ランク 1 空間近似を用いた BSS における音源及び空間モデルの考察 *

○北村大地 (総研大), 猿渡洋 (東大), 小野順貴 (NII/総研大), 澤田宏 (NTT), 亀岡弘和 (東大/NTT)

1 はじめに

ブラインド音源分離 (blind source separation: BSS) とは，音源位置や混合系が未知の条件で観測された信号のみから混合前の元信号を推定する信号処理技術である．優決定条件 (音源数 ≤ 観測チャネル数) における BSS では，独立成分分析 (independent component analysis: ICA) [1] に基づく手法が主流であり，盛んに研究されてきた [2]–[8]．一方，モノラル信号等を対象とした劣決定条件 (音源数 > 観測チャネル数) 下では，非負値行列因子分解 (nonnegative matrix factorization: NMF) [9] を応用した手法が注目を集めており，多チャネル信号用に拡張した多チャネル NMF (multichannel NMF: MNMF) [10] も提案されている．BSS は一般的に，話者分離や雑音抑圧が目的であるが，音楽を対象とした音源分離の研究も増加している [11]．

優決定条件における周波数領域 ICA (frequency-domain ICA: FDICA) や ICA の多変量モデルである独立ベクトル分析 (independent vector analysis: IVA) [12]–[14] では，時間周波数領域での線形時不変混合を仮定する．この仮定は，多チャネル観測信号の空間相関行列のランクが 1 になることから，**ランク 1 空間近似**と呼ばれ，複素スペクトログラムの各時間フレーム内で複数の音源が瞬時混合されているという混合系を想定したものである．このような仮定は，各音源から各マイクロフォンまでのインパルス応答が，短時間フーリエ変換 (short-time Fourier transform: STFT) の窓関数と比べて十分に短い場合に成立する．著者らは近年，従来の MNMF にランク 1 空間近似を導入した分解モデル (ランク 1 MNMF) [15]–[18] を提案しており，優決定条件下においては，従来の MNMF に匹敵する分離性能と 20 倍程度高速な最適化アルゴリズムを実現している．

本稿では，ランク 1 空間近似を用いた 3 つの代表的な BSS アルゴリズム (IVA, FDICA, ランク 1 MNMF) を取り上げ，それぞれの手法が仮定する音源モデルと空間モデルについて考察する．さらに，人工的な音源及び混合系を用いた場合の分離精度を比較し，各手法の仮定するモデルの違いを実験的に実証することで，3 手法の中でランク 1 MNMF が最も柔軟な音源及び空間モデルであることを示す．

2 ランク 1 空間近似を用いた BSS

2.1 ランク 1 空間近似

音源数と観測チャネル数をそれぞれ N, M とし，各時間周波数における多チャネル音源信号，多チャネル観測信号，分離信号をそれぞれ

$$\boldsymbol{s}_{ij} = (s_{ij,1} \cdots s_{ij,N})^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij,1} \cdots x_{ij,M})^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij,1} \cdots y_{ij,N})^{\mathrm{T}} \tag{3}$$

と表す (要素はすべて複素数)．ここで，$i = 1, \cdots, I$ は周波数インデックス，$j = 1, \cdots, J$ は時間インデックス，$n = 1, \cdots, N$ は音源インデックス，$m = 1, \cdots, M$ はチャネルインデックスを示し，$^{\mathrm{T}}$ は転置を表す．

混合系が線形時不変であり，時間周波数領域での複素瞬時混合で表現できると仮定すると，各時間フレームにおいて周波数毎の複素混合行列 $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \cdots \boldsymbol{a}_{i,N})$ ($\boldsymbol{a}_{i,n}$ は各音源のステアリングベクトル) が定義でき，多チャネル観測信号を次式で表現できる．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{4}$$

このとき，観測信号 $\boldsymbol{x}_{ij}$ に含まれる各音源の空間相関行列のランクは必ず 1 となる．すなわち，「混合系が線形時不変かつ複素瞬時混合」という仮定は，ランク 1 空間近似と等価であり，各音源が周波数毎の時不変なステアリングベクトル $\boldsymbol{a}_{i,n}$ 1 本で表現できるという近似を与えている．

式 (4) の混合系において $\boldsymbol{A}_i$ をフルランクとすれば，分離ベクトル $\boldsymbol{w}_{i,n}$ で表現される分離行列 $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1} \cdots \boldsymbol{w}_{i,N})^{\mathrm{H}}$ が存在し，分離信号は次式となる．

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{5}$$

但し，$^{\mathrm{H}}$ はエルミート転置を示す．ランク 1 空間近似を用いた BSS では，式 (5) 中の分離行列 $\boldsymbol{W}_i$ を推定することが最終的な目標となる．

2.2 IVA の仮定する音源及び空間モデル

IVA は複数の周波数成分を同時に取り扱う為に，ICA を多変量モデルへと拡張した手法である．周波数成分間の高次相関を考慮することで，FDICA におけるパーミュテーション問題 [3]–[5] を解決しながら同時に分離行列 $\boldsymbol{W}_i$ を推定する．ICA が非ガウス性の分布を仮定するように，IVA も非ガウスな多変量分布を仮定する．このとき，変数間の高次相関を考慮する為に，球対称の多変量分布を仮定することが重要である [13]．最もよく用いられる分布は，Fig. 1 (a) の右側に示す球状ラプラス分布である．この図では，二つの周波数成分の同時分布を示しており，原点を中心に球対称となっている．この性質から，二つの変数間に高次の相関が保証される．IVA の仮定する混合系及び分離系を Fig. 2 に示す．非ガウスの球対称分布に従う多変量の周波数ベクトルを用いることで，パーミュテーション問題を解決しながら周波数毎の分離行列を求めることができる．

IVA が仮定している音源モデルは，球状多変量分布そのものと解釈できる．この音源モデルを Fig. 1 (a) の左側に示す．各音源は周波数方向に一定の分散値 (パワー) を持っており，それらが時間的に変化するようなパワースペクトログラムを仮定している．従って，複数の周波数で同時に生起する成分を同一音源としてまとめる傾向がある．さらに，音源モデルのパワースペクトログラムを行列とみたとき，1 本の基底ベクトルで表現できる．これは 1 つの音源に対して 1 本のスペクトル基底を与えた NMF と解釈することもできる．但し，ICA と同様に周波数毎のスケールが定まらない為，必ずしもフラット (周波数方向に一様) なスペクトル基底とは限らず，任意のスペクトル構造を持つ基底 1 本とそのアクティベーションから構成されるパワースペクトログラムが，IVA の仮定する音源モデルとなる．

---

* Study on source and spatial models for BSS with rank-1 spatial approximation by Daichi Kitamura (SOKENDAI), Hiroshi Saruwatari (The University of Tokyo), Nobutaka Ono (NII/SOKENDAI), Hiroshi Sawada (NTT), Hirokazu Kameoka (The University of Tokyo/NTT)

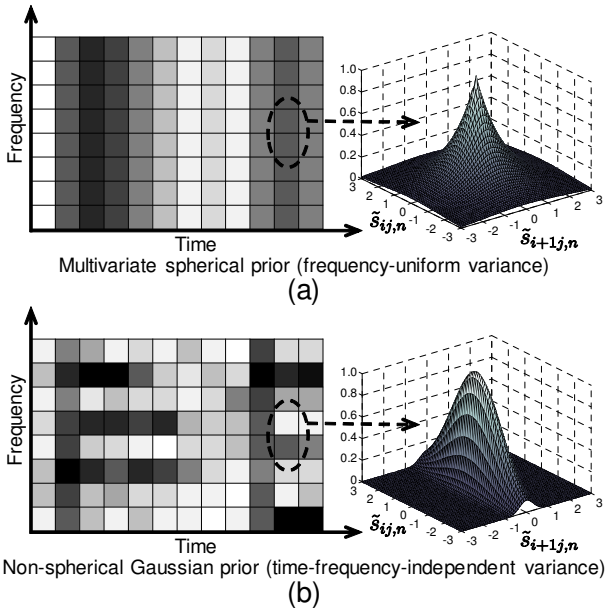

Fig. 1 Illustration of source models (model spectrograms) for one source in (a) IVA and (b) Rank-1 MNMF, where gray scale of each time-frequency slot indicates value of variance and $\tilde{s}$ denotes only real or imaginary part of complex-valued component s.

一方，IVA は空間の性質に関して具体的なモデルを与えていない．音源やマイクの位置条件に関係なく，音源モデルの統計的独立性及び多チャネルの観測信号のみから分離行列の推定を行う．

2.3 FDICA の仮定する音源及び空間モデル

時間周波数領域で，各周波数成分に独立な ICA を施す FDICA では，パーミュテーション問題の解決が極めて重要であり，これまでに多くの手法が提案されてきた．代表的なパーミュテーション問題の解決法の一つとして，周波数成分間の相関を用いる手法 [4] がある．これは，前述の IVA と本質的に等価であり，IVA が分離行列の推定と同時にパーミュテーションを解くのに対して，本手法はポスト処理としてパーミュテーションを解いている．もう一つの代表的な解決法は，音源の到来方向 (direction of arrival: DOA) の違いを活用する手法 [3] である．本手法では，推定した周波数毎の分離行列から各音源のステアリングベクトルを逆算し，位相差及び振幅比から DOA を算出して音源毎にクラスタリングすることでパーミュテーションを解いている．以後，FDICA の処理後に DOA によるパーミュテーション解決を結合した手法を FDICA+DOA と標記する．

FDICA+DOA の音源モデルは，IVA や周波数間の相関を用いたパーミュテーション解決法とは異なり，時間方向の非ガウス性制約のみである．その為，Fig. 1 のような厳密なモデルスペクトログラムは与えてられていない．一方で，FDICA で推定した DOA をパーミュテーション解決に用いる為，空間モデルに関する制約を与えている．複数の音源位置が空間的に接近した場合や残響による拡散の影響が強い場合等，音源の DOA のクラスタリングが困難な状況では，分離性能が劣化してしまう．

2.4 ランク 1 MNMF の仮定する音源及び空間モデル

従来の MNMF にランク 1 空間近似を導入したランク 1 MNMF は，優決定条件に限定した場合，MNMF と比較して高速かつ安定した音源分離性能を達成している [16]．次式はランク 1 MNMF のコスト関数を

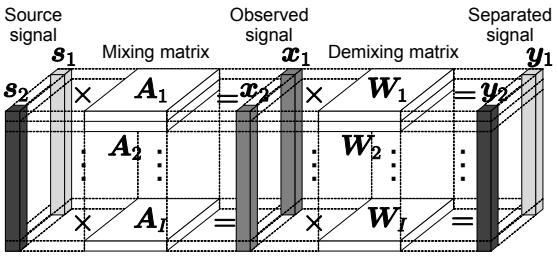

Fig. 2 Conceptual model of IVA ($N=M=2$).

示している．

$$Q = \sum_{i,j}\left[\sum_m \frac{|y_{ij,m}|^2}{\sum_l t_{il,m}v_{lj,m}} - 2\log|\det \boldsymbol{W}_i| + \sum_m \log\sum_l t_{il,m}v_{lj,m}\right] \quad (6)$$

ここで，$t_{il,m}, v_{lj,m}$ は m 番目の音源モデルに対応するスペクトル基底とアクティベーションであり，$l=1,\cdots,L$ は基底のインデックスを示す．すなわち，$\sum_l t_{il,m}v_{lj,m}$ は m 番目の音源のモデルパワースペクトログラムとなる．また，観測チャネル数と音源数の関係は $M=N$ としている．このとき，ランク 1 MNMF のコスト関数は IVA のコスト関数 (式 (6) の第一項及び第二項) と単一チャネル NMF のコスト関数 (式 (6) の第一項及び第三項) を重ね合わせた形をしている．これらの事実から，IVA はランク 1 MNMF においてスペクトル基底数が 1 の特殊ケースに相当しており，その意味でランク 1 MNMF は IVA の自然な拡張となっていると解釈できる．

ランク 1 MNMF の仮定する音源モデルを Fig. 1 (b) に示す．IVA と比較して，1 つの音源に対して L 本のスペクトル基底を用いることができる為，より複雑なパワースペクトログラムを表現可能となっている．また，各時間周波数スロットで独立な複素ガウス分布 [19] を音源モデルとして仮定しており，コスト関数 (6) は板倉齋藤擬距離の行列版である log-determinant divergence となっている．従って，時間と周波数いずれの方向にも分散が変動する分布を定義でき，より複雑な時間周波数構造を，限られた基底数で低ランク分解される音源モデルとして表現できる．

一方，空間モデルに関して，ランク 1 MNMF は，IVA と同様に具体的なモデルを与えていない．音源やマイクの位置に依存せず，観測信号と前述のモデルスペクトログラムの独立性から分離行列を推定する．

3 各手法の音源及び空間モデルに関する実験的考察

本章では，前章で考察した各手法の音源及び空間モデルの違いを実証する為に，人工的に作成した音源及び混合系を用いた実験を示す．尚，実験では簡便の為に音源数 N とチャネル数 M を 2 としている．

3.1 一定基底数の人工スペクトログラムを用いた実験

前述した IVA とランク 1 MNMF の仮定する音源モデルの違いを考慮すると，各音源のパワースペクトログラムの基底数が分離精度に影響を与えると推測できる．すなわち IVA は，1 本の基底で表現できるパワースペクトログラムを持つ音源は高精度に分離できるが，より複雑な構造を持つパワースペクトログラムに対しては，原理的に厳密なモデル化ができない為，分離精度が劣化すると考えられる．

上記の現象を実証する為に，パワースペクトログラムが任意の基底数 R で表現できる人工的な音源を生

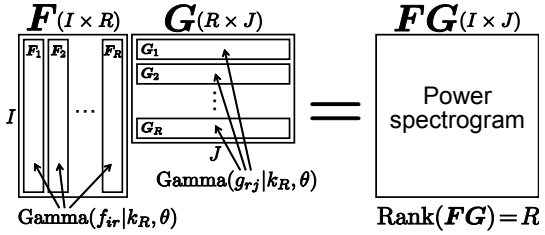

Fig. 3　Artificial source that has rank-R power spectrogram.

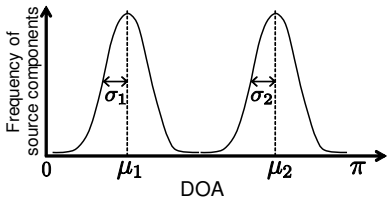

Fig. 4　Artificial DOA with Gaussian distributions.

成して分離実験を行う．生成する音源の構造を Fig. 3 に示す．非負のスパースなスペクトログラムを模擬する為，独立同一分布のガンマ分布に従う乱数 f_{ir} 及び g_{rj} を生成し，R 本の基底をもつ行列 $\boldsymbol{F}$ とそれらを生起する行列 $\boldsymbol{G}$ の積からなる行列 $\boldsymbol{FG}$ をパワースペクトログラムとする．ガンマ分布は次式で表される．

$$\mathrm{Gamma}(z|k,\theta) = z^{k-1}\frac{e^{-z/\theta}}{\Gamma(k)\theta^k} \tag{7}$$

ここで，k 及び θ はそれぞれ形状母数と尺度母数を示している．以上の手続きで生成したパワースペクトログラム $\boldsymbol{FG}$ に対し，$[0, 2\pi]$ の一様乱数に従う位相を付与することで，最終的な人工音源 (複素スペクトログラム) を生成する．従って，複素ガウス分布の分散値をガンマ分布の積に従う乱数の線形結合で模擬している．尚，人工音源のサイズは $I = J = 257$ として実験する．

上記モデルにおいては，k_R を適切な値に設定することが重要である．例えば，k_R を基底数 R にかかわらず一定値とする場合，$\boldsymbol{FG}$ の各要素は $\sum_{r=1}^{R} f_{ir}g_{rj}$ である為，R が増加すると，中心極限定理によりパワースペクトログラム $\boldsymbol{FG}$ はガウス信号に近づき，独立性に基づく手法の音源分離精度は低下する．この影響を除く為，全ての基底数の場合において $\boldsymbol{FG}$ が同一のカートシスを持つように，形状母数 k_R を R 毎に調整する．このような形状母数はモーメント-キュムラント変換を用いて求められる．厳密な証明は省略するが，次式を満たす形状母数 k_R を用いることで，一定のカートシス値 kurt を持つ $\boldsymbol{FG}$ を生成できる．

$$\frac{\zeta(k_R,R)}{\xi(k_R,R)} - \mathrm{kurt} = 0 \tag{8}$$

但し，$\zeta(k_R,R)$ と $\xi(k_R,R)$ は下記で与えられる．

$$\zeta(k_R,R) = 84k_R^3 + 174k_R^2 + 132k_R + 36 + R\left(52k_R^4 + 60k_R^3 + 19k_R^2\right) + R^2\left(12k_R^5 + 6k_R^4\right) + R^3k_R^6 \tag{9}$$

$$\xi(k_R,R) = R\left(4k_R^4 + 4k_R^3 + k_R^2\right) + R^2\left(4k_R^5 + 2k_R^4\right) + R^3k_R^6 \tag{10}$$

式 (8) を満たす k_R は解析的に求まらない為，本稿ではグリーディ探索によって求めた k_R を用い，各基底数でほぼ等しいカートシス値を持つ $\boldsymbol{FG}$ が得られることを確認している．また，本実験でのカートシス値は $\mathrm{kurt} = 500$，尺度母数は $\theta = 1$ としている．

混合系に関しては，Fig. 4 に示すガウス分布による人工的な DOA を用いる．このようなガウス分布に従うステアリングベクトルを周波数毎に生成し，混合行列 $\boldsymbol{A}_i$ を作成して式 (4) により観測信号を作成する．尚，ガウス分布の平均及び分散は $\mu_1 = 5\pi/12$，$\mu_2 = 7\pi/12$，$\sigma_1^2 = \sigma_2^2 = 0.05$，マイクの間隔は 5 cm としている．

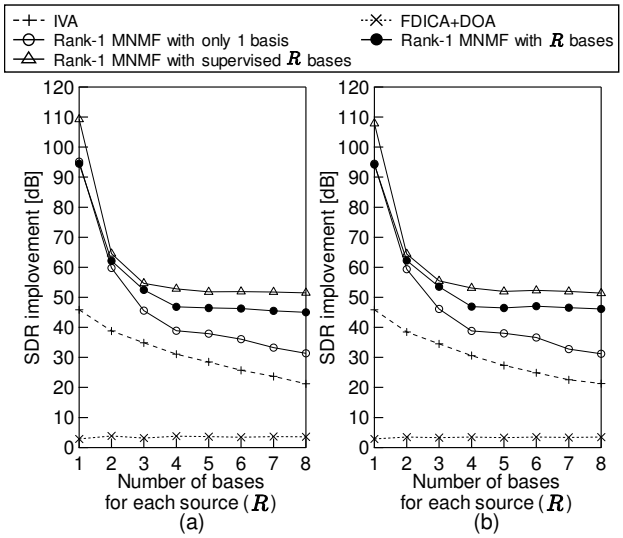

Fig. 5　Separation results of (a) source 1 and (b) source 2 with various numbers of bases.

Figure 5 は，各手法で音源分離を行った場合の SDR (signal-to-distortion ratio) [20] を，音源のパワースペクトログラムの基底数 R 毎に示した結果である．ランク 1 MNMF に関しては，音源モデルの基底数を $L = 1$ とした場合 (Rank-1 MNMF with only 1 basis)，音源モデルの基底数を $L = R$ として実際の音源の基底数と等しくした場合 (Rank-1 MNMF with R bases)，音源モデルの基底数を $L = R$ とし，各音源に対して真の基底とアクティベーションを与えた (すなわち $\boldsymbol{T} = \boldsymbol{F}$, $\boldsymbol{V} = \boldsymbol{G}$) 教師有り分離場合 (Rank-1 MNMF with supervised R bases) の 3 種類を示している．IVA と Rank-1 MNMF with only 1 basis の違いは，仮定する分布がそれぞれ球状ラプラス分布と時変ガウス分布となっている点のみである．結果を見ると，スペクトル基底数が 1 の音源モデルを持つ IVA と Rank-1 MNMF with only 1 basis では，音源のパワースペクトログラムの基底数 R の増加に伴って分離精度が低下していることが確認できる．一方，音源にとって適切な数のスペクトル基底を与える Rank-1 MNMF with R bases では，音源の基底数が増加しても高い分離精度を保っている．この事実は，IVA とランク 1 MNMF が仮定する音源モデルの違いを実証している．

3.2　人工 DOA による混合系を用いた実験

前述した FDICA+DOA とランク 1 MNMF の仮定する空間モデルの違いを考慮すると，各音源の空間的な性質が分離精度に影響を与えると推測できる．すなわち，FDICA+DOA では，音源の DOA のクラスタリングが困難な混合系において，分離性能が劣化してしまうが，IVA やランク 1 MNMF では原理的に影響を受けないと予想される．

上記の現象を実証する為に，様々な人工 DOA による混合を用いて分離実験を行う．但し，混合する音源は前節と同様の手順で生成した $\mathrm{kurt} = 500$, $R = 1$ のパワースペクトログラムを持つ信号とする．また，Fig. 4 に示すガウス分布の人工 DOA からステアリングベクトルを生成し，人工的な混合系を作成する．この時，各ガウス分布の平均や分散を変化させた場合の分離精度を実験により示す．その他の実験条件は前節と同様である．

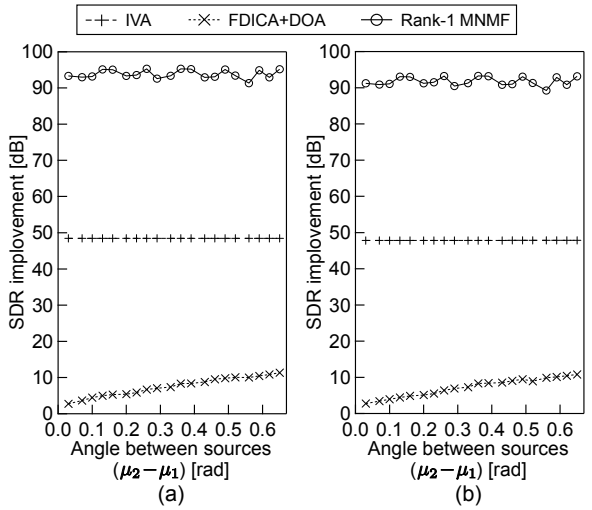

Fig. 6　Separation results of (a) source 1 and (b) source 2 with various angles.

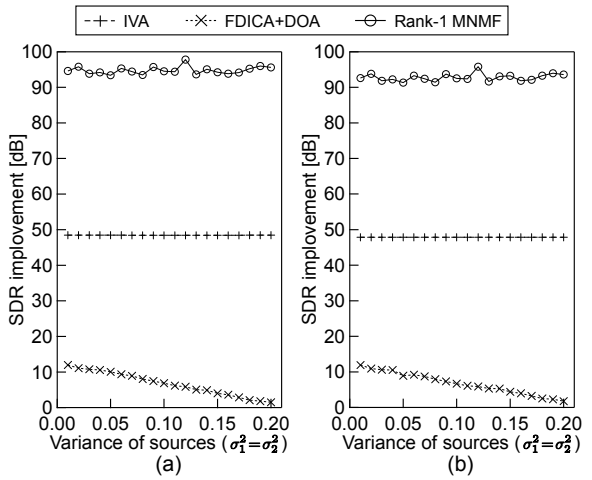

Fig. 7　Separation results of (a) source 1 and (b) source 2 with various variances.

Figure 6 は分散を $\sigma_1^2=\sigma_2^2=0.05$ に固定し，平均 μ_1 と μ_2 を変化させた場合の分離精度の変化を示している．尚，グラフの横軸は $\mu_2-\mu_1$ のラジアン値である．さらに，Fig. 7 は平均を $\mu_1=5\pi/12$，$\mu_2=7\pi/12$ に固定し，分散 σ_1^2 と σ_2^2 を同じ値で等しく変化させたときの分離精度の変化を示している．これらの結果から，FDICA+DOA は混合系に依存して分離精度が大きく劣化していることがわかる．これは，複数の音源位置が空間的に接近した場合 (Fig. 6 横軸の 0.0 付近) や残響による拡散の影響が強い場合 (Fig. 7 横軸の 0.20 付近) 等で，推定した音源の DOA のクラスタリングが困難な為，パーミュテーション問題がうまく解けていないことが原因である．一方，IVA やランク 1 MNMF では，空間モデルの具体的な制約を用いていないことから，いかなる混合系に関しても頑健な音源分離を実現していることが確認できる．

4　おわりに

本稿では，ランク 1 空間近似を用いた 3 つの BSS の音源及び空間モデルに関して考察し，それらの違いを実証する人工的な混合音源の分離実験を示した．従来の代表的な BSS である IVA 及び FDICA+DOA は，音源の性質あるいは空間の性質に起因して分離精度が劣化する問題があり，これらは両手法が仮定する音源及び空間モデルに依存している．一方，ランク 1 MNMF では，任意の基底数による NMF 分解を用いた効果的な音源モデルと，特定の制約を与えない空間モデルに基づいていることから，非常に柔軟な音源及び空間モデルであることが実験的に立証された．

謝辞　本研究の一部は JSPS 特別研究員奨励費 26·10796 の助成を受けたものである．

References

[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, pp.21–34, 1998.
[3] S. Kurita, H. Saruwatari, S. Kajita K. Takeda and F. Itakura, “Evaluation of blind signal separation method using directivity pattern under reverberant conditions,” *Proc. ICASSP*, vol.5, pp.3140–3143, 2000.
[4] 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.
[5] 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 Trans. ASLP*, vol.12, no.5, pp.530–538, 2004.
[6] H. Sawada, R. Mukai, S. Araki and S. Makino, “Convolutive blind source separation for more than two sources in the frequency domain,” *Proc. ICASSP*, pp.Ⅲ-885–Ⅲ-888, 2004.
[7] 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 Trans. ASLP*, vol.14, no.2, pp.666–678, 2006.
[8] N. Ono and S. Miyabe, “Auxiliary-function-based independent component analysis for super-Gaussian sources,” *Proc. LVA/ICA*, pp.165–172, 2010.
[9] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Proc. Advances in Neural Information Processing Systems*, vol.13, pp.556–562, 2001.
[10] H. Sawada, H. Kameoka, S. Araki and N. Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE Trans. ASLP*, vol.21, no.5, pp.971–982, 2013.
[11] 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,” *Proc. ICASSP*, pp.5365–5368, 2012.
[12] A. Hiroe, “Solution of permutation problem in frequency domain ICA using multivariate probability density functions,” *Proc. ICA*, pp.601–608, 2006.
[13] T. Kim, H. T. Attias, S.-Y. Lee and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. ASLP*, vol.15, no.1, pp.70–79, 2007.
[14] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. WASPAA*, pp.189–192, 2011.
[15] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, “Efficient multichannel nonnegative matrix factorization with rank-1 spatial model,” *Proc. Autumn Meeting of ASJ*, pp.579–582, 2014 (in Japanese).
[16] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, “Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model,” *Proc. ICASSP*, pp.276–280, 2015.
[17] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, “Relaxation of rank-1 spatial model in overdetermined BSS,” *Proc. Spring Meeting of ASJ*, pp.629–632, 2015 (in Japanese).
[18] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, “Relaxation of rank-1 spatial constraint in overdetermined blind source separation,” *Proc. EUSIPCO*, 2015 (in press).
[19] 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.
[20] E. Vincent, R. Gribonval and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol.14, no.4, pp.1462–1469, 2006.