インパルス応答のスパース性を用いた教師あり独立低ランク行列分析*

☆高橋 悠希，△北原 大地，平林 晃 (立命館大)

1 はじめに

音源分離とは，複数音源からなる混合音を，混合前の各音源信号に分離する技術である [1]–[4]．この技術は，自動採譜や楽器カラオケの作成，音声認識の精度向上等への活用が期待されている．音源分離問題は，学習データを用いない「ブラインド音源分離」と学習データを用いる「教師あり音源分離」に大別される．ブラインド音源分離では，音源や混合過程に関する情報がほぼ未知であるため，良好な分離結果を得ることは難しい．一方で，教師あり音源分離では，学習データを利用することで高精度な分離が可能となる．

音源信号の混合過程は，時間領域では畳み込みとなり取り扱いが難しいため，周波数領域で乗算としてモデル化することが多い．過決定条件下 (マイク数 ≥ 音源数) における音源分離では，周波数ビンごとに，混合系 (混合行列) の逆写像である分離系 (分離行列) を求める手法が数多く提案されている [1]–[4]．これらの手法では，各音源信号の生成モデルを統計的独立性と優ガウス性に基づいて設計した上で，最尤推定により分離行列を決定している．例えば，独立低ランク行列分析 (Independent Low-Rank Matrix Analysis: ILRMA) [3] では，音源信号のパワースペクトログラムが非負値行列因子分解 (Nonnegative Matrix Factorization: NMF) [5] を用いて低ランク近似できることに着目して，時変複素ガウス分布を音源信号の生成モデルとしている．更に，混合過程に関する一般的な性質である「音源からマイクへのインパルス応答のスパース性」を考慮したアルゴリズム [4] も提案されており，ILRMA よりも優れた分離性能を示している．

教師情報がある場合は，劣決定条件下 (マイク数 < 音源数) においても，学習した音源信号の情報を利用して各音源を高精度に分離できる．特に，マイク数が 1 つの場合には，教師あり NMF や深層学習を用いた手法が提案されている [6]–[9]．教師あり NMF [6]–[8] は，スペクトルパターンが低ランクとなる楽器等の音源に対して，基底を事前に学習する手法である．

本研究では，楽曲の譜面情報や使用される楽器の種類が事前に分かっていることを想定して，ILRMA 内で用いられる NMF を教師あり NMF に置き換えることを提案する．提案法では，ILRMA における非負値基底行列を事前に学習しておくことで，推定変数の数が減り，より優れた分離行列が求まる．基底行列の学習には文献 [10] の手法を用い，音源信号の各単音のパワースペクトログラムから 1 つ以上の基底ベクトルを作成する．更に，文献 [4] と同様にインパルス応答のスパース性を用いたアルゴリズムも提案する．MIDI 音源と公開されているインパルス応答を用いた数値実験により提案法の優れた分離性能を示す．

2 過決定条件下でのブラインド音源分離

2.1 観測モデル

音源数を $N\ (\geq 2)$ で表し，マイク数を $M\ (\geq N)$ で表す．n 番目 $(n = 1, 2, \ldots, N)$ の音源で時刻 t に生成される音を $s_n[t]$ とし，m 番目 $(m = 1, 2, \ldots, M)$ のマイクで時刻 t に観測される音を $x_m[t]$ とする．n 番目の音源から m 番目のマイクへのインパルス応答を $h_{m,n}[\tau]\ (\tau = 0, 1, \ldots, T-1)$ とすると，観測音は

$$x_m[t] = \sum_{n=1}^{N} (h_{m,n} * s_n)[t] = \sum_{n=1}^{N} \sum_{\tau=0}^{T-1} h_{m,n}[\tau] s_n[t-\tau] \tag{1}$$

と表される．ここで，$*$ は畳み込み演算である．離散時間信号 $s_n[t]$ の短時間フーリエ変換結果を

$$\hat{s}_{i,j,n} := \sum_{\tau=0}^{L-1} \psi[\tau] s_n[(j-1)\eta + \tau] \exp\left(-\iota \frac{2\pi(i-1)\tau}{L}\right)$$

と定義し，$x_m[t]$ の短時間フーリエ変換結果も同様に $\hat{x}_{i,j,m}$ と定義する．ここで，L はフレーム長，$\psi[\tau]$ は窓関数，η はフレームシフト量，ι は虚数単位，$i = 1, 2, \ldots, L$ は周波数インデックス，$j = 1, 2, \ldots, J$ はフレームインデックスを表す．インパルス応答 $h_{m,n}[\tau]$ の離散時間フーリエ変換 $\sum_{\tau=0}^{T-1} h_{m,n}[\tau] \exp(-\iota\omega\tau)$ を $\omega_i := \frac{2\pi(i-1)}{L}$ でサンプリングしたものを

$$\hat{a}_{i,m,n} := \sum_{\tau=0}^{T-1} h_{m,n}[\tau] \exp\left(-\iota \frac{2\pi(i-1)\tau}{L}\right)$$

と定義する．短時間フーリエ変換のフレーム長 L がインパルス応答長 T より十分長いならば，フレームごとに式 (1) の「時間領域における畳み込み演算」を「周波数領域における乗算」に変換することができる．ここで，$\hat{\epsilon}_{i,j,m}$ を短時間フーリエ変換で得られる周波数成分と真の周波数成分の差に起因するモデル誤差とし，ベクトル $\boldsymbol{s}_{i,j}$, $\boldsymbol{x}_{i,j}$, $\boldsymbol{\epsilon}_{i,j}$, $\boldsymbol{a}_{i,j}$ と混合行列 A_i を

$$\begin{cases} \boldsymbol{s}_{i,j} := (\hat{s}_{i,j,1}, \hat{s}_{i,j,2}, \ldots, \hat{s}_{i,j,N})^{\mathrm{T}} \in \mathbb{C}^N \\ \boldsymbol{x}_{i,j} := (\hat{x}_{i,j,1}, \hat{x}_{i,j,2}, \ldots, \hat{x}_{i,j,M})^{\mathrm{T}} \in \mathbb{C}^M \\ \boldsymbol{\epsilon}_{i,j} := (\hat{\epsilon}_{i,j,1}, \hat{\epsilon}_{i,j,2}, \ldots, \hat{\epsilon}_{i,j,M})^{\mathrm{T}} \in \mathbb{C}^M \\ \boldsymbol{a}_{i,n} := (\hat{a}_{i,1,n}, \hat{a}_{i,2,n}, \ldots, \hat{a}_{i,M,n})^{\mathrm{T}} \in \mathbb{C}^M \\ A_i := (\boldsymbol{a}_{i,1}, \boldsymbol{a}_{i,2}, \ldots, \boldsymbol{a}_{i,N}) \in \mathbb{C}^{M \times N} \end{cases}$$

と定義すれば，式 (1) の観測モデルは周波数領域で

$$\boldsymbol{x}_{i,j} = A_i \boldsymbol{s}_{i,j} + \boldsymbol{\epsilon}_{i,j} \qquad (i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, J) \tag{2}$$

のように表される．ここで，i 番目と $(L-i+2)$ 番目の周波数成分は複素共役であるため $I = \lfloor \frac{L+2}{2} \rfloor$ となる．ブラインド音源分離は，A_i が未知の状態で教師情報も用いずに，$\boldsymbol{x}_{i,j}$ から $\boldsymbol{s}_{i,j}$ を推定する問題である．

---

*Supervised independent low-rank matrix analysis using the sparsity of impulse responses. By Yuki TAKAHASHI, Daichi KITAHARA, Akira HIRABAYASHI (Ritsumeikan University).

2.2 独立低ランク行列分析 (ILRMA)

過決定条件下においては，A_i の一般逆行列

$$W_i := (\boldsymbol{w}_{i,1}, \boldsymbol{w}_{i,2}, \ldots, \boldsymbol{w}_{i,N})^{\mathrm{H}} := A_i^{\dagger} \in \mathbb{C}^{N\times M}$$

を求めることで，式 (2) の $\boldsymbol{s}_{i,j}$ を

$$\begin{aligned}\boldsymbol{y}_{i,j} &:= (\hat{y}_{i,j,1}, \hat{y}_{i,j,2}, \ldots, \hat{y}_{i,j,N})^{\mathrm{T}} \in \mathbb{C}^N \\ &:= (\boldsymbol{w}_{i,1}^{\mathrm{H}}\boldsymbol{x}_{i,j}, \boldsymbol{w}_{i,2}^{\mathrm{H}}\boldsymbol{x}_{i,j}, \ldots, \boldsymbol{w}_{i,N}^{\mathrm{H}}\boldsymbol{x}_{i,j})^{\mathrm{T}} = W_i\boldsymbol{x}_{i,j}\end{aligned}$$

のように推定する手法が主流である [1]–[4]．ここで，W_i は分離行列と呼ばれる．これらの手法では，「複素スペクトログラム $\bar{S}_n := (\hat{s}_{i,j,n}) \in \mathbb{C}^{I\times J}$ が優ガウス的な分布から独立に生成される」と仮定して分離行列 W_i を決定している．独立低ランク行列分析 (ILRMA) [3] では，パワースペクトログラム $|\bar{S}_n|^2 := (|s_{i,j,n}|^2) \in \mathbb{R}_+^{I\times J}$ が非負値の基底行列 $B_n := (b_{i,k,n}) \in \mathbb{R}_+^{I\times K_n}$ と係数行列 $C_n := (c_{k,j,n}) \in \mathbb{R}_+^{K_n\times J}$ の積で低ランク近似できることに着目して，「分散が NMF で定義された時変複素ガウス分布」を $\bar{S}_n$ の生成モデルとしている．ここで，$K_n \ll \min(I, J)$ は基底ベクトルの数である．

$\boldsymbol{\mathcal{W}} := (W_1, W_2, \ldots, W_I) \in \mathbb{C}^{N\times M\times I}$ と非負値行列 $B := (B_1, B_2, \ldots, B_N)$, $C := (C_1^{\mathrm{T}}, C_2^{\mathrm{T}}, \ldots, C_N^{\mathrm{T}})^{\mathrm{T}}$ を観測値 $\boldsymbol{\mathcal{X}} := (\hat{x}_{i,j,m}) \in \mathbb{C}^{I\times J\times M}$ の確率密度関数のパラメータ $\Theta = (\boldsymbol{\mathcal{W}}, B, C)$ とすると，負対数尤度は

$$-\log(p(\boldsymbol{\mathcal{X}}|\Theta)) := \sum_{n=1}^{N}\sum_{i=i}^{I}\sum_{j=1}^{J}\left[\frac{\boldsymbol{w}_{i,n}^{\mathrm{H}}\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^{\mathrm{H}}\boldsymbol{w}_{i,n}}{\sum_{k=1}^{K_n} b_{i,k,n}c_{k,j,n}} + \log\sum_{k=1}^{K_n} b_{i,k,n}c_{k,j,n} + \log\pi\right] - J\sum_{i=1}^{I}\log\det(W_iW_i^{\mathrm{H}})$$

と表される．最尤推定の考えに基づけば，最小化問題

$$\underset{\boldsymbol{\mathcal{W}},B,C}{\mathrm{minimize}}\ -\log(p(\boldsymbol{\mathcal{X}}|\Theta)) \quad \text{s.t.}\ \forall n\ \|\bar{Y}_n\|_{\mathrm{F}}^2 = IJ \quad (3)$$

を解くことで各 W_i が求まる．ここで，$\bar{Y}_n := (\hat{y}_{i,j,n}) \in \mathbb{C}^{I\times J}$ に関する制約は，混合系に対する定数倍の任意性を解消するために必要となる．まず変数 B と C を

$$\begin{cases} b_{i,k,n}^{(l+1)} = b_{i,k,n}^{(l)}\sqrt{\dfrac{\sum_{j=1}^{J}|\hat{y}_{i,j,n}^{(l)}|^2\, c_{k,j,n}^{(l)}\big(\sum_{k'=1}^{K_n} b_{i,k',n}^{(l)}c_{k',j,n}^{(l)}\big)^{-2}}{\sum_{j=1}^{J} c_{k,j,n}^{(l)}\big(\sum_{k'=1}^{K_n} b_{i,k',n}^{(l)}c_{k',j,n}^{(l)}\big)^{-1}}} \\ c_{k,j,n}^{(l+1)} = c_{k,j,n}^{(l)}\sqrt{\dfrac{\sum_{i=1}^{I}|\hat{y}_{i,j,n}^{(l)}|^2\, b_{i,k,n}^{(l+1)}\big(\sum_{k'=1}^{K_n} b_{i,k',n}^{(l+1)}c_{k',j,n}^{(l)}\big)^{-2}}{\sum_{i=1}^{I} b_{i,k,n}^{(l+1)}\big(\sum_{k'=1}^{K_n} b_{i,k',n}^{(l+1)}c_{k',j,n}^{(l)}\big)^{-1}}} \end{cases} \quad (4)$$

のように更新し，次に変数 $\boldsymbol{\mathcal{W}}$ を

$$\begin{cases} U_{i,n}^{(l+1)} = \dfrac{1}{J}\displaystyle\sum_{j=1}^{J}\frac{\boldsymbol{x}_{i,j}\boldsymbol{x}_{i,j}^{\mathrm{H}}}{\sum_{k=1}^{K_n} b_{i,k,n}^{(l+1)}c_{k,j,n}^{(l+1)}} \\ \boldsymbol{v}_{i,n}^{(l+1)} = \big(U_{i,n}^{(l+1)}\big)^{-1}\boldsymbol{a}_{i,n}^{(l)} \\ \boldsymbol{w}_{i,n}^{(l+1)} = \dfrac{\boldsymbol{v}_{i,n}^{(l+1)}}{\sqrt{\boldsymbol{v}_{i,n}^{(l+1)\mathrm{H}}U_{i,n}^{(l+1)}\boldsymbol{v}_{i,n}^{(l+1)}}} \end{cases} \quad (5)$$

のように更新する．その後，制約を満たすように変数のスケールを調整する．これらを複数回繰り替えすことで，式 (3) の問題を解く．最終的な分離結果は，$(\hat{\hat{y}}_{i,j,1,n}, \hat{\hat{y}}_{i,j,2,n}, \ldots, \hat{\hat{y}}_{i,j,M,n})^{\mathrm{T}} := \hat{y}_{i,j,n}\boldsymbol{a}_{i,n}$ を計算し，$(\hat{\hat{y}}_{i,j,m,n}) \in \mathbb{C}^{I\times J}$ を逆短時間フーリエ変換して求める．

2.3 インパルス応答のスパース性を用いた ILRMA

ブラインド音源分離の多くの手法が，「スペクトログラムの各成分 $\hat{s}_{i,j,n}$ は高い頻度で 0 になる」という音源信号の先験情報に基づいており，$\bar{S}_n$ の確率密度関数を複雑に設計することで分離精度の向上を図っていた．文献 [4] で筆者らは混合過程の先験情報も利用することを考え，音源やマイクの位置に依らずに成り立つ混合過程の性質として，「インパルス応答 $h_{m,n}[\tau]$ のスパース性」に着目した．新たにベクトル

$$\boldsymbol{h}_{m,n} := (h_{m,n}[0], h_{m,n}[1], \ldots, h_{m,n}[T-1])^{\mathrm{T}} \in \mathbb{R}^T$$

を定義し，全てをまとめて $\boldsymbol{\mathcal{H}} := (\boldsymbol{h}_{m,n}) \in \mathbb{R}^{T\times M\times N}$ とする．$i = 2, 3, \ldots, \lfloor\frac{L+1}{2}\rfloor$ に対して，$A_i = W_i^{\dagger}$ の各成分 $\hat{a}_{i,m,n}$ の複素共役として $\hat{a}_{L-i+2,m,n}$ を定義し，$\bar{\boldsymbol{a}}_{m,n} := (\hat{a}_{1,m,n}, \hat{a}_{2,m,n}, \ldots, \hat{a}_{L,m,n})^{\mathrm{T}} \in \mathbb{C}^L$ とする．

筆者らは，Θ の負対数尤度 $-\log(p(\boldsymbol{\mathcal{X}}|\Theta))$ に加えてインパルス応答のスパース性も考慮した最適化問題

$$\begin{aligned}&\underset{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{H}},B,C}{\mathrm{minimize}}\ -\log(p(\boldsymbol{\mathcal{X}}|\Theta)) + \lambda J[\mathcal{F}(\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{H}}) + L\mathcal{G}(\boldsymbol{\mathcal{H}})] \\ &\text{s.t.}\ \forall n \sum_{m=1}^{M}\|\bar{\boldsymbol{a}}_{m,n}\|_2^2 = L \ \text{ and } \ \forall n \sum_{m=1}^{M}\|\boldsymbol{h}_{m,n}\|_2^2 = 1\end{aligned} \quad (6)$$

を解くことを提案した．ここで，$\mathcal{F}$ は $\boldsymbol{\mathcal{W}}$ と $\boldsymbol{\mathcal{H}}$ の整合性を評価する特殊な関数であり，インパルス応答 $\boldsymbol{\mathcal{H}}$ から計算される混合行列 $\widetilde{A}_i$ と分離行列 $\widetilde{W}_i$ を用いて

$$\mathcal{F}(\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{H}}) := \begin{cases} \displaystyle\sum_{i=1}^{I}\big\|W_i - \widetilde{W}_i\big\|_{\mathrm{F}}^2 & (\boldsymbol{\mathcal{W}}\text{ の更新時}) \\ \displaystyle\sum_{i=1}^{L}\big\|A_i - \widetilde{A}_i\big\|_{\mathrm{F}}^2 & (\boldsymbol{\mathcal{H}}\text{ の更新時}) \end{cases}$$

と定義される．$\mathcal{G}$ はインパルス応答のスパース性及び振幅の単調減少傾向を評価する関数であり，2 値関数 $\Gamma : \mathbb{R}\backslash\{0\} \ni h \mapsto 1 : 0 \mapsto 0$ と単調増加列 $\kappa[\tau]$ を用いて

$$\mathcal{G}(\boldsymbol{\mathcal{H}}) := \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{\tau=0}^{T-1}\kappa[\tau]\Gamma(h_{m,n}[\tau])$$

と定義される．$\bar{\boldsymbol{a}}_{m,n}$ と $\boldsymbol{h}_{m,n}$ に関する制約は，定数倍の任意性解消のために必要となる．$\lambda > 0$ が適切であれば，ILRMA の分離性能を上回ることができる．

3 インパルス応答のスパース性を用いた教師あり独立低ランク行列分析

ブラインド音源分離では，音源や混合過程に関する情報がほぼ未知であるため，各音源を高精度に分離することが難しく，しばしば分離に失敗してしまう．一方，教師あり音源分離では，音源信号の情報を学習できるため，マイク数が少ない場合でも高精度な分離が可能となる．本研究では，使用される楽器の種類が既知である状況を想定し，ILRMA で用いる基底行列 B_n を事前に学習することを提案する．基底ベクトルは，文献 [10] の手法により，音源信号の単音からなるパワースペクトログラムを任意精度で近似するように作成される．更に，文献 [4] と同様にインパルス応答のスパース性を考慮したアルゴリズムも提案する．

3.1 ランク r 近似による基底行列の作成

教師あり NMF において音源信号を簡単に学習するには，各単音のパワースペクトログラム $|S_{\mathrm{mono}}|^2 \in \mathbb{R}_+^{I\times J}$ に対して，基底ベクトル $\tilde{\boldsymbol{b}}_{\mathrm{mono}}$ をランク 1 近似

$$(\tilde{\boldsymbol{b}}_{\mathrm{mono}}, \tilde{\boldsymbol{c}}_{\mathrm{mono}}) = \underset{\tilde{\boldsymbol{b}}\in\mathbb{R}_+^I, \tilde{\boldsymbol{c}}\in\mathbb{R}_+^J}{\operatorname{argmin}} \left\| |S_{\mathrm{mono}}|^2 - \tilde{\boldsymbol{b}}\tilde{\boldsymbol{c}}^{\mathrm{T}} \right\|_{\mathrm{F}}$$

により求めればよい．具体的に，$\tilde{\boldsymbol{b}}_{\mathrm{mono}}$ は $|S_{\mathrm{mono}}|^2$ を特異値分解した際の第 1 左特異ベクトルとして計算される．文献 [10] では，音の鳴り始め等を詳細に表現するために，単音 $|S_{\mathrm{mono}}|^2$ から複数の基底ベクトル $(\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \ldots, \tilde{\boldsymbol{b}}_{r_*}) =: \widetilde{B}_{r_*} \in \mathbb{R}_+^{I\times r_*}$ が設計された．基底ベクトルの数 r_* は，$\epsilon > 0$ により設定される近似精度

$$\Upsilon(\widetilde{B}_r) := \min_{\widetilde{C}_r\in\mathbb{R}_+^{r\times J}} \left\| |S_{\mathrm{mono}}|^2 - \widetilde{B}_r\widetilde{C}_r \right\|_{\mathrm{F}} \le \epsilon \tag{7}$$

を満たす最小の r とする．具体的な手順を以下に記す．

ステップ 1. 式 (7) を満たす最小の r の厳密計算は困難であるため，代わりに $|S_{\mathrm{mono}}|^2$ を特異値分解して

$$\begin{aligned}\min_{\widetilde{B}_r\in\mathbb{R}_+^{I\times r}} \Upsilon(\widetilde{B}_r) &\ge \min_{\widehat{B}_r\in\mathbb{R}^{I\times r}, \widehat{C}_r\in\mathbb{R}^{r\times J}} \left\| |S_{\mathrm{mono}}|^2 - \widehat{B}_r\widehat{C}_r \right\|_{\mathrm{F}} \\ &= \left\| |S_{\mathrm{mono}}|^2 - |S_{\mathrm{mono}}|_r^2 \right\|_{\mathrm{F}} = \|\Sigma - \Sigma_r\|_{\mathrm{F}}\end{aligned}$$

を考える．$|S_{\mathrm{mono}}|_r^2$ は $|S_{\mathrm{mono}}|^2$ のランク r 近似であり，Σ と Σ_r は $|S_{\mathrm{mono}}|^2$ と $|S_{\mathrm{mono}}|_r^2$ の特異値行列である．$\|\Sigma - \Sigma_r\|_{\mathrm{F}} \le \epsilon$ を満たす最小の r の値を r_* とする．

ステップ 2. まず，$|S_{\mathrm{mono}}|^2$ を特異値分解した際の第 1 左特異ベクトルを第 1 基底ベクトル $\tilde{\boldsymbol{b}}_1$ とする．残りの基底ベクトル $\tilde{\boldsymbol{b}}_2, \tilde{\boldsymbol{b}}_3, \ldots, \tilde{\boldsymbol{b}}_{r_*}$ はランク r_* 近似行列 $|S_{\mathrm{mono}}|_{r_*}^2$ の列ベクトルから貪欲法的に選出する．具体的には，既に求めた基底ベクトル $\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \ldots, \tilde{\boldsymbol{b}}_k$ が張る部分空間 $\mathrm{span}\{\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \ldots, \tilde{\boldsymbol{b}}_k\}$ と $|S_{\mathrm{mono}}|_{r_*}^2$ の各列ベクトルがなす角度を直交射影を用いて計算し，角度が最大となる列ベクトルを新たに $\tilde{\boldsymbol{b}}_{k+1}$ とする．

ステップ 3. ステップ 2 で得られた各基底ベクトル $\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \ldots, \tilde{\boldsymbol{b}}_{r_*}$ に含まれる負の成分を 0 に置き換える．

3.2 提案法における各変数の更新アルゴリズム

提案する教師あり ILRMA では，式 (3) において B が変数ではなくなるため，式 (4) の C と式 (5) の $\boldsymbol{\mathcal{W}}$ の更新，及びスケール調整を繰り返す．式 (6) の問題の B を固定する場合には，式 (4) の C を更新した後に，$\boldsymbol{\mathcal{W}}$ は $\widetilde{W}_i^{(l)}$ と $\widetilde{U}_{i,n}^{(l+1)} := U_{i,n}^{(l+1)} + \lambda E_M$ を用いて

$$\begin{cases} \boldsymbol{v}_{i,n}^{(l+1)} = \big(\widetilde{U}_{i,n}^{(l+1)}\big)^{-1}\boldsymbol{a}_{i,n}^{(l)} \\ \widetilde{\boldsymbol{v}}_{i,n}^{(l+1)} = \lambda\big(\widetilde{U}_{i,n}^{(l+1)}\big)^{-1}\widetilde{\boldsymbol{w}}_{i,n}^{(l)} \\ d_{i,n}^{(l+1)} = \boldsymbol{v}_{i,n}^{(l+1)\mathrm{H}}\widetilde{U}_{i,n}^{(l+1)}\boldsymbol{v}_{i,n}^{(l+1)} \\ \widetilde{d}_{i,n}^{(l+1)} = \boldsymbol{v}_{i,n}^{(l+1)\mathrm{H}}\widetilde{U}_{i,n}^{(l+1)}\widetilde{\boldsymbol{v}}_{i,n}^{(l+1)} \\ \boldsymbol{w}_{i,n}^{(l+1)} = \begin{cases} \dfrac{\boldsymbol{v}_{i,n}^{(l+1)}}{\sqrt{d_{i,n}^{(l+1)}}} + \widetilde{\boldsymbol{v}}_{i,n}^{(l+1)} & \text{if } \widetilde{d}_{i,n}^{(l+1)} = 0, \\ \dfrac{\widetilde{d}_{i,n}^{(l+1)}}{2d_{i,n}^{(l+1)}}\left[\sqrt{1 + \dfrac{4d_{i,n}^{(l+1)}}{|\widetilde{d}_{i,n}^{(l+1)}|^2}} - 1\right]\boldsymbol{v}_{i,n}^{(l+1)} + \widetilde{\boldsymbol{v}}_{i,n}^{(l+1)} & \text{otherwise}, \end{cases} \end{cases} \tag{8}$$

のように更新される．その後，$\forall n\ \sum_{m=1}^M \|\bar{\boldsymbol{a}}_{m,n}\|_2^2 = L$ を満たすように $\boldsymbol{\mathcal{W}}$ と C のスケールを調整する．$\boldsymbol{\mathcal{H}}$ は，分離行列 $W_i^{(l+1)}$ から計算されるインパルス応答 $\widetilde{\boldsymbol{h}}_{m,n}^{(l+1)} := \Phi^{\mathrm{H}}\bar{\boldsymbol{a}}_{m,n}^{(l+1)}/L$ (Φ はフーリエ変換) を用いて

$$h_{m,n}^{(l+1)}[\tau] = \begin{cases} \widetilde{h}_{m,n}^{(l+1)}[\tau] & \text{if } \left|\widetilde{h}_{m,n}^{(l+1)}[\tau]\right| \ge \sqrt{\kappa[\tau]}, \\ 0 & \text{if } \left|\widetilde{h}_{m,n}^{(l+1)}[\tau]\right| < \sqrt{\kappa[\tau]}, \end{cases} \tag{9}$$

のように更新される．その後，$\forall n\ \sum_{m=1}^M \|\boldsymbol{h}_{m,n}\|_2^2 = 1$ を満たすように $\boldsymbol{\mathcal{H}}$ のスケールを調整する．式 (4)，(8)，(9) を複数回繰り返すことで，式 (6) の問題を解く．

4 シミュレーション

4.1 実験概要

2 音源 2 マイクロホンの状況を想定した数値実験で提案法の有効性を検証する．実験では，従来法である ILRMA [3] 及びインパルス応答のスパース性を利用した ILRMA (ILRMA+sparse) [4] と，提案法である教師あり ILRMA (提案法 1) 及びインパルス応答のスパース性を利用した教師あり ILRMA (提案法 2) の各分離性能を比較する．提案法に関しては，各単音の基底ベクトルをランク 1 近似により作成した場合と，近似誤差 10% ($\epsilon = 0.1\||S_{\mathrm{mono}}|^2\|_{\mathrm{F}}$) のランク r 近似により作成した場合の 2 通りを検証するため，合計で 6 つの手法の評価を行った．

音源信号には MIDI で作成したサンプリング周波数 16,000 Hz のピアノ音 $s_1[t]$ とベース音 $s_2[t]$ を用いた．インパルス応答には RWCP データベース [11] に収録されている E2A (残響時間 0.9 秒) を用いた．ここで，ピアノとベースはそれぞれ 130° と 50° の位置に存在するとし，2 つのマイクは 17 番と 30 番とした．図 1 から図 4 に，実際のインパルス応答 $\boldsymbol{h}_{m,n}$ ($m = 1, 2$; $n = 1, 2$) を示す．これらの図から，インパルス応答のスパース性及び振幅の単調減少傾向が確認できる．提案法では，音源信号 $s_1[t], s_2[t]$ に含まれている全ての音程を 1 音ずつ第 3.1 節の方法で事前に学習した．

短時間フーリエ変換では，フレーム長 $L = 8192$，フレームシフト量 $\eta = 2048$ とし，ハミング窓 $\psi[\tau] := 0.54 - 0.46\cos(2\pi\tau/L)$ を用いた．従来法の基底ベクトル数は $K_1 = K_2 = 30$ とした．ILRMA+sparse のパラメータは $\kappa[\tau] = -\log_{10}(1 - \exp(-432/(\tau+1)))$，$T = 4096$，$\lambda = 0.075$ とし，提案法 2 では $\kappa[\tau]$ と T は共通で，$\lambda = 0.09$ とした．全ての手法で分離行列の初期値を単位行列 $W_i^{(0)} = E_2$ ($i = 1, 2, \ldots, I$) とし，ILRMA+sparse と提案法 2 でインパルス応答の初期値を $\boldsymbol{h}_{m,n}^{(0)} = \boldsymbol{0}$ ($m = 1, 2$; $n = 1, 2$) とした．従来法における基底行列と係数行列の初期値 $B_n^{(0)}$ と $C_n^{(0)}$ は $[0, 1]$ の一様分布からランダムに生成させた．提案法でも同様に $C_n^{(0)}$ の初期値をランダムに生成させた．各手法で変数の更新を 100 回繰り返したときの分離性能を評価した．評価指標には文献 [12] の source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), sources-to-artifacts ratio (SAR) を用い，$B_n^{(0)}$ と $C_n^{(0)}$ がランダムのため 10 回の平均値を比較した．

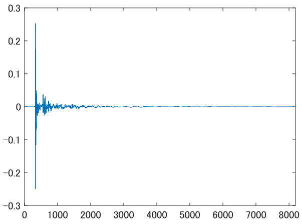

図 1 $\boldsymbol{h}_{1,1}$

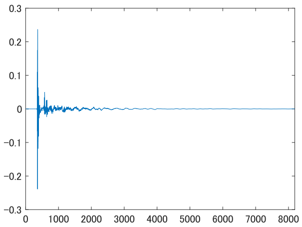

図 2 $\boldsymbol{h}_{1,2}$

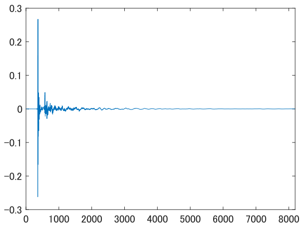

図 3 $\boldsymbol{h}_{2,1}$

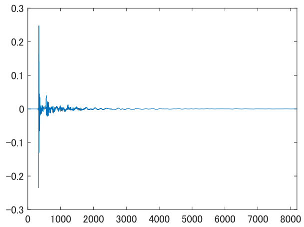

図 4 $\boldsymbol{h}_{2,2}$

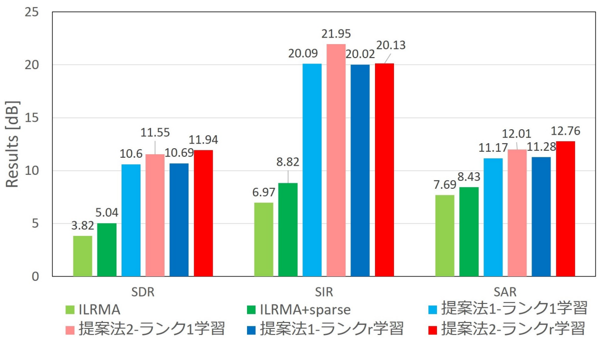

図 5 従来法と提案法によるピアノ音の分離結果

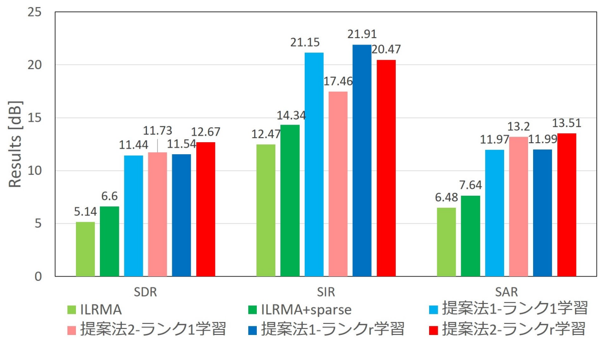

図 6 従来法と提案法によるベース音の分離結果

4.2 実験結果

ピアノ音の分離結果を図 5 に，ベース音の分離結果を図 6 に示す．これらの図から，提案法は教師情報を用いることで，ILRMA や ILRMA+sparse の性能を全ての指標で大きく向上させていることが分かる．実際に分離結果を聴いてみると，従来法ではピアノ音とベース音が依然として混合している印象を受けるのに対して，提案法ではいずれの手法もピアノ音とベース音を概ね分離できていると感じる．

総合的な分離性能を示す SDR の比較では，ピアノ音とベース音の両方に関して，ランク r 近似による学習を行った場合の提案法 2 が最も高い値を示した．また，従来法と提案法の両方に関して，インパルス応答のスパース性を用いることで SDR 値が向上しており，式 (6) の問題の有効性が確認できた．分離結果を聴くことでも，インパルス応答を考慮しない場合と比較して，音が明瞭となることを実感できる．提案法で使用する学習方法の比較では，ランク 1 近似による学習と比べて，ランク r 近似による学習のほうが僅かに高い SDR 値を示した．しかし，実際に分離結果を聴くと，ほぼ同じであるという印象を受けた．

非目的音の除去性能を示す SIR に関しては，ピアノ音では提案法 2 が最も高い値を示したが，ベース音では提案法 1 が最も高い値を示した．歪みの少なさを示す SAR に関しては SDR と同様の傾向が見られた．

5 おわりに

本論文では，教師情報を用いて，ILRMA における基底行列を事前に学習することを提案した．各基底ベクトルは，各単音のスペクトログラムを低ランク近似することで作成された．更に，インパルス応答のスパース性を用いたアルゴリズムも提案し，提案法の有効性を数値実験で示した．今後は，教師情報と実楽器の音色の違いも考慮したモデル [8] を導入する．

参考文献

[1] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[2] T. Kim *et al.*, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2007.

[3] D. Kitamura *et al.*, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[4] 北原 大地, 小田 亮太, 平林 晃, “混合過程推定にスパース性を利用したブラインド音源分離,” 第 62 回システム制御情報学会研究発表講演会, 2018, 7 pages.

[5] D. D. Lee and H. S. Seung, “Algorithms for nonnegative matrix factorization,” in *Proc. NIPS*, 2001, pp. 535–541.

[6] P. Smaragdis *et al.*, “Supervised and semi-supervised separation of sounds from single-channel mixtures,” in *Proc. ICA*, 2007, pp. 414–421.

[7] D. Kitamura *et al.*, “Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing,” in *Proc. ISSPIT*, 2013, pp. 392–397.

[8] D. Kitamura *et al.*, “Music signal separation by supervised nonnegative matrix factorization with basis deformation,” in *Proc. DSP*, 2013, 6 pages.

[9] S. Mogami *et al.*, “Independent deeply learned matrix analysis for multichannel audio source separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 15 pages, 2019, to appear.

[10] T. Fujiwara *et al.*, “Reduced-rank modeling of time-varying spectral patterns for supervised source separation,” in *Proc. ICASSP*, 2015, pp. 3307–3311.

[11] S. Nakamura *et al.*, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” in *Proc. LREC*, 2000, pp. 965–968.

[12] E. Vincent *et al.*, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.