マルチチャネル非負値行列因子分解における チャネル数増加に伴う逐次的初期値設定法*

☆浦本昂伸 (大分大), 太刀岡勇気, 成田知宏 (三菱電機),
三浦伊織, 上ノ原進吾, 古家賢一 (大分大)

1　はじめに

非負値行列因子分解 (Nonnegative Matrix Factorization: NMF)[1] とは非負値の行列を分解し、解析を行う手法である。行列表現できるデータならば分解可能であるため、音や画像、文書など多種多様なものに利用できる。音響分野ではマルチチャネル拡張によって空間情報を活用することで音源分離を行う手法が提案されている [2, 3]。しかし、従来のマルチチャネル NMF (MNMF) は自由度の高いモデルであるため、局所最適解に陥りやすく、分離性能の初期値依存性が課題となっている [4, 5]。また、本稿の実験で示すように、チャネル数が増加するほど、この初期値依存性が顕在化するため、音源分離が困難となる。

本稿は、3 チャネル以上を用いた MNMF に有効な逐次的初期値設定法を提案する。初期値にランダムな値を設定する従来法に対して、分離性能の比較を行い、提案法の有効性を検証していく。

2　MNMF

2.1　概要

MNMF[2, 3] とは、NMF をマルチチャネル拡張したものであり、観測行列 $\mathbf{X}$ を 4 つの行列 $\mathbf{H}$、$\mathbf{Z}$、$\mathbf{T}$、$\mathbf{V}$ に分解する。MNMF では空間情報を用いてスペクトル基底を L 個の音源にクラスタリングすることで事前の学習なしで音源分離を実現する。位相情報を扱うために複素数を用いるので、複素数における非負性に対応するものとして、エルミート半正定値行列を用いる [2]。

2.2　定式化

M をマイクロホン数として入力ベクトルを $\tilde{\mathbf{x}} = [\tilde{x}_1, \cdots, \tilde{x}_M]^\top$ とする。ただし、$^\top$ は転置を表す。$\tilde{x}_m$ は m 番目のマイクロホンでの Short Time Fourier Transform (STFT) の複素係数であり、スペクトログラムを指す。周波数 i $(1 \leq i \leq I)$、時間 j $(1 \leq j \leq J)$ のとき $\tilde{\mathbf{x}}_{ij}$ で表すと行列 $\mathbf{X}$ は $\mathsf{X}_{ij} = \tilde{\mathbf{x}}_{ij}\tilde{\mathbf{x}}_{ij}^H$ もしくは i、j それぞれについて

$$\mathsf{X} = \tilde{\mathbf{x}}_m\tilde{\mathbf{x}}_m^H = \begin{bmatrix} |\tilde{x}_1|^2 & \cdots & \tilde{x}_1\tilde{x}_M^* \\ \vdots & \ddots & \vdots \\ \tilde{x}_M\tilde{x}_1^* & \cdots & |\tilde{x}_M|^2 \end{bmatrix} \tag{1}$$

で表される。ただし、^{H} はエルミート転置を表す。すなわち、I 行 J 列の行列 $\mathbf{X}$ はそれぞれの要素が $M \times M$ の複素行列を持つ階層的なエルミート半正定値行列となる。この行列 $\mathbf{X}$ を MNMF で分解すると、式 (2) で表されるように、K 個の基底から成る基底行列 $\mathbf{T}(\in \mathbb{R}^{I\times K})$、アクティベーション行列 $\mathbf{V}(\in \mathbb{R}^{K\times J})$、音源の空間情報を示す空間相関行列 $\mathbf{H}$ と音源の空間情報と各基底を関連付ける潜在変数行列 $\mathbf{Z}(\in \mathbb{R}^{L\times K})$ という 4 つの行列に分解できる。

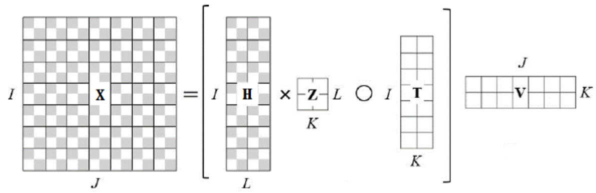

Fig. 1　MNMF で分解された行列の例

$$\mathbf{X} = (\mathbf{HZ} \circ \mathbf{T})\mathbf{V} \tag{2}$$

ただし、$\circ$ はアダマール積を表す。行列 $\mathbf{H}$ は行列 $\mathbf{X}$ と同様にそれぞれの要素が $M \times M$ の複素行列を持つ I 行 L 列の階層的なエルミート半正定値行列である。Fig. 1 は式 (2) を図式化したものである。このとき、右辺は

$$\hat{\mathsf{X}}_{ij} = \sum_{k=1}^{K}\left(\sum_{l=1}^{L}\mathsf{H}_{il}z_{lk}\right)t_{ik}v_{kj} \tag{3}$$

と表すことができ、理想的には行列 X と $\hat{\mathsf{X}}_{\mathrm{ij}}$ を要素に持つ行列 $\hat{\mathsf{X}}$ は等しくなる。しかし、一般的には誤差が生じるため、MNMF では行列 X と行列 $\hat{\mathsf{X}}$ との距離 $D_*(\mathsf{X}, \hat{\mathsf{X}})$ を定義し、この距離を最小化する行列 $\mathbf{H}$、$\mathbf{Z}$、$\mathbf{T}$、$\mathbf{V}$ を求める。今回はダイナミックレンジが大きい音楽や音声に適している Itakura-Saito (IS) divergence[7] を用いて以下のように定義する。

$$D_{IS}(\mathsf{X}_{ij}, \hat{\mathsf{X}}_{ij}) = tr(\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}) - \log\det\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1} - M \tag{4}$$

ただし、$tr\ (\cdot)$ は対角要素の和を表している。

2.3　行列分解アルゴリズム

$D_{IS}(\mathsf{X}, \hat{\mathsf{X}})$ を最小化するために、Multiplicative update rule[8] と呼ばれる反復アルゴリズムを、ランダムな非負の値で初期化した行列 $\mathbf{T}$、$\mathbf{V}$、$\mathbf{Z}$ ならびに各要素へ単位行列を持たせた行列 $\mathbf{H}$ に繰り返し適用する。IS divergence を用いた場合、更新式は以下のよ

*Sequential initialvalue setting associated with increasing number of channels in Multi-channel Non-negative Matrix Factorization. by Takanobu Uramoto (Oita University), Yuuki Tachioka, Tomohiro Narita (Mitsubishi Electric), Iori Miura, Shingo Uenohara, and Ken'ichi Furuya (Oita University)

うになる。

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_l z_{lk} \sum_j v_{kj} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}{\sum_l z_{lk} \sum_j v_{kj} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}} \tag{5}$$

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_l z_{lk} \sum_i t_{ik} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}{\sum_l z_{lk} \sum_i t_{ik} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}} \tag{6}$$

$$z_{lk} \leftarrow z_{lk} \sqrt{\frac{\sum_{i,j} t_{ik} v_{kj} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}{\sum_{i,j} t_{ik} v_{kj} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}} \tag{7}$$

H_{il} については次式の A，B を係数に持つ代数リッカチ方程式を解くことで求めることができる。

$$A = \sum_k z_{lk} t_{ik} \sum_j v_{kj} \hat{\mathsf{X}}_{ij}^{-1} \tag{8}$$

$$B = \mathsf{H}'_{il} \left(\sum_k z_{lk} t_{ik} \sum_j v_{kj} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \mathsf{X}_{ij}^{-1} \right) \mathsf{H}'_{il} \tag{9}$$

ただし、H'_{il} は更新前の行列 H_{il} を表している。

2.4 正規化

行列 $\mathbf{H}$ と行列 $\mathbf{Z}$ については、更新毎に発散を防ぐために正規化を行わなければならない。正規化は以下の式で行った。

$$\mathsf{H}_{il} = \frac{\mathsf{H}_{il}}{tr(\mathsf{H}_{il})}, \quad z_{lk} = \frac{z_{lk}}{\sum_l z_{lk}} \tag{10}$$

2.5 音源分離

音源分離を行うために次式で表されるウィナーフィルタを用いる。

$$Y = \frac{\hat{S}}{\hat{S} + N} X \tag{11}$$

ただし、Y は目的信号、$\hat{S}$ は目的信号の推定値、N は雑音信号、X は雑音信号を含んだ目的信号を示す。$\tilde{y}_{ij}^{(l)}$ を分離後の音源としたとき、$Y = \tilde{y}_{ij}^{(l)}$、$\hat{S} = (\sum_{k=1}^{K} z_{lk} t_{ik} v_{kj})\ \mathsf{H}_{il}$、$\hat{S} + N = \hat{\mathsf{X}}_{ij}$、$X = \mathsf{X}_{ij}$ を代入すると、次式のマルチチャネルウィナーフィルタとなり、各音源に対応した分離信号を得られる。

$$\tilde{y}_{ij}^{(l)} = \left(\sum_{k=1}^{K} z_{lk} t_{ik} v_{kj} \right) \mathsf{H}_{il} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \tag{12}$$

3 MNMF の課題を示す実験

MNMF は自由度の高いモデルであるため、局所最適解が増え、初期値依存による分離性能のばらつきが問題となることが報告されている [4]。チャネル数を増加させた場合の初期値依存性について実験的に分析を行う。

3.1 実験条件

実験に用いた混合信号は Table 3[11] の音楽データに Fig. 2 の環境で測定した RWCP 実環境音声・音響データベースのインパルス応答を畳み込み作成した。Fig. 2 においてマイクロホンは右から順に 1-14 まで番号が付いており、今回の実験で使用したマイクロホン番号を Table 1 に示す。ここで、チャネル数を増やした際に、元のマイクロホンが含まれているようにした。例えば 3 チャネルのマイクロホンの組には 2 チャネルのマイクロホンの組が含まれている。分離処理に用いたパラメータを Table 2 に示す。なお、使用した隣接するマイクロホン間の距離は 5.66cm である。また、MNMF での IS divergence の計算 (4) において行列式が 0 になるのを防ぐために X_{ij} の対角要素に 10^{-10} を足している。プログラムは Sawada らのアルゴリズム [2] を MATLAB で実装した。ただし、音源数は既知として pairwize-merge は導入せず、Multiplicative update rule の反復適用のみ行っている。また、文献 [2] に倣い、初めの 20 回は空間相関行列 $\mathbf{H}$ と潜在変数行列 $\mathbf{Z}$ を更新せず、その他の変数のみを更新した。一様分布から生成した生成した 10

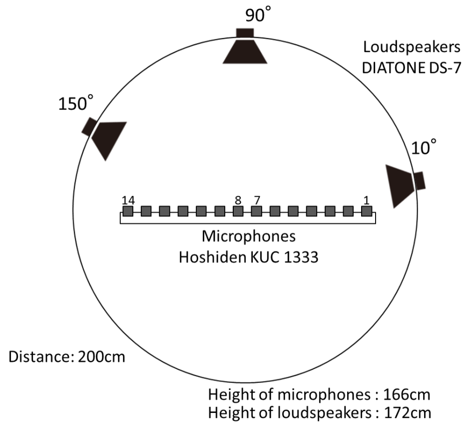

Fig. 2 音源の配置図

Table 1 使用マイクロホン番号

2ch	6,8
3ch	6,8,10
4ch	4,6,8,10
5ch	4,6,8,10,12
6ch	2,4,6,8,10,12

Table 2 分離処理に用いるパラメータ

インパルス応答長	10885
サンプリング周波数	16kHz
フレームサイズ	1024
シフトサイズ	256
基底数	30
音源数	3
更新回数	500

Table 3 実験に用いた音楽データ

ID	Author/Song	Snip	Part
1	Bearlin Roads	85-99 (14 sec)	piano ambient vocals
2	Another Dreamer The Ones We Love	69-94 (25 sec)	drums vocals guitar
3	Fort Minor Remamber The Name	69-94 (24 sec)	drums vocals violin_synth
4	Ultimate Nz Tour	54-78 (18 sec)	drums guitar synth

個の初期値パターンを用意し、音源分離を実行する。ただし、文献 [2] と同様に空間相関行列 $\mathbf{H}$ には各要素の対角成分が 1/M の対角行列を持たせ、潜在変数行列 $\mathbf{Z}$ は 0.2 から 0.4 の一様乱数の値を持たせた。分離性能の評価基準は次式の Signal-to-Distortion Ratio (SDR)[3] を用いた。

$$\mathrm{SDR} = 10\log_{10}\frac{\sum_t s^{\mathrm{img}}(t)^2}{\sum_t y^{\mathrm{spat}}(t)^2 + y^{\mathrm{int}}(t)^2 + y^{\mathrm{artif}}(t)^2} \tag{13}$$

ただし、s^{est} は目的音源の推測信号、s^{img} は目的音源の正解信号、y^{spat} は空間（フィルタリング）歪み、y^{int} は目的音源以外の音源の信号、y^{artif} は分離処理による信号の歪みを表す。

3.2 チャネル数増加に伴う初期値依存性

初期値ランダムの従来法において、単純にマイクロホン数を増やした場合の分離性能を示す。Fig. 3 は各音楽データとチャネル毎の分離後における 3 音源の平均 SDR を示したものである。エラーバーは標準偏差を示す。この図から、音源にもよるが 3 チャネルよりも 4、5、6 チャネルの方が分離性能が低下している。これは、局所最適解による初期値依存性がチャネル数増加に伴って、顕在化するため分離性能が低下したと考えられる。

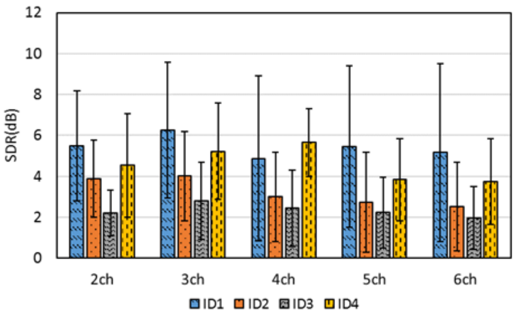

Fig. 3 チャネル数増加に伴う初期値依存性

4 提案手法

4.1 教師有り逐次的初期値設定法

従来の著者らの検討により、MNMF の分離性能は空間相関行列 $\mathbf{H}$ に対する初期値依存性が大きいということが分かっている [6]。そこで、空間相関行列 $\mathbf{H}$ に着目する。m チャネルで分離を行い、m チャネルの空間相関行列 $\mathbf{H}$ は、$m+1$ チャネルの空間相関行列 $\mathbf{H}$ の部分行列になっていることを利用して、SDR が最も高い時の分離後の空間相関行列 $\mathbf{H}$ を次の $m+1$ チャネルの空間相関行列 $\mathbf{H}$ の初期値に設定し、分離を行う。$m = 2, 3, 4, 5$ とし、チャネル数増加に伴い逐次的にこの処理を行う。始めに分離を行う 2 チャネルの初期値には、従来法と同様にランダムな値を設定する。Fig. 4 に示すように、チャネル数を順次増やしていき、逐次的に初期値を設定する。初期値の設定箇所以外には、従来法と同様に単位行列を設定している。SDR を計算するためには、教師情報として元の音源信号を用いる。

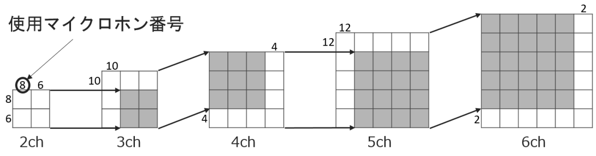

Fig. 4 初期値の設定方法

4.2 教師無し逐次的初期値設定法

教師無しで初期値を設定し、チャネル数増加に伴う分離性能の分析を行う。これまでにバイナリマスクであらかじめ分離したデータからクロススペクトル法を用いて空間相関行列 $\mathbf{H}$ を計算し、MNMF の初期値に設定することで、従来法と比べ分離性能が向上し、MNMF とバイナリマスクを組み合わせることの有効性が分かっている [6]。今回は、始めに分離を行う 2 チャネルの初期値に以下に示すバイナリマスクとクロススペクトル法により求められた空間相関行列 $\mathbf{H}$ を設定し、教師有り提案法と同様に分離後の空間相関行列 $\mathbf{H}$ を逐次的に設定する。この時、SDR を計算せずにランダムに選択し、設定する。教師無し提案法を以下の 2 つの指標 (上限・下限) と比較する。

- 最も高い SDR が得られた分離後の空間相関行列 $\mathbf{H}$ を逐次的に設定 (Fig. 6 で “上限” と示す)
- 最も低い SDR が得られた分離後の空間相関行列 $\mathbf{H}$ を逐次的に設定 (“下限”)

なお、音楽データは Table 3 の ID4 を使用。

4.3 バイナリマスク [10]

バイナリマスクとは、各音源の到来時間差に基づいて時間周波数上でマスキングを行い、音源分離を行う手法である。例えば、目的音源が正面方向である場合、マイクロホン間の位相差は 0 である。雑音が 0 度方向から到来する場合、位相差は大きくなるので、マイクロホン間の位相差がゼロから離れた時間周波数ビンのパワーをマスキングすれば、目的音源を強調することができる。マスク W は以下のように閾値を用いて設定される。

$$W_{i,j} = \begin{cases} \epsilon & \text{if } |\theta_{i,j}| > \theta_{\mathrm{c}}, \\ 1 & \text{if } |\theta_{i,j}| \leq \theta_{\mathrm{c}}, \end{cases}$$

ϵ は十分小さい定数、$\theta_{i,j}$ は時間周波数ビンの位相差、θ_c は事前に定めておく閾値である。事前に音源方向が分かっていれば、それぞれの音源が強調されるようにマスキングすることができる。

4.4 クロススペクトル法 [9]

音源データのスペクトルをフーリエ変換することで

$$A_i = \begin{bmatrix} a_{i,1} & \ldots & a_{i,M} \end{bmatrix}^{\top} \tag{14}$$

M 行 1 列のステアリングベクトル A_i が与えられる。A_i と、そのエルミート転置 (1 行 M 列) の積

$$\mathsf{H}_i = A_i A_i^H = \begin{bmatrix} |a_{1,1}|^2 & \cdots & a_{i,1}a_{i,M}^* \\ \vdots & \ddots & \vdots \\ a_{i,M}a_{i,1}^* & \cdots & |a_{i,M}|^2 \end{bmatrix} \tag{15}$$

は周波数ビン i における空間相関を表す。L 個の各音源から H_i を作成することで、MNMF における I 行 L 列の空間相関行列 $\mathbf{H}$ として設定出来る [9]。本稿では、各マイクロホンのスペクトル成分を要素に持つ M 行 1 列の行列とそのエルミート転置の積から空間相関行列 $\mathbf{H}$ を算出する手法をクロススペクトル法と呼ぶ。ここでは、データの全区間から空間相関行列 $\mathbf{H}$ を計算できるように、フレームサイズおよびシフトサイズを 1024 として、STFT を行う。各フレームからクロススペクトル法で空間相関行列 $\mathbf{H}$ を計算し、全フレームの空間相関行列 $\mathbf{H}$ の平均の値を MNMF の初期値とした。

5 実験

提案法の有効性を確認するために従来法を比較して実験を行う。実験条件は 3 節と同じである。

5.1 実験結果

Fig. 5 は教師有り提案法で分離を行った時の結果である。SDR が最も高い時の分離後の空間相関行列 $\mathbf{H}$ をチャネル数増加に伴い逐次的に設定することで、Fig. 3 の従来法よりも SDR が向上し、標準偏差が小さくなることから、分離性能が向上していることが分かる。また、チャネル数増加に伴い SDR が向上している。

Fig. 6 は教師無し提案法で分離を行った時の結果で、従来法よりも分離性能が向上しており、推定される SDR の範囲内に概ね収まっている。ただし、提案法の 4、5、6 チャネルを比較するとチャネル数が増加しても必ずしも SDR が向上していない。

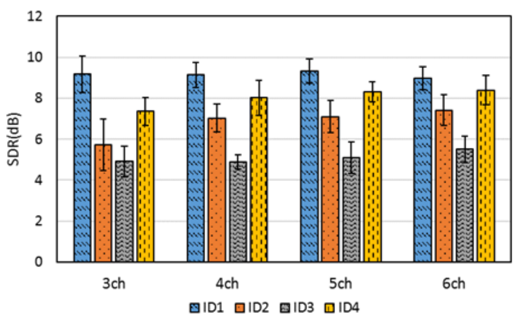

Fig. 5 教師有り提案法による実験結果

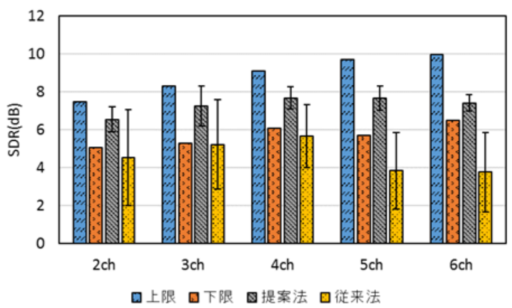

Fig. 6 教師無し提案法による実験結果 (ID4)

5.2 考察

Fig. 5 から提案法では、チャネル数増加に伴い SDR が向上していることが分かる。従来法では、チャネル数と共に行列の自由度が増加するため、局所最適解に陥りやすくなる。しかし、良いパラメータを推定できている行列を逐次的に設定することで、局所最適解に陥るのを避け、マイクロホン数の増加に伴う多くの情報量を適切に扱えるため SDR が向上したと考えられる。

Fig. 6 から、各チャネルで常に最良の初期値を設定することが出来れば上限のように SDR は向上するが、ランダムに選択して設定するとチャネル数増加に伴い必ずしも分離性能が改善しない場合も見られた。ただし、ランダムに設定する従来法よりは分離性能が向上した。

6 まとめ

本稿では、MNMF のチャネル数増加に伴う初期値依存性を解決するために逐次的初期値設定法を提案した。教師有りと教師無しの 2 つの方法を提案し、両方とも、従来法よりも分離性能が向上することから提案法の有効性を確認した。ただし、教師無し提案法では、教師有りの場合の上限に達していないことから、初期値の設定に何らかの基準を設けることで、分離性能が向上する余地がある。

参考文献

[1] D.D. Lee *et al.*, “Learning the Parts of Objects with Nonnegative Matrix Factorization,” Nature, vol. 401, pp. 788-791,1999.

[2] H. Sawada *et al.*, “Multichannel Extensions of Non-Negative Matrix Factorization with Complex-Valued Data,” IEEE Trans. ASLP, vol.21, no.5, pp. 971-982, 2013.

[3] E. Vincent *et al.*, “First Stereo Audio Source Separation Evaluation Campaigh: Data Algprithm and Results,” Independent Component Analysis and Signal Separation(Springer, Bearlin, 2007), pp. 552-559.

[4] 吉山 文教, 他: “マルチチャネル非負値行列因子分解における分離性能の高い初期値の判別法” 音講論集, pp. 777-780, 2014.

[5] 三浦 伊織, 他: “ マルチチャネル非負値行列因子分解における初期値依存性の挙動解析” 音講論集, pp. 669-672, 2016.

[6] 三浦 伊織, 他: “ マルチチャネル非負値行列因子分解におけるバイナリマスクを用いた初期値設定法” 音講論集, pp. 425-428, 2016.

[7] C. Fevotte, N. Bertin *et al.*, “Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis,” Neural Comput., vol. 21, no. 3, pp. 793-830, 2009.

[8] M. Nakano *et al.*, “Convergence-Guaranteed Multiplicative Algorithms for Non-Negative Matrix Factorization with Beta-Divergence,” In Proc.MLSP 2010, pp. 283-288, 2010.

[9] 北村 大地, 他: “ランク 1 空間モデルを用いた効率的な多チャネル非負値行列因子分解” 音講論集, pp. 579-582, 2014.

[10] H Sawada *et al.*, “Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment” IEEE Trans. Audio, Speech, Language Process., vol. 19, pp. 516-527, Mar. 2011.

[11] S. Araki *et al.*, “The 2011 Signal Separation Evaluation Campaign (SiSEC2011): -Audio Source Separation,” Latent Variable Analysis and Signal Separation(Springer, Bearlin, 2012), pp. 414-422.