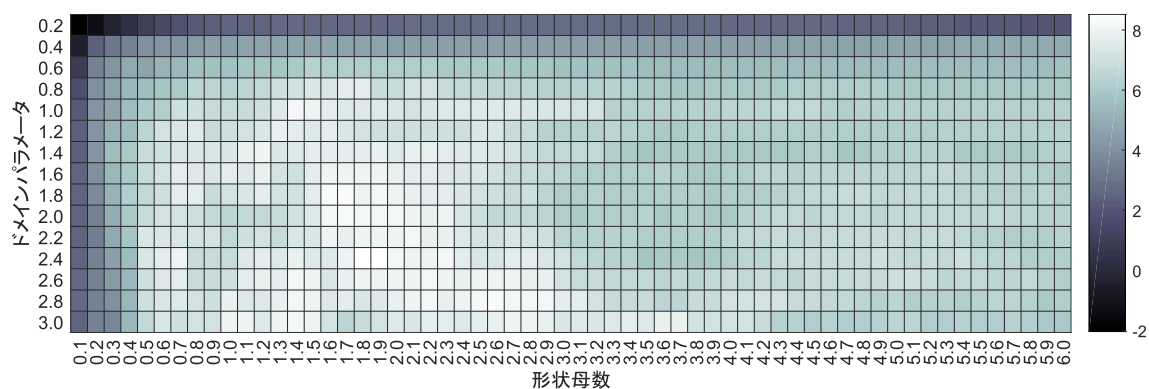


図 29 Cl.と Hp.の混合信号の実験結果 (SDR 値のヒートマップ)

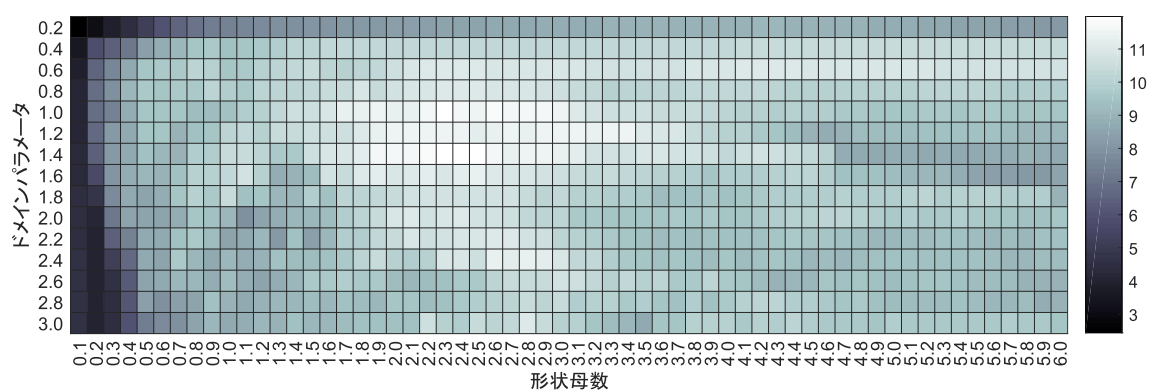


図 30 Cl.と Pf.の混合信号の実験結果 (SDR 値のヒートマップ)

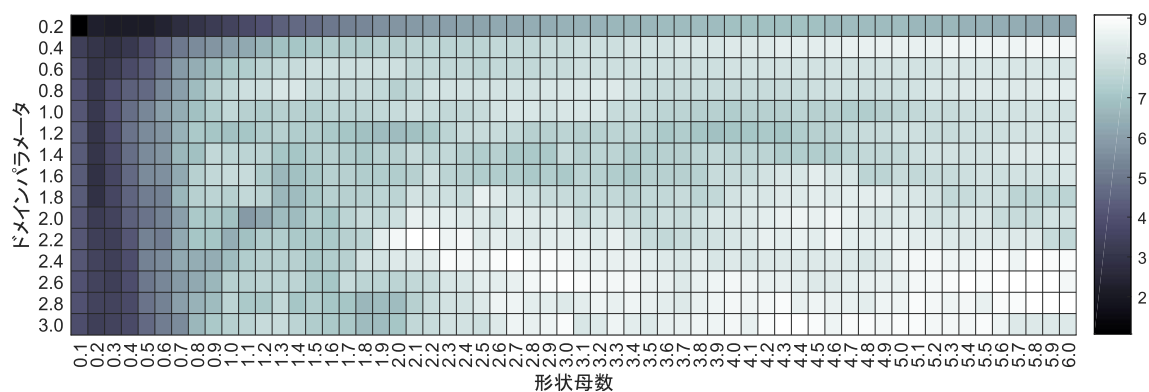


図 31 Cl.と Tb.の混合信号の実験結果 (SDR 値のヒートマップ)

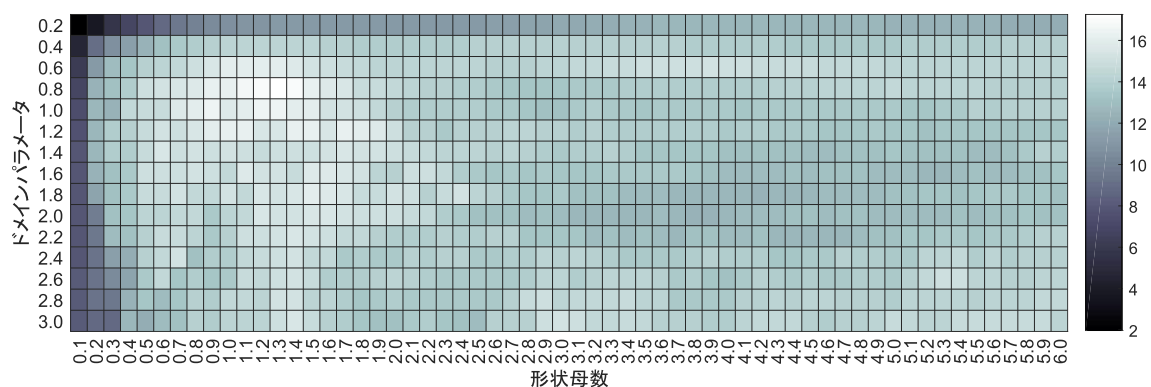


図 32 Cl.と Vc.の混合信号の実験結果 (SDR 値のヒートマップ)

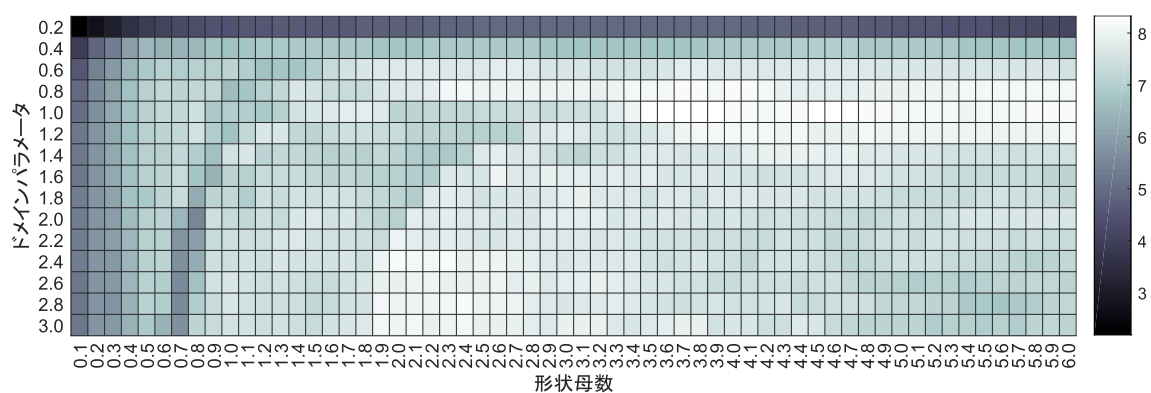


図 33 Vn.と Hp.の混合信号の実験結果 (SDR 値のヒートマップ)

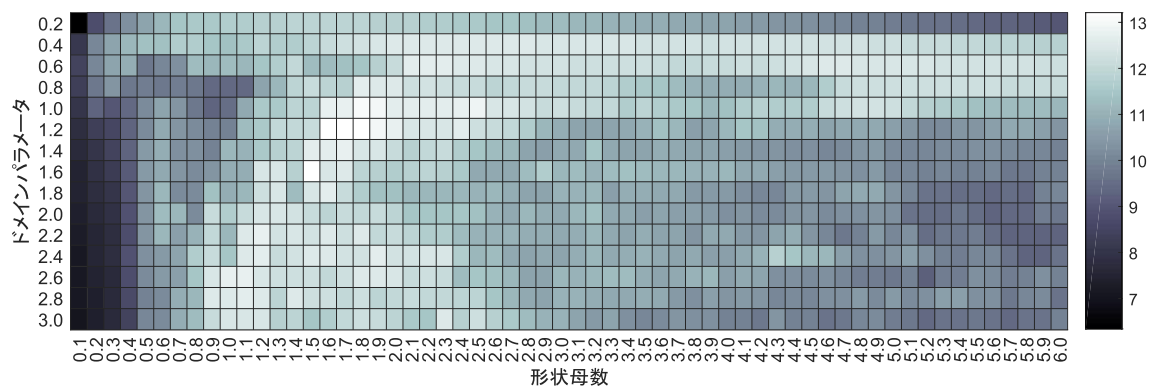


図 34 Vn.と Pf.の混合信号の実験結果 (SDR 値のヒートマップ)

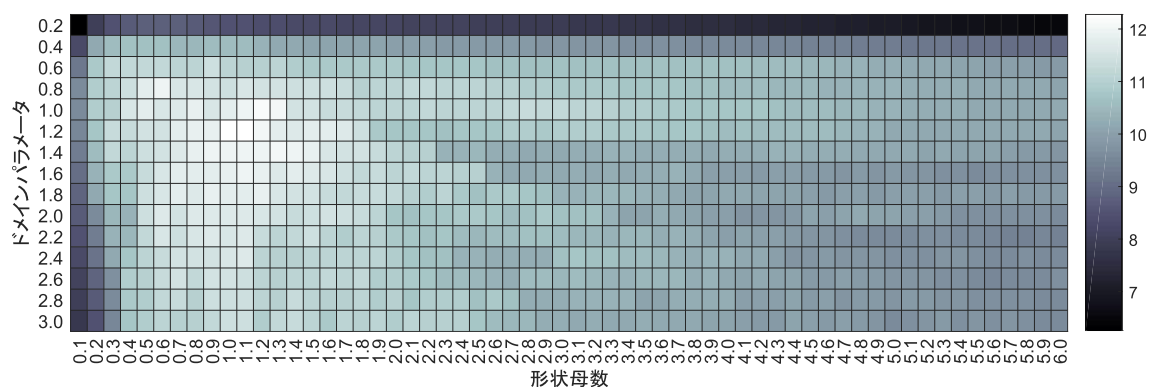


図 35 Hp.と Tb.の混合信号の実験結果 (SDR 値のヒートマップ)

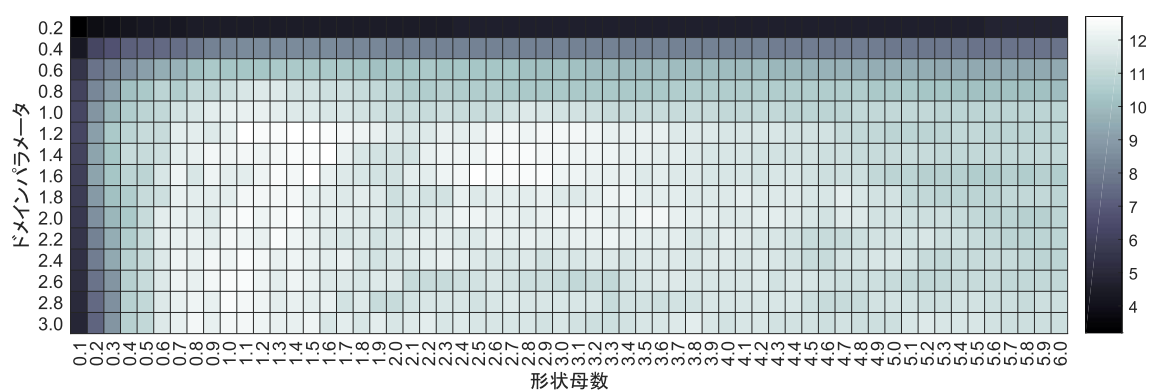


図 36 Hp.と Vc.の混合信号の実験結果 (SDR 値のヒートマップ)

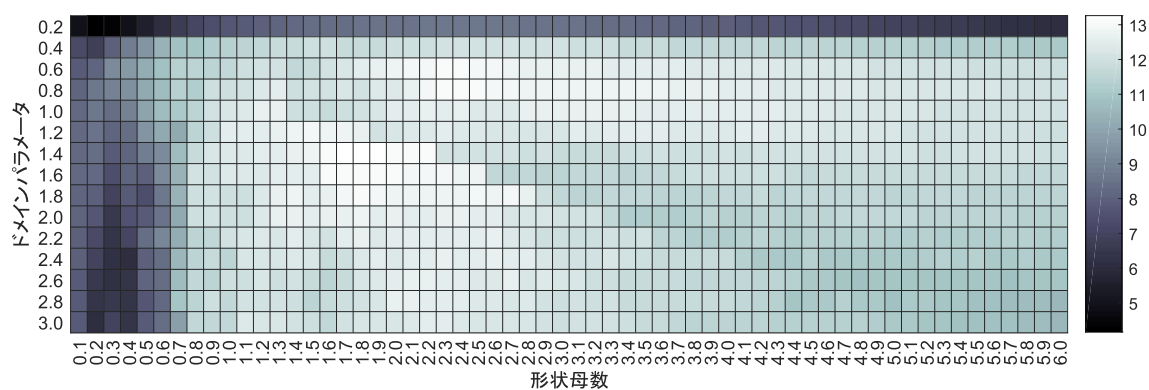


図 37 Pf.と Tb.の混合信号の実験結果 (SDR 値のヒートマップ)

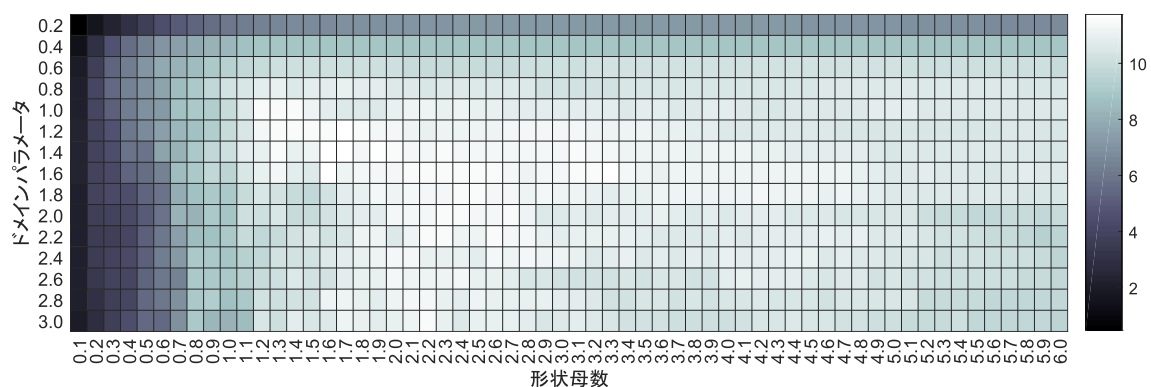


図 38 Pf.と Vc.の混合信号の実験結果 (SDR 値のヒートマップ)

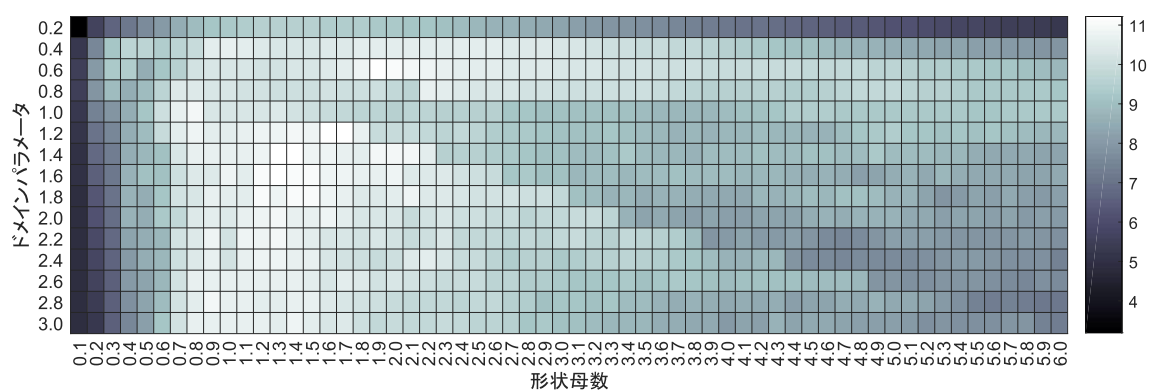


図 39 Vn.と Tb.の混合信号の実験結果 (SDR 値のヒートマップ)

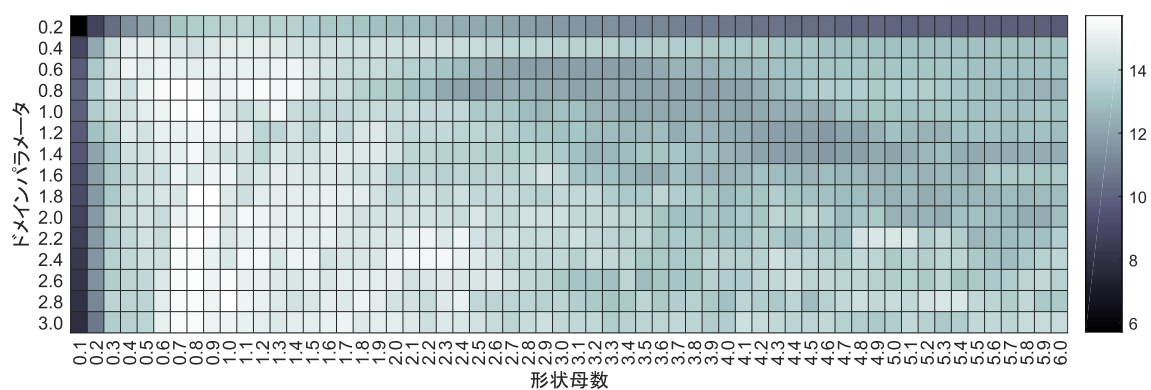


図 40 Vn.と Vc.の混合信号の実験結果 (SDR 値のヒートマップ)

5. GGD-NMF の最適パラメータ推定法の検討

4 章の結果から、GGD-NMF による教師あり音源分離は、従来の IS-NMF を用いる手法の性能を上回る例が散見された。しかしながら、SDR 値が最も高くなる時の GGD-NMF のパラメータ（形状母数 ρ 及びドメインパラメータ p ）は、混合されている音源の種類に強く依存しており、最良の分離性能を得るためには様々な条件で音源分離を行う必要がある。また、アプリケーションにおいては、混合信号中の「混合前の各音源信号（音源分離における正解信号）」は知りえないため、SDR のような客観評価指標を算出することはできない。この問題を解決するため、本章では、学習ステージを終えた段階で、音源分離に最適な GGD-NMF のパラメータを推定する方法を検討する。

5.1 音源分離性能の予測

分離ステージの音源分離が失敗する大きな要因は、学習ステージで得た各音源の音色の辞書である基底行列 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ が「似たような基底ベクトル（音色）を含む」ことである。例えば、 $\mathbf{W}_1 = \mathbf{W}_2$ として全く同じ基底行列が学習されてしまった場合、式(5)に示す分離ステージは原理的に 2 つの音源を分離することができない。なぜなら、分離ステージにおけるアクティベーション行列 $\mathbf{G}_1$ 及び $\mathbf{G}_2$ の推定は、式(24)のように「 $\mathbf{Y}$ と $\mathbf{W}_1\mathbf{G}_1 + \mathbf{W}_2\mathbf{G}_2$ を近づける」という基準しか考慮されておらず、 $\mathbf{W}_1 = \mathbf{W}_2$ ならば 2 つの音源の成分が $\mathbf{W}_1\mathbf{G}_1$ と $\mathbf{W}_2\mathbf{G}_2$ に分離される保証がないからである。

この事実に基づくと、学習ステージで得た $\mathbf{W}_1$ 及び $\mathbf{W}_2$ の類似度を測ることで、分離ステージでの音源分離性能を事前に予測することができると考えられる。もし $\mathbf{W}_1$ 及び $\mathbf{W}_2$ の類似度が高い場合（ $\mathbf{W}_1$ 及び $\mathbf{W}_2$ が似た基底ベクトルを多く含む場合）、分離ステージでの音源分離は困難になる。逆に類似度が低い場合、高性能な音源分離ができると予測される。

学習ステージで推定される $\mathbf{W}_1$ 及び $\mathbf{W}_2$ は、同じ 2 種類の音源を学習する場合であっても、NMF で仮定する生成モデル（すなわち形状母数 ρ やドメインパラメータ p ）に依存して変化する。従って、あらゆるパラメータで $\mathbf{W}_1$ 及び $\mathbf{W}_2$ の学習を行い、逐一類似度を計算し、最も類似度が低くなるパラメータとそのときの $\mathbf{W}_1$ 及び $\mathbf{W}_2$ を用いて分離ステージを行えば、最良の音源分離結果が得られる可能性が高い。

以上の仮説を立証するために、「 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ の類似度」と「音源分離性能」に強い相関関係があるか否かを実験的に調査する。両者に強い相関がある場合には、各音源のサンプル信号のみを用いて音源分離性能を予測することが可能となる。

5.2 音源分離性能と相関を持つ類似度の提案

5.1 節で述べた目的のためには、 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ の持つ基底ベクトル間の大きさ（音量）の違いではなく角度（音色）の違いのみを反映した類似度であることが望ましい。従って、次式に示す $\cos$ 類似度を用いることを提案する。

$$D_{\cos}(\mathbf{a}||\mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} \quad (52)$$

ここで、 $\mathbf{a}$ 及び $\mathbf{b}$ は同じ次元数の任意ベクトルである。この $\cos$ 類似度を、 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ 内の全基底ベクトル間で測り、それらを総和した値を $\mathbf{W}_1$ 及び $\mathbf{W}_2$ 間の類似度に用いる。従って、類似度は次式となる。

$$D_{\cos}(\mathbf{W}_1||\mathbf{W}_2) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} D_{\cos}(\mathbf{w}_{1k_1}||\mathbf{w}_{2k_2}) \quad (53)$$

ここで、 $\mathbf{w}_{1k_1}$ 及び $\mathbf{w}_{2k_2}$ は基底行列 $\mathbf{W}_1$ 及び $\mathbf{W}_2$ の基底ベクトル（列ベクトル）である。式(53)は行列形式にすると次式で表せる。

$$D_{\cos}(\mathbf{W}_1||\mathbf{W}_2) = \frac{\|\mathbf{W}_1^T \mathbf{W}_2\|_1}{\|\mathbf{W}_1\|_{\text{Fr}} \cdot \|\mathbf{W}_2\|_{\text{Fr}}} \quad (54)$$

ここで、 $\|\cdot\|_1$ 及び $\|\cdot\|_{\text{Fr}}$ はそれぞれ L_1 ノルム及びフロベニウスノルムであり、次式で表す。

$$\|\mathbf{C}\|_1 = \sum_{i=1}^I \sum_{j=1}^J |c_{ij}| \quad (55)$$

$$\|\mathbf{C}\|_{\text{Fr}} = \sqrt{\sum_{i=1}^I \sum_{j=1}^J |c_{ij}|^2} \quad (56)$$

ここで、 $\mathbf{C}$ はサイズが $I \times J$ の任意の行列であり、 c_{ij} は $\mathbf{C}$ の要素である。

5.3 相関関係の実験的調査

本実験の条件は、4.1 節に述べたものと同様である。各混合信号における式(54)の類似度と SDR 値との相関係数を図 41 に示す。これを見ると、ほとんどの混合信号において負の相関関係があることがわかる。すなわち、 $\cos$ 類似度が大きいほど（ $\mathbf{W}_1$ 及び $\mathbf{W}_2$ が似ているほど）音源分離性能が劣化する事実を示している。従って、 $\cos$ 類似度が小さくなるパラメータを選ぶことで高性能な音源分離結果が得られる可能性が高い。ただし、相関係数の平均値は-0.47 となっており、必ずしも強い相関とはいえないため、 $\cos$ 類似度以外の基準についても検討の余地が残る。

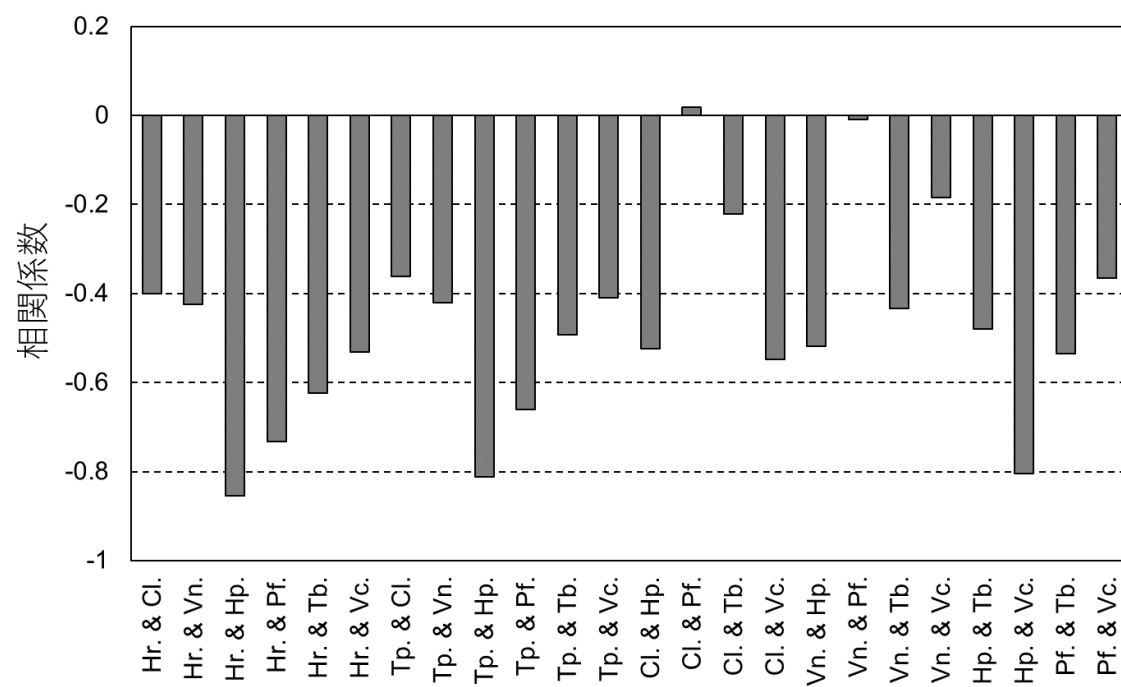


図 41 各楽器の組み合わせ毎の相関係数

6. まとめ

本論文では、省サンプルの教師あり音源分離を対象として、現時点で調査が不十分である複素 GGD に基づく教師あり NMF の実験的な性能調査を行った。特に、既存の一般的な手法である IS-NMF の場合と比較した。多くの混合信号において、GGD-NMF による音源分離が IS-NMF の分離性能を上回ることを実験により示した。ただし、音源分離にとって最適な複素 GGD のパラメータは、混合されている音源の種類や組み合わせに強く依存する結果となった。この問題を解決するために、音源分離性能を予測するための基準として、cos 類似度を用いることを新たに提案した。実験の結果、cos 類似度と音源分離性能は負の相関関係を持つことが明らかとなり、音源分離性能予測のための指標となることを示した。ただし、あまり強い相関とはいえないため、今後 cos 類似度を追及するならば、楽器音の種類等、混合されている音源にとって重要な周波数をより考慮するような重みづけ cos 類似度を用いる手法等が考えられる。cos 類似度はベクトルの角度にしか着目しないため、 $I = 2049$ のように高い次元のベクトルはほとんど直行していることから不適切である可能性も残る。従って、ベクトルの距離の長さも考慮した尺度についても検討する必要がある。今後の課題として、GGD-NMF の音声信号に対する音源分離への適用や、音源分離性能とより強い相関関係を持つ基準の検討等が挙げられる。

謝辞

本研究を実施するにあたり、ご多忙の中にも関わらずご指導ご鞭撻を賜り、本論文をまとめる際も親身になって御助言を頂いた北村大地助教に心より厚く御礼申し上げます。そして本論文を査読していただいた重田和弘教授にも厚く御礼申し上げます。最後に、今まで生活を支えてくれた両親に心より感謝いたします。

参考文献

- [1] 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.
- [2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Proc. Advances in neural information processing systems*, pp. 556–562, 2001.
- [3] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.
- [4] A. Ozerov, C. Fevotte and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 381–386, 2009.
- [5] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama, "Constrained and regularized variants of nonnegative matrix factorization incorporating music-specific constraints," *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 5365–5368, 2012.
- [6] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. International Conference on Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.
- [7] 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 Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E97-A, no. 5, pp. 1113–1118, 2014.
- [8] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," *Proc. International Conference on Independent Component Analysis and Signal Separation*, pp. 3734–3738, 2014.
- [9] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," *Proc. International Conference on Independent Component Analysis and Signal Separation*, pp. 2135–2139, 2015.

- [10] 北村大地, 高宗典玄, 最上伸一, 三井祥幹, 猿渡洋, 高橋祐, 近藤多伸, "ヘビーテイルな分布に基づく非負値行列因子分解を用いたスパース雑音除去," *日本音響学会講演論文集*, pp. 441–444, 2018.
- [11] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, p. 28, 2018.
- [12] C. Fevotte, 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.
- [13] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with b-divergence," *Proc. IEEE International Workshop on Machine Learning and Signal Processing*, pp. 283–288, 2010.
- [14] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.