ランク制約付き空間共分散モデル推定を用いた多チャネル補聴器システムの評価*

☆宇根昌和（筑波大学）, 久保優騎（東京大学）, 高宗典玄（東京大学）,
北村大地（香川高専）, 猿渡洋（東京大学）, 牧野昭二（筑波大学）

1 はじめに

補聴器を利用する場合，周囲の雑音の影響で目的音声の品質が劣化するため，目的音声の抽出処理が必要となる．補聴器システムにおける音声抽出処理には，目的話者の位置や空間情報が未知であっても頑健に動作するブラインド音源分離 (blind source separation: BSS) が有効である．これまで，様々な BSS 手法 [1–3] が提案されており，中でも独立低ランク行列分析 (independent low-rank matrix analysis: ILRMA) [3] は独立ベクトル分析の音源モデルに非負値行列因子分解 (nonnegative matrix factorization: NMF) [4] を導入することで，従来手法と比べて高品質な音源分離を達成している．一方，雑音が全方位から到来する状況 (拡散性雑音) を対象とした音声抽出法である，ランク制約付き空間共分散モデル推定法 (rank-constrained spatial covariance matrix (SCM) estimation) [5] が提案されている．この手法は多チャネル MNF (multichannel NMF: MNMF) [6] と同様に，各音源の空間伝達特性を表現する空間相関行列 [7] を推定する．MNMF は推定するパラメータの数が多く計算コストが大きい一方，ランク制約付き空間共分散モデル推定法は，ILRMA で推定された高精度な空間パラメータを用いることで推定すべきパラメータの数を削減し，その際に不足する空間情報を補う手法である．この手法により，拡散性雑音下のように空間相関行列がフルランクとなる場合においても MNMF より効率的かつ初期値に頑健な分離を可能にしている．しかし，ランク制約付き空間共分散モデル推定法の実環境データにおける有効性は明らかになっていない．

本研究では，両耳のマイクロホンだけでなくスマートフォンに内蔵されているマイクロホンも含めた分散マイクロホンアレー補聴器システムを新たに提案する．補聴器を装着するユーザの頭や耳の形は人によってそれぞれ異なり，またスマートフォンの位置も特定できない．BSS はこれらの不確定な要素の多い状況に対しても柔軟に処理を行うことができるが，提案する補聴器システムに対する適用例は多くない．提案するシステムの構築に際し，スマートフォンを持った人間を模したダミーヘッドを用意する．ダミーヘッドの胸部にスマートフォンを取り付け，スマートフォンと両耳を含めた複数のマイクロホンで実環境を想定した収録を行う．提案するシステムの音声抽出手法として，拡散性雑音下でも有効とされるランク制約付き共分散モデル推定法を採用する．収録にしたデータ対してランク制約付き共分散モデル推定法の音声抽出性能を評価し，実環境においても ILRMA に比べ有効な手法であることを示す．

2 定式化及び BSS 手法

2.1 定式化

N 個の音源信号を M 個のマイクロホンで収録し，観測した信号を分離することを考える．複素時間周波数成分における音源信号，観測信号，及び分離信号をそれぞれ，$\boldsymbol{s}_{ij} = (s_{ij,1}, \ldots, s_{ij,N})^\top$，$\boldsymbol{x}_{ij} = (x_{ij,1}, \ldots, x_{ij,M})^\top$，及び $\boldsymbol{y}_{ij} = (y_{ij,1}, \ldots, y_{ij,N})^\top$ とする．ここで，$i = 1, \ldots, I$，$j = 1, \ldots, J$，及び $n = 1, \ldots, N$ はそれぞれ周波数ビン，時間フレーム，及び音源信号のインデクスである．各音源が方向性の点音源であり，短時間フーリエ変換 (short-time Fourier transform: STFT) の窓長が残響時間より十分短い場合，各周波数ビンにおいて混合行列 $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \cdots \boldsymbol{a}_{i,N}) \in \mathbb{C}^{M \times N}$ が存在し，次のように書ける．

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij} \tag{1}$$

ただし，$\boldsymbol{a}_{i,n}$ は周波数 i における音源 n のステアリングベクトルである．$M = N$ かつ $\boldsymbol{A}_i$ が正則である場合，$\boldsymbol{A}_i$ の逆行列 $\boldsymbol{W}_i \in \mathbb{C}^{N \times M}$ を推定することで，次のように分離信号が得られる．

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij} \tag{2}$$

2.2 ILRMA [3]

ILRMA では，各時間周波数フレームにおける音源 n の成分が

$$s_{ij,n} \sim \mathcal{N}_c\left(0, \sum_l t_{il,n} v_{lj,n}\right) \tag{3}$$

なる単変量複素ガウス分布に従い生起する確率生成モデルを仮定する．ここで，$t_{il,n} \geq 0$，$v_{lj,n} \geq 0$ は NMF 変数であり，$l = 1, \ldots, L$ は NMF 基底のインデクス，L は NMF の基底数である．この時 $\boldsymbol{s}_{ij}$ は多変量複素ガウス分布に従い，式 (1) と多変量複素ガウス分布の再生性より，$\boldsymbol{x}_{ij}$ も多変量複素ガウス分布

$$\boldsymbol{x}_{ij} \sim \mathcal{N}_c\left(\boldsymbol{0}, \sum_n r_{ij,n} \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{H}}\right) \tag{4}$$

$$r_{ij,n} = \sum_l t_{il,n} v_{lj,n} \tag{5}$$

に従う．ここで，$r_{ij,n}$ は音源 n の音源モデルに相当し，非負実数である NMF 変数の $t_{il,n}$ と $v_{lj,n}$ を用いて音源パワーのスペクトログラムを低ランク近似したものである．また，$\boldsymbol{a}_{i,n}$ はステアリングベクトル，即ち音源 n における空間基底から構成されるランク

*Evaluation of multichannel hearing aid system using rank-constrained spatial covariance matrix estimation by Masakzu Une (The University of Tsukuba), Yuki Kubo (The University of Tokyo), Norihiro Takamune (University of Tokyo), Daichi Kitamura (National Institute of Technology, Kagawa College), Hiroshi Saruwatari (The University of Tokyo), Shoji Makino (The University of Tsukuba).

1 空間相関行列であり，音源 n の空間モデルに相当する．NMF 変数 $t_{il,n}$，$v_{lj,n}$ 及び分離行列 $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} = (\boldsymbol{w}_{i,1} \cdots \boldsymbol{w}_{i,N})^{\mathrm{H}}$ は音源間の統計的独立性最大化基準に則り，最尤推定を用いて推定される．

2.3 ランク制約付き空間共分散モデル推定法 [5]

ランク制約付き空間共分散モデル推定法は 1 個の方向性目的音源と拡散性雑音が混合している状況を対象とした手法である．方針として，ILRMA によって得られた各音源の空間基底，及びそれによって構成される M 個のランク 1 空間相関行列を用い，拡散性雑音のフルランク空間相関行列を推定する．まず，ILRMA を観測信号 $\boldsymbol{x}_{ij}$ に適用し，1 個の目的音と雑音が混ざった信号と $M-1$ 個の雑音のみの信号を得る (この現象の詳細は [8] 参照)．次に，得られた信号から空間相関行列を推定するが，ILRMA で得られる拡散性雑音の空間相関行列は目的音源の分だけランクが不足する．これを補うよう確率的定式化を行い，パラメータを推定する．最後に多チャネルウィーナーフィルタを構成し，目的音源方向の拡散性雑音を低減する．アルゴリズムの概要を以下に示す．

ランク制約付き空間共分散モデル推定法のモデルは観測信号 $\boldsymbol{x}_{ij}$ を目的音源のソースイメージ $\boldsymbol{h}_{ij} = (h_{ij,1}, \ldots, h_{ij,M})^{\top}$ と拡散性音源のソースイメージ $\boldsymbol{u}_{ij} = (u_{ij,1}, \ldots, u_{ij,M})^{\top}$ の和として次のように表す．

$$\boldsymbol{x}_{ij} = \boldsymbol{h}_{ij} + \boldsymbol{u}_{ij} \tag{6}$$

目的音源のソースイメージ $\boldsymbol{h}_{ij}$ は，ILRMA によって得られた空間基底 $\boldsymbol{a}_{i,1}, \ldots, \boldsymbol{a}_{i,N}$ のうち目的音源に対応するベクトル $\boldsymbol{a}_i^{(h)} =: \boldsymbol{a}_{i,n_h}$ と，目的音源のドライソース $s_{ij}^{(h)}$ を用いて次のように表す．

$$\boldsymbol{h}_{ij} = \boldsymbol{a}_i^{(h)} s_{ij}^{(h)} \tag{7}$$

$$s_{ij}^{(h)} \sim \mathcal{N}_c\left(0, r_{ij}^{(h)}\right) \tag{8}$$

ここで，n_h は目的音源に対応する音源インデクス，$r_{ij}^{(h)}$ は目的音源の分散 (パワースペクトログラム) である．

目的音源の分散 $r_{ij}^{(h)}$ はスパース性を有するとし，事前分布として逆ガンマ分布を仮定する．

$$p(r_{ij}^{(h)}; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \left(r_{ij}^{(h)}\right)^{-\alpha-1} \exp\left(-\frac{\beta}{r_{ij}^{(h)}}\right) \tag{9}$$

ここで，$\alpha > 0$ は形状母数，$\beta > 0$ は尺度母数，$\Gamma(\cdot)$ はガンマ関数を表す．一方，拡散性音源のソースイメージ $\boldsymbol{u}_{ij}$ は目的音源のソースイメージ $\boldsymbol{h}_{ij}$ とは独立な多変量複素ガウス分布に従うと仮定する．

$$\boldsymbol{u}_{ij} \sim \mathcal{N}_c\left(\boldsymbol{0}, r_{ij}^{(u)} \boldsymbol{R}_i^{(u)}\right) \tag{10}$$

ここで，$r_{ij}^{(u)}$ と $\boldsymbol{R}_i^{(u)}$ はそれぞれ拡散性音源の分散と空間相関行列である．ILRMA によって推定された N 個の分離音 $\hat{\boldsymbol{y}}_{ij,1}, \ldots, \hat{\boldsymbol{y}}_{ij,N}$ が得られているため，拡散性音源の空間相関行列 $\boldsymbol{R}_i^{(u)}$ は次のように表現できる．

$$\boldsymbol{R}_i^{(u)} = \boldsymbol{R'}_i^{(u)} + \lambda_i \boldsymbol{b}_i \boldsymbol{b}_i^{\mathrm{H}} \tag{11}$$

$$\boldsymbol{R'}_i^{(u)} = \frac{1}{J} \sum_j \hat{\boldsymbol{y}}_{ij}^{(u)} \left(\hat{\boldsymbol{y}}_{ij}^{(u)}\right)^{\mathrm{H}} \tag{12}$$

$$\hat{\boldsymbol{y}}_{ij}^{(u)} = \sum_{n \neq n_h} \hat{\boldsymbol{y}}_{ij,n} \tag{13}$$

ここで，$\boldsymbol{R'}_i^{(u)}$ は ILRMA によって推定された雑音のランク $M-1$ 空間相関行列であり，$\boldsymbol{b}_i$ は $\boldsymbol{R'}_i^{(u)}$ の零固有値に対応する単位固有ベクトル，λ_i はスカラー変数である．ここで，$\boldsymbol{R}_i^{(u)}$ において推定すべき変数は λ_i だけであり，ILRMA によって推定された $\boldsymbol{R'}_i^{(u)}$ と $\boldsymbol{b}_i$ を固定して最適化を行う．最後に，式 (9) の目的音源の分散の事前分布を考慮したランク制約付き空間共分散モデルの負対数尤度関数 $\mathcal{L}$ は次のように表される．

$$\mathcal{L}(r_{ij}^{(h)}, r_{ij}^{(u)}, \lambda_i) = \sum_{i,j} \left[\boldsymbol{x}_{ij}^{\mathrm{H}} (\mathbf{R}_{ij}^{(x)})^{-1} \boldsymbol{x}_{ij} + \log\det \mathbf{R}_{ij}^{(x)} + (\alpha+1)\log r_{ij}^{(h)} + \frac{\beta}{r_{ij}^{(h)}} \right] + \text{const.} \tag{14}$$

ここで，const. は目的変数に依存しない定数である．この負対数尤度関数 $\mathcal{L}$ は expectation-maximization (EM) アルゴリズムを用いて最適化される [5]．

3 多チャネル補聴器システム

3.1 システムの仕様

本研究では実環境下で対面する人との会話シーンを想定し，8 個のマイクロホンを用いてインパルス応答と拡散性雑音の収録を行う．収録のため，Fig. 1 (a) のように，スマートフォンを持った人間を模したダミーヘッドを作成した．ダミーヘッドの両耳には Figs. 1 (b), (d) のように，片耳に 3 個ずつ，両耳を合わせて計 6 個の無指向性マイクロホンを取り付けた．スマートフォンは，Fig. 1 (c) のようにダミーヘッドの胸部から 20 cm の位置に取り付け，胸部側に先端が向くよう 2 個の無指向性マイクロホンを 4 cm の間隔で取り付けた．合計 8 ch のマイクロホンを装備したダミーヘッドを用いて収録を行う．便宜上，Figs. 1 (b)，(c)，(d) のように各マイクロホンに対してナンバリングを行った．ダミーヘッドの身長は 170 cm とし，高さを調節した台に Fig. 1 (a) のダミーヘッドを乗せて収録した．また，ダミーヘッドと同身長の人との対話を想定し，床から口元までの高さを測り，スピーカの高さを 152 cm とした．

3.2 インパルス応答と拡散性雑音の収録

インパルス応答の計測方法として，時間引き伸ばしパルス (time stretched pulse: TSP) 信号を用いた．収録環境及び TSP 信号の収録条件を Table 1 に示す．Table 1 の仕様に基づき，ダミーヘッドからスピーカへの距離を 75 cm, 100 cm, 150 cm に，角度は正面方向 (0°) に加え左右にそれぞれに 20° 変化させ，計

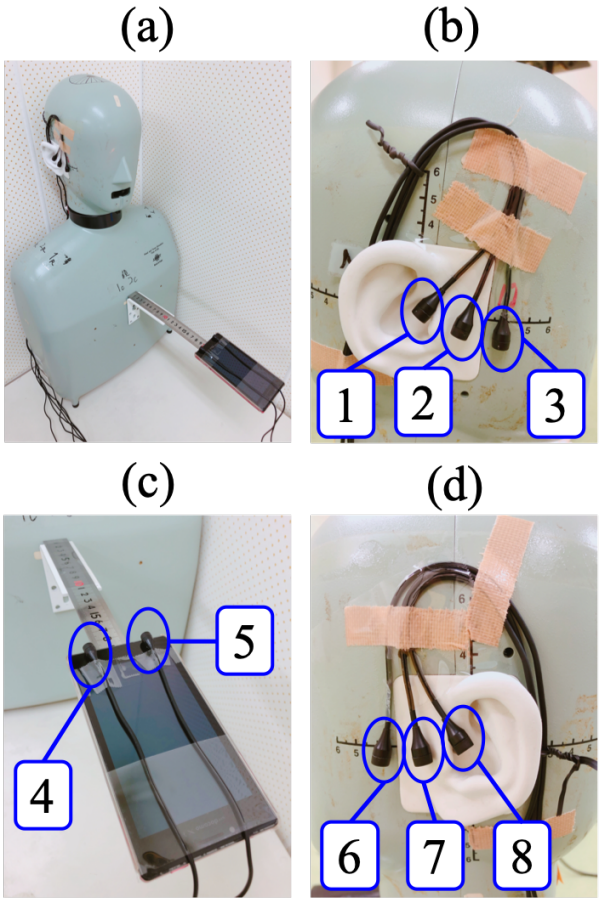

Fig. 1　(a) Overall view of dummy head and torso, (b) right-ear microphone array, (c) smartphone's microphones, and (d) left-ear microphone array.

Table 1　Recording conditions

Recording location	Studio
Reverberation time (T_{60})	300 ms
Microphone	C417 PP (AKG)
Loudspeaker	ADIVA11 (Anthony Gallo)
TSP length	65536 samples
Recording sampling freq.	48 kHz

9 パターンのスピーカ位置における TSP 信号を計測した．計測する 9 パターンのスピーカの位置の概略図を Fig. 2 に示す．

雑音データの作成として，数人が自由に移動・会話している状況を想定し収録を行った．雑音源は目的話者より外に存在するとしているため，協力者にはダミーヘッド前方半径 150 cm の半円より外側を周回させた．

4　評価実験

4.1　実験条件

本評価実験の目的は，両耳とスマートフォンのマイクロホンを用いた補聴器体系において，実環境下での ILRMA とランク制約付き空間共分散モデル推定法を比較し，それらの有効性について調査することである．音声データベース JNAS [9] の女声データ 1 文 に 3.2 節で収録したインパルス応答を畳み込んだものを

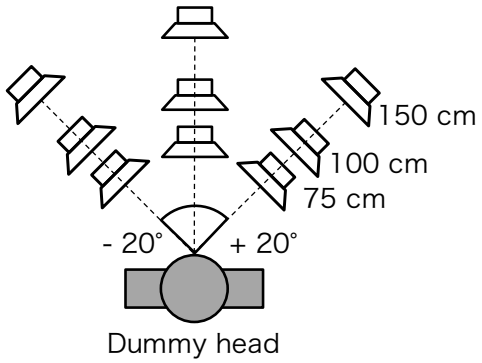

Fig. 2　Position of loudspeaker (mouth of conversation partner) for nine recording cases.

Table 2　Experimental conditions for BSS

Sampling freq.	16 kHz
FFT length	1024 sample (50% overlap)
Window	Hamming window
Number of bases in low-rank model	10
Number of iterations in ILRMA	50
Initialization of $\boldsymbol{W}_i$ in ILRMA	Identity matrix
Number of iterations in rank-constrained SCM estimation	10

目的信号とした．拡散性雑音には 3.2 節で収録した雑音を用いた．ただし，使用したコーパスデータのサンプリング周波数が 16 kHz であったため，48 kHz で収録したインパルス応答及び雑音をダウンサンプリングした．実験するにあたり，入力 SNR は −10 dB, −5 dB, 0 dB，ランク制約付き空間共分散モデル推定法における形状母数パラメータ α は 0.5, 1.1, 10, 20 と変化させ，尺度母数パラメータ β は 10^{-16} とした．ILRMA とランク制約付き空間共分散モデル推定法において観測信号を主成分分析を用いて白色化を行い，異なる乱数初期値で 10 回試行した．その他の条件を Table 2 に示す．以上の条件で，評価尺度として source-to-distortion ratio (SDR) 改善量 [10] を用いて分離性能を比較した．

4.2　反復による SDR 改善量の傾向

入力 SNR が −10 dB，マイクロホン 1 (右耳外耳道付近) の平均 SDR 改善量の反復による変化を Fig. 3 を示す．全ての場合においてランク制約付き空間共分散モデル推定法が ILRMA を上回っていることがわかる．また，内部パラメータ α の値によって SDR 改善量の変化に大きく違いが現れることがわかった．今回調査した範囲では，2，3 回程度の少ない反復で高い SDR 改善量を達成することが分かった．これにより，ランク制約付き空間共分散モデル推定法の収束が大変速いことが示された．

4.3　様々な入力 SNR における SDR 改善量

次に，角度を 0° に限定し，各入力 SNR における SDR 改善量の傾向について，ILRMA とランク制約付き空間共分散モデル推定法の性能を調査する．ただし，ランク制約付き空間共分散モデル推定法は 4.2 節の結果に基づき，SDR 改善量が概ね大きい反復 2 回

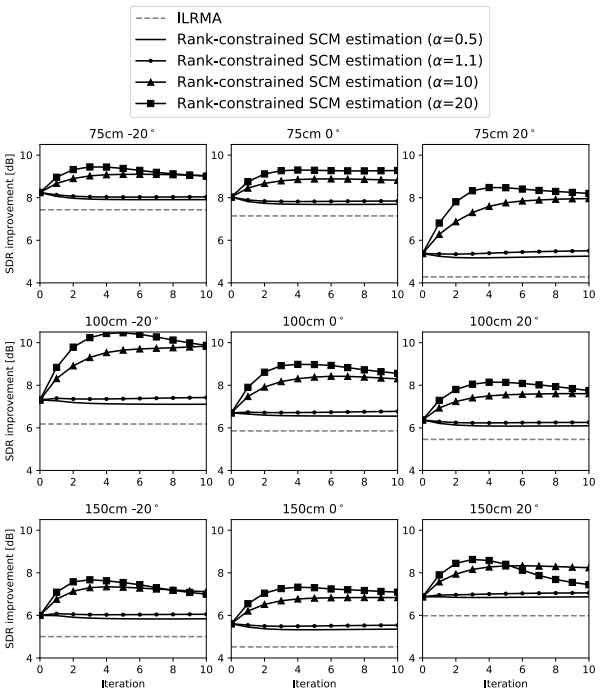

Fig. 3 Average SDR improvements for each iteration at microphone 1 under −10 dB input SNR condition.

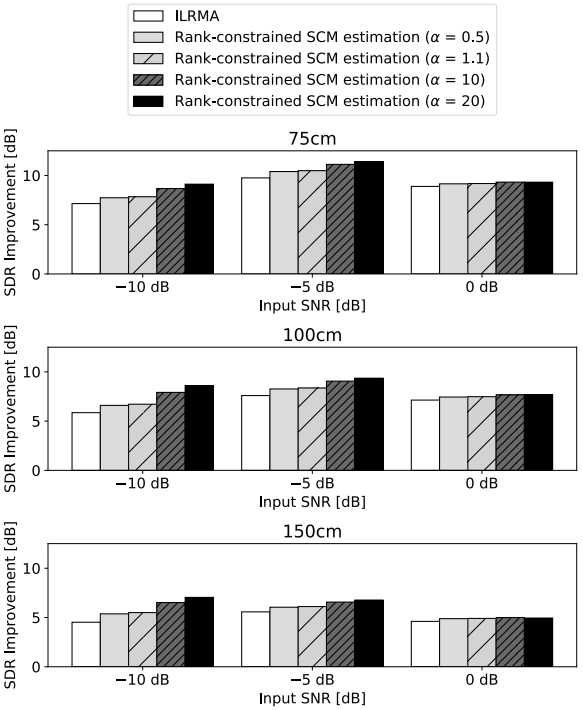

Fig. 4 Average SDR improvements of ILRMA and rank-constrained SCM estimation after two iterations at microphone 1 when target source is located at 0°.

目の結果を用いて比較を行う．

右耳外耳道付近のマイクロホンでの平均 SDR 改善量の結果を Fig. 4 に示す．ILRMA は観測信号に対して十分な分離性能を達成しており，今回の実験体系において有効であると言える．ランク制約付き空間共分散モデル推定法について，入力 SNR が −10 dB，−5 dB の場合の SDR 改善量が ILRMA と比較して大きい．さらに，パラメータ α が大きいほど性能が良いことから，事前分布により強いスパース性を仮定することで性能が向上することを確認した．

5 おわりに

本研究ではスマートフォンのマイクロホンを含めた分散マイクロホンアレー補聴器システムを提案し，提案システムにおける実環境下データの収録を行った．並びに，実環境データに対する ILRMA とランク制約付き空間共分散モデル推定法の有効性についても調査した．結果から，提案した実験体系においてもランク制約付き空間共分散モデル推定法は ILRMA より有効な手法であることを確認した．また，ランク制約付き空間共分散モデル推定法は低い入力 SNR であるほど効果的に動作することが分かった．

謝辞 本研究の一部は，SECOM 科学技術支援財団，JSPS 科研費 19H01116 の助成を受けたものである．

参考文献

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

[2] A. Hiroe, “Solution of permutation problem in frequency domain ICA using multivariate probability density functions,” in *Proc. of ICA*, 2006, pp. 601–608.

[3] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE/ACM Trans. on ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[4] 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.

[5] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, “Efficient full-rank spatial covariance estimation using independent low-rank matrix analysis for blind source separation,” in *Proc. EUSIPCO*, 2019.

[6] 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.

[7] N. Q. K. Duong, E. Vincent, and R. Gribonval, “Underdetermined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[8] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, “Blind spatial subtraction array for speech enhancement in noisy environment,” *IEEE Trans. ASLP*, vol. 17, no. 4, pp. 650–664, 2009.

[9] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” *The Journal of Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[10] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.