Vol.2019-MUS-123 No.29
Vol.2019-SLP-127 No.29
2019/6/22

深層音声生成モデルと同時対角化可能な空間相関行列に基づく高速マルチチャネル音声強調

關口 航平[1,2,a)]　Aditya Arie Nugraha[1]　坂東 宜昭[3]　吉井 和佳[1,2]

概要：本稿では，フルランク空間相関行列に基づくマルチチャネル音源分離を高速に実行するための，収束保証付きの汎用的なアルゴリズムについて述べる．代表的な音源分離法であるマルチチャネル非負値行列因子分解 (MNMF) では，各音源スペクトログラムのパワースペクトル密度が低ランク構造を持つと仮定している．音声スペクトログラムに対してこの仮定は成り立たないため，最近では，音声に対しては事前学習した深層生成モデルを用い，雑音に対しては NMF に基づく低ランクモデルを用いた音声強調法が提案されている．これらの手法は，フルランクの空間相関行列を直接取り扱う上で計算量が大きく，実用上の課題となっていた．本稿では，各周波数において，各音源に対応する空間相関行列が同時対角化可能であるという制約のもとでは，観測スペクトログラムを線形変換することで，各チャネルを独立化でき，共分散行列演算が回避できることを示す．具体的には，独立ベクトル分析 (IVA) で提案された反復射影法 (IP) を用いた変換行列の推定と, 変換後の空間での非負値テンソル分解 (NTF) との反復を行うことで, 収束保証付きの最適化アルゴリズムを導出できる．提案する同時対角化可能フルランク空間モデルは，独立低ランク行列分析 (ILRMA) で用いられるランク 1 空間モデルと深い関係がある．実験では，ILRMA と同等の計算量に削減しつつ, 初期値依存性が小さく, より高精度な音声強調ができることを確認した．

1. はじめに

マルチチャネル音声強調は雑音存在下での頑健な音声認識を行うための重要な技術である．マルチチャネル音声強調法の一つとして，対象話者と雑音の空間相関行列から計算されるビームフォーマやウィナーフィルタを用いた手法が広く用いられている．これらの空間相関行列を観測音から求めるために，観測音の各時間周波数ビンを音声と雑音とに分類する深層ニューラルネットワーク (DNN) が用いられるが，学習データと異なる未知の環境においては動作が不安定になる問題がある．本研究では，雑音環境を仮定せず環境に適応可能なブラインド音源分離法と，その拡張である半教師有り音声強調法に着目する．

ブラインド音源分離法では，観測音のみから音の混合過程と音源信号を推定することが目的である．このような不良設定問題を解くための方法として，音の伝達過程を表す空間モデルと各音源のパワースペクトル密度を表す音源モデルを統合した確率モデルに基づく統計的アプローチがある．[1] では，音源スペクトルが複素ガウス分布に従うことを仮定した音源モデルとフルランクの空間モデルを統合したフルランク空間共分散分析 (FCA) と呼ばれる手法が提案されている．FCA の持つ周波数間のパーミュテーション問題を緩和するため，非負値行列因子分解 (NMF) に基づく低ランクな音源モデルをフルランク空間モデルと統合したマルチチャネル NMF (MNMF) が提案されている [2–4]．NMF に基づく音源モデルは音声に対しては不適であるため，最近では，クリーンな音声のみを用いて学習した DNN に基づく音声モデル (Deep Prior, DP) と，NMF に基づく雑音モデルを統合した MNMF-DP と呼ばれる手法が提案されている [5–7]．

これらの手法はフルランクの空間相関行列を用いており，逆行列計算に伴う計算量の多さが深刻な問題となっていた．[8] では，MNMF の空間相関行列をランク 1 に制限することで，MNMF の初期値依存性と計算量の問題を改善しているが，空間モデルの自由度が大きく制限されていた．これらの問題を解決するため，[9,10] では，フルランクの空間相関行列を同時対角化可能な行列に制限することで，FCA や MNMF を高速化した手法が提案されている．しかし，これらの手法では，不動点反復法を用いており，収束の理論的な保証がなされていなかった．

[1] 理化学研究所 AIP
Tokyo 103-0027, Japan
[2] 京都大学　大学院情報学研究科
Kyoto 606-8501, Japan,
[3] 産業技術総合研究所
Tokyo, 135-0064, Japan,
[a)] kouhei.sekiguchi@riken.jp

Vol.2019-MUS-123 No.29
Vol.2019-SLP-127 No.29
2019/6/22

本稿では，同時対角化可能なフルランク空間相関行列の収束保証付き推定法と，本手法を MNMF，MNMF-DP に適応した FastMNMF，FastMNMF-DP を提案する．通常，観測音の各チャネル間には相関があるが，特別な行列で観測スペクトログラムを変換することで，チャネル間を独立にすることができる．これにより，元の複素スペクトログラムに対する MNMF は，変換後の空間でのパワースペクトログラムに対する非負値テンソル分解 (NTF) と等価になる．この変換行列の推定には，IVA や ILRMA で分離行列を推定する際に用いられる収束保証付きの反復射影法 (IP) [11] を用いる．提案法は NTF と IP を反復する手法であり，NMF と IP を反復する ILRMA と類似している．実際には，FastMNMF は MNMF と ILRMA の中間的なモデルになっており，それらとの関係性についても述べる．

2. 関連研究

本章では，本研究の基礎となるランク 1 空間モデルとフルランク空間モデルについて述べる．また，ランク 1 空間モデルに NMF に基づく音源モデルを導入した独立低ランク行列分析 (ILRMA) [8]，フルランク空間モデルに NMF に基づく音源モデルを導入したマルチチャネル非負値行列因子分解 (MNMF) [2–4]，フルランク空間モデルに深層生成モデルに基づく音声モデルと NMF に基づく雑音モデルを導入した MNMF-DP [6] について述べる．

2.1 ランク 1 空間モデルとフルランク空間モデル

音源からマイクロホンまでの音の伝搬過程を表現する空間モデルとして，ランク 1 空間モデルとフルランク空間モデルがある．ある音源 n の音源信号の複素スペクトログラムの周波数ビン f，時間フレーム t を $s_{ftn} \in \mathbb{C}$ とし，s_{ftn} が以下の複素ガウス分布に従うと仮定する．

$$s_{ftn} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{ftn}\right) \tag{1}$$

ここで，$\mathcal{N}_{\mathbb{C}}(\mu, \sigma)$ は平均 μ，分散 σ を持つ複素ガウス分布，λ_{ftn} は音源 n の周波数ビン f，時間フレーム t でのパワースペクトル密度を表す．線形時不変システムを仮定し，音源 n のみが存在すると仮定すると，s_{ftn} と M チャネルマイクロホンアレイでの音源 n の観測音 $\mathbf{x}_{ft}^{(n)} \in \mathbb{C}^M$ の関係は次式で与えられる．

$$\mathbf{x}_{ft}^{(n)} = \mathbf{a}_{nf} s_{ftn} \tag{2}$$

ここで，$\mathbf{a}_{nf} \in \mathbb{C}^M$ は音源 n の周波数ビン f におけるステアリングベクトルを表す．式 (1) と式 (2) より，$\mathbf{x}_{ft}^{(n)}$ は以下の複素ガウス分布に従う．

$$\mathbf{x}_{ft}^{(n)} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_M, \lambda_{ftn}\mathbf{\Phi}_{nf}\right) \tag{3}$$

ここで，$\mathbf{\Phi}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathrm{H}}$ は音源 n の周波数 f における空間相関行列を表す．音源が N 個存在するときのマイクロホンアレイでの観測音の複素スペクトログラム $\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{x}_{ft}^{(n)} \in \mathbb{C}^M$ は，混合行列 $\mathbf{A}_f = [\mathbf{a}_{1f}, \cdots, \mathbf{a}_{Nf}] \in \mathbb{C}^{M\times N}$ を用いて次式で与えられる．

$$\mathbf{x}_{ft} = \mathbf{A}_f \mathbf{s}_{ft} \tag{4}$$

ここで，$\mathbf{s}_{ft} = [s_{ft1}, \cdots, s_{ftN}]^{\mathrm{T}} \in \mathbb{C}^N$ である．また，$\mathbf{x}_{ft}$ はガウス分布の再生性より以下の分布に従う．

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_M, \sum_{n=1}^{N} \lambda_{ftn}\mathbf{\Phi}_{nf}\right) \tag{5}$$

式 (5) で表される空間モデルをランク 1 空間モデルと呼ぶ．

フルランク空間モデルでは，音源の移動や短時間フーリエ変換の窓長を超える長さの残響の影響などにより各空間相関行列がフルランクであると仮定する．このフルランクの空間相関行列をランク 1 の空間相関行列 $\mathbf{\Phi}_{nf}$ と区別して $\mathbf{G}_{nf}$ と表すことにする．

2.2 独立低ランク行列分析 (ILRMA)

ILRMA はランク 1 空間モデルに対し，NMF に基づく音源モデルを導入したモデルである．$\mathbf{\Phi}_{nf}$ はランク 1 の行列であるが，音源数とマイク数が等しいという制約 ($M = N$) を導入することで式 (5) の分散共分散行列をフルランクにしている．NMF に基づく音源モデルでは，各音源スペクトログラムが低ランク構造を持つと仮定し，音源 n のパワースペクトル密度 λ_{ftn} は次式で与えられる．

$$\lambda_{ftn} = \sum_{k=1}^{K} w_{nkf} h_{nkt} \tag{6}$$

ここで，w_{nkf} は音源 n の周波数ビン f での k 番目の基底，h_{nkt} は音源 n の時間フレーム t での基底 k のアクティベーション，K は基底数を表す．式 (5), (6) より，観測音の対数尤度関数は次式で与えられる．

$$\begin{aligned}
&\log p(\mathbf{x}|\mathbf{W}, \mathbf{H}, \mathbf{A}) \\
&= \sum_{f=1}^{F}\sum_{t=1}^{T} \log \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{ft}|\mathbf{0}_M, \sum_{n=1}^{N}\sum_{k=1}^{K} w_{nkf}h_{nkt}\mathbf{\Phi}_{nf}\right) \\
&\propto -\sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{n=1}^{N}\left(\frac{|\hat{s}_{ftn}|^2}{\sum_{k=1}^{K} w_{nkf}h_{nkt}} + \log\sum_{k=1}^{K} w_{nkf}h_{nkt}\right) \\
&\quad + T\sum_{f=1}^{F} \log|\mathbf{D}_f\mathbf{D}_f^{\mathrm{H}}|
\end{aligned} \tag{7}$$

ここで，$\mathbf{D}_f = \mathbf{A}_f^{-1} = [\mathbf{d}_{1f}, \cdots, \mathbf{d}_{Nf}]^{\mathrm{H}}$ は分離行列であり，$\hat{\mathbf{s}}_{ft} = \mathbf{D}_f\mathbf{x}_{ft}$ は分離音を表す．

パラメータの推定は最尤法を用いて，各変数を順番に 1 つずつ更新する．式 (7) の第 1 項と第 2 項は $|\hat{s}_{ftn}|^2$ と $\lambda_{ftn} = \sum_{k=1}^{K} w_{nkf}h_{nkt}$ の（負の）Itakura-Saito（IS）ダイバージェンスとなっているため，$\mathbf{W}$ と $\mathbf{H}$ の更新は IS ダイ

Vol.2019-MUS-123 No.29
Vol.2019-SLP-127 No.29
2019/6/22

バージェンスに基づく NMF (IS-NMF) と同様の手法で音源ごとに行う．第 1 項と第 3 項は IVA の対数尤度関数と同一になっているため，$\mathbf{D}$ の更新は IVA の更新に用いられる，収束性が保証された反復射影法 (IP) を用いる．従って，ILRMA のパラメータの更新は NMF と IP の反復とみることができる．

パラメータを推定した後に実際の分離音 $\mathbf{x}_{ft}^{(n)}$ を得る際には，$\mathbf{D}$ の周波数間でのスケールの曖昧性を解決するため Projection Back 法を用いる．

$$\mathbf{x}_{ft}^{(n)} = \mathbf{a}_{nf} s_{ftn} = \mathbf{a}_{nf} \mathbf{d}_{nf}^{\mathrm{H}} \mathbf{x}_{ft} \tag{8}$$

2.3 多チャネル非負値行列因子分解 (MNMF)

MNMF はフルランク空間モデルに対し，NMF に基づく音源モデルを導入したモデルである．ILRMA と同様に各音源が低ランク構造を持つと仮定し，パワースペクトル密度 λ_{ftn} は式 (6) で表される．ILRMA との違いは，空間相関行列 $\mathbf{G}_{nf}$ がフルランク行列になる点である．ILRMA では 1 つの空間相関行列のパラメータ数が M であったのに対し，MNMF では $(M^2+M)/2$ であり，自由度が大きく増えている．MNMF は ILRMA をその特殊形として含むが，初期値依存性が高いため，実際上は ILRMA より精度が低くなる傾向がある [8]．

MNMF の対数尤度関数は次式で与えられる．

$$\begin{aligned}
&\log p(\mathbf{x}|\mathbf{W},\mathbf{H},\mathbf{G}) \\
&= \sum_{f=1}^{F}\sum_{t=1}^{T} \log \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{ft}|\mathbf{0}_M, \sum_{n=1}^{N}\sum_{k=1}^{K} w_{nkf}h_{nkt}\mathbf{G}_{nf}\right) \\
&\propto \sum_{f=1}^{F}\sum_{t=1}^{T}\left(-\mathrm{tr}\left(\mathbf{Y}_{ft}^{-1}\mathbf{X}_{ft}\right) - \log|\mathbf{Y}_{ft}|\right)
\end{aligned} \tag{9}$$

$$\mathbf{X}_{ft} = \mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathrm{H}} \tag{10}$$

$$\mathbf{Y}_{ft} = \sum_{n=1}^{N} \lambda_{ftn}\mathbf{G}_{nf} = \sum_{n=1}^{N}\sum_{k=1}^{K} w_{nkf}h_{nkt}\mathbf{G}_{nf} \tag{11}$$

式 (9) は，$\mathbf{X}_{ft}$ と $\mathbf{Y}_{ft}$ の（負の）Log-Det (LD) ダイバージェンスとなっているため，対数尤度の最大化は，LD ダイバージェンスの最小化と等価である．ILRMA の場合には，$\mathbf{D}$ を更新した後は，$N(=M)$ 個の $F\times T$ 非負値行列を近似する $\mathbf{W}$ と $\mathbf{H}$ を求めるのに対し，MNMF では，FT 個の $M\times M$ 複素行列 $\{\mathbf{X}_{ft}\}_{f,t=1}^{F,T}$ を近似する $\mathbf{W}$, $\mathbf{H}$, $\mathbf{G}$ を求める必要があるため計算量が多い．

パラメータ推定後はマルチチャネルウィナーフィルタ (MWF) を用いて分離音 $\mathbf{x}_{ft}^{(n)}$ を推定する．

$$\mathbf{x}_{ft}^{(n)} = \mathbb{E}[\mathbf{x}_{ft}^{(n)}|\mathbf{x}_{ft}] = \mathbf{Y}_{ftn}\mathbf{Y}_{ft}^{-1}\mathbf{x}_{ft} \tag{12}$$

ここで，$\mathbf{Y}_{ftn} = \lambda_{ftn}\mathbf{G}_{nf} = \sum_{k=1}^{K} w_{nkf}h_{nkt}\mathbf{G}_{nf}$ である．

2.4 深層生成モデルと MNMF の統合 (MNMF-DP)

MNMF や ILRMA では，全ての音源が低ランク構造を持つことを仮定していたが，音声信号などに対してこの過程は不適である．NMF に基づく音源モデルの代わりに，VAE などを用いて学習した深層生成モデルを音源モデルとして用いる手法が提案されている．MNMF-DP は，フルランクの空間モデルに対して，観測音が 1 つの音声と複数の非音声雑音を含むと仮定して，音声に対して深層生成モデルに基づく音源モデル，雑音に対して NMF に基づく音源モデルを導入したモデルである．音源 $n=1$ を音声，音源 $n\geq 2$ を雑音と仮定し，音声に対して深層生成モデルに基づく音源モデルを用いると，パワースペクトル密度は次式で与えられる．

$$\lambda_{ft1} = u_f v_t [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_t)]_f \tag{13}$$

ここで，$\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\cdot)$ はパラメータ $\boldsymbol{\theta}$ を持つ非線形関数（DNN）であり，潜在変数 $\mathbf{z}_t\in\mathbb{R}^D$ をパワースペクトル密度 $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_t)\in\mathbb{R}_+^F$ に変換する．$u_f\geq 0$ は周波数 f のスケールパラメータ，$v_t\geq 0$ は時間フレーム t のアクティベーションを表す．

MNMF-DP の対数尤度関数は式 (9) で与えられ，$\mathbf{Y}_{ft}$ は次式で与えられる．

$$\mathbf{Y}_{ft} = u_f v_t [\boldsymbol{\sigma}_{\theta}^2(\mathbf{z}_t)]_f \mathbf{G}_{1f} + \sum_{n=2}^{N}\sum_{k=1}^{K} w_{nkf}h_{nkt}\mathbf{G}_{nf} \tag{14}$$

3. 同時対角化可能な空間相関行列に基づく高速マルチチャネル音声強調

本章では，フルランク空間モデルの空間相関行列に対して同時対角化可能という制約を導入することで，MNMF，MNMF-DP を高速化した FastMNMF，FastMNMF-DP について述べる．

3.1 同時対角化可能制約付きフルランク空間モデル

N 個の空間相関行列 $\{\mathbf{G}_{nf}\}_{n=1}^{N}$ が行列 $\mathbf{Q}_f = [\mathbf{q}_{f1},\cdots,\mathbf{q}_{fM}]^{\mathrm{H}}\in\mathbb{C}^{M\times M}$ を用いて次式のように同時対角化可能であるという制約を導入する．

$$\mathbf{Q}_f\mathbf{G}_{nf}\mathbf{Q}_f^{\mathrm{H}} = \mathrm{Diag}(\tilde{\mathbf{g}}_{nf}) \tag{15}$$

ここで，$\mathrm{Diag}(\tilde{\mathbf{g}}_{\mathrm{nf}})$ は $\tilde{\mathbf{g}}_{nf} = [\tilde{g}_{nf1},\cdots,\tilde{g}_{nfM}]\in\mathbb{R}_+^M$ を対角成分に持つ対角行列である．$\mathbf{Q}_f\mathbf{G}_{nf}\mathbf{Q}_f^{\mathrm{H}}$ は $\mathbf{Q}_f\mathbf{x}_{ft}$ の分散共分散行列であるため，$\mathbf{x}_{ft}$ を $\mathbf{Q}_f$ で変換した空間では各チャネルが独立になることを意味する．

3.2 FastMNMF-DP

FastMNMF-DP は，同時対角化可能制約付きフルランク空間モデルに対して，観測音が 1 つの音声と複数の非音声雑音を含むと仮定して，音声に対して深層生成モデルに基づく音源モデル (式 (13))，雑音に対して NMF に基づく音源モデル (式 (6)) を導入したモデルである．音源 $n=1$ を音声，音源 $n\geq 2$ を雑音と仮定すると，式 (6), (13), (15)

Vol.2019-MUS-123 No.29
Vol.2019-SLP-127 No.29
2019/6/22

より，対数尤度関数は次式で与えられる．

$$\begin{aligned}
&\log p(\mathbf{X}|\mathbf{Q},\tilde{\mathbf{G}},\mathbf{W},\mathbf{H},\mathbf{Z},\mathbf{U},\mathbf{V}) \\
&= \sum_{f}^{F}\sum_{t=1}^{T}\log\mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{ft}\middle|\mathbf{0},\sum_{n=1}^{N}\lambda_{ftn}\mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_{nf})\mathbf{Q}_f^{-\mathrm{H}}\right) \\
&\propto \sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{m=1}^{M}\left(-\frac{\tilde{x}_{ftm}}{\tilde{y}_{ftm}}-\log\tilde{y}_{ftm}\right)+T\sum_{f=1}^{F}\log\left|\mathbf{Q}_f\mathbf{Q}_f^{\mathrm{H}}\right| \qquad (16)
\end{aligned}$$

$$\tilde{x}_{ftm} = [|\mathbf{Q}_f\mathbf{x}_{ft}|^{\circ 2}]_m \qquad (17)$$

$$\begin{aligned}
\tilde{y}_{ftm} &= \sum_{n=1}^{N}\lambda_{ftn}\tilde{g}_{nfm} \\
&= u_f v_t[\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t)]_f\tilde{g}_{1fm}+\sum_{n=2}^{N}\sum_{k=1}^{K}w_{nkf}h_{nkt}\tilde{g}_{nfm} \qquad (18)
\end{aligned}$$

ここで，$|\cdot|^{\circ 2}$ は要素ごとの絶対値の 2 乗を表し，$\tilde{x}_{ftm}$ は $\mathbf{x}_{ft}$ を $\mathbf{Q}_f$ で変換した後のチャネル m のパワーを表す．音声に対しても NMF に基づく音源モデルを用いることで，MNMF を高速化した FastMNMF を導くことができる．

パラメータ推定後は MWF を用いて分離音 $\mathbf{x}_{ft}^{(n)}$ を推定する．式 (15) を用いることで，MWF は次式で表される．

$$\mathbf{x}_{ft}^{(n)} = \mathbf{Q}_f^{-1}\mathrm{Diag}\left(\frac{\lambda_{ftn}\tilde{\mathbf{g}}_{nf}}{\sum_{n=1}^{N}\lambda_{ftn}\tilde{\mathbf{g}}_{nf}}\right)\mathbf{Q}_f\mathbf{x}_{ft} \qquad (19)$$

これは，観測音を $\mathbf{Q}_f$ で変換した空間で分離を行い，$\mathbf{Q}_f^{-1}$ で元の空間に戻す操作であり，式 (12) よりも高速である．

3.3 推論

最尤推定を用いてパラメータ $\mathbf{Q}$, $\tilde{\mathbf{G}}$, $\mathbf{W}$, $\mathbf{H}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{Z}$ を推定する方法について述べる．

3.3.1 空間モデルの更新

式 (16) の第 1 項と第 3 項は，ILRMA と同様に，$\mathbf{Q}$ について IVA の対数尤度関数と同一になっているため，ILRMA と同様に収束保証付き IP 法を用いて次式で更新する．

$$\mathbf{V}_{fm} \overset{\mathrm{def}}{=} \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_{ft}\tilde{y}_{ftm}^{-1}, \qquad (20)$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{Q}_f\mathbf{V}_{fm})^{-1}\mathbf{e}_m, \qquad (21)$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{V}_{fm}\mathbf{q}_{fm})^{-\frac{1}{2}}\mathbf{q}_{fm} \qquad (22)$$

ここで，$\mathbf{e}_m$ は one-hot ベクトルを表す．

式 (16) の第 1 項と第 2 項は $\tilde{x}_{ftm}$ と $\tilde{y}_{ftm}$ の IS ダイバージェンスとなっている．ILRMA の場合には，各 n ごとに $F\times T$ の行列に対し IS-NMF を行ったが，FastMNMF と FastMNMF-DP では $F\times T\times M$ のテンソルを λ_{ftn}, $\tilde{g}_{nfm}$ で分解する IS-非負値テンソル分解 (IS-NTF) を行う．IS-NTF では，IS-NMF の場合と同様に，対数尤度関数の下限を最大化する Majorization Minimization (MM) 法を用いる．対数尤度関数の下限は Jensen の不等式と Taylor の一次近似を用いて次式で与えられる．

$$\begin{aligned}
&\log p(\mathbf{X}|\mathbf{Q},\tilde{\mathbf{G}},\mathbf{W},\mathbf{H},\mathbf{Z},\mathbf{U},\mathbf{V}) \\
&\geq -\sum_{f,t,m=1}^{F,T,M}\left(\frac{\tilde{x}_{ftm}\phi_{ftm1}^2}{u_f v_t[\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t)]_f\tilde{g}_{1fm}}+\sum_{n=2}^{N}\sum_{k=1}^{K}\frac{\tilde{x}_{ftm}\phi_{ftmnk}^2}{w_{nkf}h_{nkt}\tilde{g}_{nfm}}\right) \\
&\quad -\sum_{f,t,m=1}^{F,T,M}\left(\log\psi_{ftm}+\frac{\sum_{n=1}^{N}\lambda_{ftn}\tilde{g}_{nfm}}{\psi_{ftm}}\right) \\
&\quad +T\sum_{f=1}^{F}\log|\mathbf{Q}_f\mathbf{Q}_f^{\mathrm{H}}|^2+\mathrm{const} \qquad (23) \\
&\overset{\mathrm{def}}{=} \mathcal{L} \qquad (24)
\end{aligned}$$

ここで，$\phi_{ftm1}+\sum_{n=2}^{N}\sum_{k=1}^{K}\phi_{ftmnk}=1$ であり，等号は以下の時に成り立つ．

$$\phi_{ftm1} = \frac{u_f v_t[\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t)]_f\tilde{g}_{1fm}}{\tilde{y}_{ftm}} \qquad (25)$$

$$\phi_{ftmnk} = \frac{w_{nkf}h_{nkt}\tilde{g}_{nfm}}{\tilde{y}_{ftm}} \qquad (26)$$

$$\psi_{ftm} = \tilde{y}_{ftm} \qquad (27)$$

下限 $\mathcal{L}$ の $\tilde{g}_{nfm}$ についての偏微分が 0 となる $\tilde{g}_{nfm}$ を求め，式 (25), (26), (27) を代入することで，乗法更新式は以下で与えられる．

$$\tilde{g}_{nfm} \leftarrow \tilde{g}_{nfm}\sqrt{\frac{\sum_{t=1}^{T}\lambda_{ftn}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{t=1}^{T}\lambda_{ftn}\tilde{y}_{ftm}^{-1}}} \qquad (28)$$

3.3.2 雑音モデルの更新

$\tilde{g}_{nfm}$ の場合と同様にして，w_{nkf} と h_{nkt} の乗法更新式は次式で与えられる．

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{\sum_{t,m=1}^{T,M}h_{nkt}\tilde{g}_{nfm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{t,m=1}^{T,M}h_{nkt}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}}} \qquad (29)$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{\sum_{f,m=1}^{F,M}w_{nkf}\tilde{g}_{nfm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,m=1}^{F,M}w_{nkf}\tilde{g}_{nfm}\tilde{y}_{ftm}^{-1}}} \qquad (30)$$

3.3.3 音声モデルの更新

$\tilde{g}_{nfm}$ の場合と同様にして，u_f と v_t の乗法更新式は次式で与えられる．

$$u_f \leftarrow u_f\sqrt{\frac{\sum_{t,m=1}^{T,M}v_t[\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t)]_f\tilde{g}_{1fm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{t,m=1}^{T,M}v_t[\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t)]_f\tilde{g}_{1fm}\tilde{y}_{ftm}^{-1}}} \qquad (31)$$

$$v_t \leftarrow v_t\sqrt{\frac{\sum_{f,m=1}^{F,M}u_f[\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t)]_f\tilde{g}_{1fm}\tilde{x}_{ftm}\tilde{y}_{ftm}^{-2}}{\sum_{f,m=1}^{F,M}u_f[\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t)]_f\tilde{g}_{1fm}\tilde{y}_{ftm}^{-1}}} \qquad (32)$$

$\mathbf{z}_t$ の更新には，Metropolis サンプリング法を用いる方法と，誤差逆伝播法を用いる方法の二通りがある．サンプリング法では，まず現在の推定値 $\mathbf{z}_t$ を平均に持つガウス分布 $\mathcal{N}_{\mathbb{C}}(\mathbf{z}_t,\xi\mathbf{I}_D)$ から $\mathbf{z}_t^{\mathrm{new}}$ をサンプリングする．次に $\mathbf{z}_t^{\mathrm{new}}$ の採択率 γ_t を次式で計算する．

$$\log \gamma_t = -\sum_{f,m=1}^{F,M}\left(\frac{\tilde{x}_{ftm}}{\lambda_{ft1}^{\mathrm{new}}\tilde{g}_{1fm}+\tilde{y}_{ftm}^{\neg 1}}-\frac{\tilde{x}_{ftm}}{\lambda_{ft1}\tilde{g}_{1fm}+\tilde{y}_{ftm}^{\neg 1}}\right) - \sum_{f,m=1}^{F,M}\log\frac{\lambda_{ft1}^{\mathrm{new}}\tilde{g}_{1fm}+\tilde{y}_{ftm}^{\neg 1}}{\lambda_{ft1}\tilde{g}_{1fm}+\tilde{y}_{ftm}^{\neg 1}}, \tag{33}$$

ここで，$\tilde{y}_{ftm}^{\neg 1}=\sum_{n'=2}^{N}\lambda_{ftn'}\tilde{g}_{n'fm}$，$\lambda_{ft1}^{\mathrm{new}}=u_f v_t \boldsymbol{\sigma}_\theta^2(\mathbf{z}_t^{\mathrm{new}})$ である．最後に，採択率に基づいて $\mathbf{z}_t^{\mathrm{new}}$ を採択するかどうかを決定する．誤差逆伝播法では，現在の $\mathbf{z}_t$ を用いて対数尤度関数を式 (16) で計算し，対数尤度関数を最大化するように Adam を用いて $\mathbf{z}_t$ を更新する．

3.4 ILRMA との関係

ILRMA の更新は NMF と IP 法を交互に適用するのに対して，FastMNMF と FastMNMF-DP は NTF と IP 法を交互に適用しており，二つの手法は類似性があることがわかる．実際，ILRMA は FastMNMF の特殊系になっていることを示す．式 (15) は以下のように書ける．

$$\begin{aligned}\mathbf{G}_{nf} &= \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_{nf})\mathbf{Q}_f^{-\mathrm{H}} \\ &= \sum_{m=1}^{M}\tilde{g}_{nfm}\tilde{\mathbf{q}}_{fm}\tilde{\mathbf{q}}_{fm}^{\mathrm{H}}\end{aligned} \tag{34}$$

ここで，$\tilde{\mathbf{q}}_{fm}$ は $\mathbf{Q}_f^{-1}$ の m 列目を表す．これは，N 個の空間相関行列が，共通の M 個のステアリングベクトルから構成されており，各ステアリングベクトルの重みが $\tilde{g}_{nfm}$ で表されるということを意味している．一方，ランク 1 空間モデルは，各音源の空間相関行列が 1 つのステアリングベクトルで構成されているとみなすことができる．従って，同時対角化可能制約付きフルランク空間モデルにおいて，$\tilde{\mathbf{g}}_{nf}$ が n 番目の要素のみが 1 となる one-hot ベクトルのとき，$\mathbf{G}_{nf}=\tilde{\mathbf{q}}_{fn}\tilde{\mathbf{q}}_{fn}^{\mathrm{H}}$ となり，ランク 1 モデルと完全に一致する．また，FastMNMF の $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{H}$ の更新式と ILRMA の $\mathbf{D}$, $\mathbf{W}$, $\mathbf{H}$ の更新式も一致する．

4. 評価実験

4.1 実験設定

CHiME3 evaluation データセットから 100 発話をランダムに選び，音声強調の精度と実効速度を評価した．各発話はタブレットに搭載された 6 チャネルマイクロホンアレイで録音された 16 kHz の観測音であり，2 番目のマイクロホンのみタブレットの裏側に搭載されているため，このマイクロホンを除いた 5 チャネル ($M=5$) を使用した．短時間フーリエ変換（STFT）の窓幅は 1024 ($F=513$)，シフト長は 256 サンプルを使用した．音声強調の精度は Signal-to-Distortion Ratio (SDR) を用いて評価した [12, 13]．実行時間の計測は，Intel Xeon W-2145 (3.70 GHz) を搭載した計算機を使用した．

FastMNMF と FastMNMF-DP を ILRMA [8]，MNMF [3]，MNMF-DP [6]，不動点反復法 (FPI) を用いた FastMNMF-FPI [10] と比較した．ILRMA の音源数 N はマイク数（$M=5$）と等しいという制約があるため 5 とし，他の手法については 2 から 5 を使用した．基底数は 4, 16, 64 を使用した．MNMF-DP と FastMNMF-DP では，潜在変数 $\mathbf{z}_t$ の次元 D は 16 とし，全体の更新 1 回につき $\mathbf{Z}$ を 30 回更新した．推定法は，MNMF-DP ではサンプリング法のみ，FastMNMF-DP ではサンプリング法と誤差逆伝播法を用いた．DNN のパラメータは WSJ-0 コーパスに含まれる約 15 時間のクリーンな音声データを用いて，VAE により学習した．ただし，学習データと評価データの話者は異なる．DNN は [7] と同様の構造を用いた．

MNMF と MNMF-DP の音声の空間相関行列 $\mathbf{G}_{1f}$ は，観測音の空間相関行列で初期化し，雑音の空間相関行列 $\mathbf{G}_{(n\geq 2)f}$ は単位行列で初期化した．FastMNMF と FastMNMF-DP については，観測音の空間相関行列に対して固有値分解を行い，$\tilde{\mathbf{g}}_{1f}$ はその固有値，$\mathbf{Q}_f^{\mathrm{H}}$ は固有ベクトル，$\tilde{\mathbf{g}}_{(n\geq 2)f}$ は $\mathbf{1}_M$ で初期化した．ILRMA については，観測音の空間相関行列の最大固有値に対応する固有ベクトルを音声のステアリングベクトルとみなし，$\mathbf{A}_f=\mathbf{I}_M$ の 1 列目に代入し，逆行列を計算することで分離行列 $\mathbf{D}_f$ を初期化した．

4.2 実験結果

表 1 に CPU を使った場合の実行時間，表 2 に 100 発話の平均の SDR を示す．FastMNMF を MNMF と比較すると，実行速度は大幅に短くなり，SDR はすべての場合において向上している．2 つの半正定値行列は厳密に同時対角化可能であるため，音源数が 2 の場合，空間モデルの自由度は同じになる．分離精度が大きく向上している要因として初期化の違いが考えられる．MNMF では音声の空間相関行列のみを初期化しているのに対し，FastMNMF では観測の空間相関行列の固有値分解で $\mathbf{Q}_f$ を初期化している．3.4 節で述べたように，$\mathbf{Q}_f^{-1}$ は各音源のステアリングベクトルを並べた行列になることが理想である．観測の空間相関行列の固有ベクトルは各音源のステアリングベクトルと近い値になるため，この初期化が適切であり，FastMNMF の精度が向上したと考えられる．FastMNMF と FastMNMF-FPI を比較すると，実行速度，SDR ともにほとんど変わらないが，わずかに提案法が上回った．MNMF-DP と FastMNMF-DP(sampling) を比較すると，音源数が 2 の場合には分離精度が同程度となった．これは MNMF-DP は初期値依存性が低いためだと考えられる．音源数が 3 以上の場合には FastMNMF-DP が上回っており，同時対角化制約により空間モデルの自由度が適切に制限されているためだと考えられる．

FastMNMF-DP においてサンプリング法と誤差逆伝播法を比較すると，サンプリング法の方が実行時間が短く，

表 1: 5 チャネル 8 秒のデータを CPU で処理する際の 1 回の更新あたりの実行時間 [秒].

Method		ILRMA [8]			MNMF [3] / FastMNMF-FPI [10] / FastMNMF			MNMF-DP (Sampling) [6] / FastMNMF-DP (Sampling / Backprop)		
Number of bases K		4	16	64	4	16	64	4	16	64
Number of sources N	2	-	-	-	5.1/ 0.67 /0.66	5.2/ 0.74/ 0.74	5.5/ 1.2/ 1.2	12 / 1.9 / 3.1	12 / 1.9 / 3.1	12 / 2.1 / 3.2
	3	-	-	-	6.1/ 0.71/ 0.71	6.2/ 0.84/ 0.83	6.8/ 1.7/ 1.7	14 / 1.9 / 3.1	14 / 2.1 / 3.2	14 / 2.6 / 3.6
	4	-	-	-	7.0/ 0.81/ 0.81	7.3/ 0.99/ 0.99	8.0/ 2.2/ 2.2	16 / 2.1 / 3.2	16 / 2.2 / 3.4	16 / 3.1 / 4.1
	5	0.51	0.61	1.0	8.1/ 0.88/ 0.87	8.3 / 1.1 / 1.1	9.2/ 2.7/ 2.7	18 / 2.2 / 3.4	18 / 2.4 / 3.6	18 / 3.7 / 4.8

表 2: 100 発話に対する SDR[dB] の平均

Method		ILRMA [8]			MNMF [3] / FastMNMF-FPI [10] / FastMNMF			MNMF-DP (Sampling) [6] / FastMNMF-DP (Sampling / Backprop)		
Number of bases K		4	16	64	4	16	64	4	16	64
Number of sources N	2	-	-	-	11.4/ 15.4/ 15.5	11.1/ 15.6/ 15.5	10.5/ 15.1/ 15.1	17.5/ 17.5/ 16.8	18.1/ 18.2/ 17.6	18.5/ 18.6/ 18.1
	3	-	-	-	12.3/ 16.0/ 16.1	12.0/ 16.4/ 16.4	11.3/ 15.9/ 15.8	18.0/ 18.3/ 17.5	18.4/ 18.6/ 18.1	18.6/ 18.8/ 18.5
	4	-	-	-	13.0/ 16.1/ 16.2	12.7/ 16.7/ 16.7	11.9/ 16.1/ 16.1	18.0/ 18.4/ 17.9	18.4/ **18.9**/ 18.4	18.4/ **18.9**/ 18.6
	5	15.1	15.1	14.9	13.2/ 16.2/ 16.4	13.1/ 16.7/ 16.8	12.4/ 16.2/ 16.3	18.2/ 18.6/ 18.0	18.2/ 18.8/ 18.4	18.1/ 18.8 /18.6

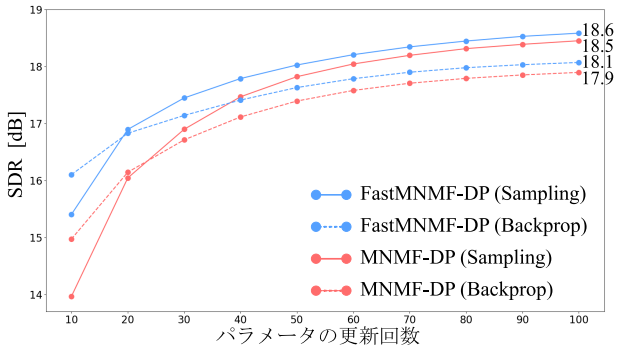

図 1: $N = 2, K = 64$ での MNMF-DP と FastMNMF-DP の比較

SDR も上回っている．図 1 に $N = 2, K = 64$ において，MNMF-DP と FastMNMF-DP にサンプリング法と誤差逆伝播法を用いて，10 回更新ごとに SDR を評価した結果を示す．どちらの手法でも最初の 10 回更新後は誤差逆伝播法が上回っていることから，計算時間に制限がある場合には誤差逆伝播法を用いることが有効であると考えられる．MNMF-DP と FastMNMF-DP を比較すると，前半は FastMNMF-DP が大きく上回っていることからも，FastMNMF(-DP) の初期化が適切であることがわかる．

5. まとめ

本稿では，同時対角化可能制約付きフルランク空間モデルの収束保証付き推定手法と，本手法を MNMF と MNMF-DP に適応し高速化した FastMNMF と FastMNMF-DP を提案した．提案法は変換行列を推定する IP 法と，変換先での NTF を反復して実行する手法である．IP 法と NMF を反復する ILRMA と類似しており，提案法との関連性について議論した．実験では，従来法と比較して，分離精度と実行速度の両方の点で提案法の性能が向上していることを確認した．また，FastMNMF-DP の音源モデルにおける潜在変数の推定方法としてサンプリング法と誤差逆伝播法を比較し，サンプリング法の有効性を示した．

謝辞:本研究の一部は基盤研究 (B) No. 19H04137 の支援を受けた．

参考文献

[1] Duong, N. Q. K. et al.: Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model, *IEEE TASLP*, Vol. 18, No. 7, pp. 1830–1840 (2010).

[2] Ozerov, A. and Févotte, C.: Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation, *IEEE TASLP*, Vol. 18, No. 3, pp. 550–563 (2010).

[3] Sawada, H. et al.: Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data, *IEEE TASLP*, Vol. 21, No. 5, pp. 971–982 (2013).

[4] Nikunen, J. and Virtanen, T.: Multichannel Audio Separation by Direction of Arrival Based Spatial Covariance Model and Non-negative Matrix Factorization, *IEEE ICASSP*, pp. 6677–6681 (2014).

[5] Bando, Y. et al.: Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-negative Matrix Factorization, *IEEE ICASSP*, pp. 716–720 (2018).

[6] Sekiguchi, K. et al.: Bayesian Multichannel Speech Enhancement with a Deep Speech Prior, *APSIPA*, pp. 1233–1239 (2018).

[7] Leglaive, S., Girin, L. and Horaud, R.: Semi-supervised Multichannel Speech Enhancement with Variational Autoencoders and Non-negative Matrix Factorization, *IEEE ICASSP*, pp. 101–105 (2019).

[8] Kitamura, D. et al.: Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization, *IEEE TASLP*, Vol. 24, No. 9, pp. 1626–1641 (2016).

[9] Ito, N. et al.: FastFCA: A Joint Diagonalization Based Fast Algorithm for Audio Source Separation Using A Full-rank Spatial Covariance Model, *EUSIPCO*, pp. 1667–1671 (2018).

[10] Ito, N. and Nakatani, T.: FastMNMF: Joint Diagonalization Based Accelerated Algorithms for Multichannel Nonnegative Matrix Factorization, *ICASSP*, pp. 371–375 (2019).

[11] Ono, N.: Stable and Fast Update Rules for Independent Vector Analysis Based on Auxiliary Function Technique, *IEEE WASPAA*, pp. 189–192 (2011).

[12] Vincent, E. et al.: Performance Measurement in Blind Audio Source Separation, *IEEE TASLP*, Vol. 14, No. 4, pp. 1462–1469 (2006).

[13] Raffel, C. et al.: mir_eval: A Transparent Implementation of Common MIR metrics, *ISMIR*, pp. 367–372 (2014).