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*

☆三浦伊織 (大分大), 太刀岡勇気, 成田知宏, 石井純 (三菱電機), 吉山文教, 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、多くの局所最適解が存在し、分離性能に対する初期値依存性が課題となっている [4, 5]。

本稿は、通常ランダムに設定することの多い分離行列の初期値に対して、あらかじめバイナリマスクで分離したデータから計算した初期値を設定することで、分離性能を向上させることを目的とする。

2 MNMF[2, 3]

2.1 概要

MNMF とは、NMF をマルチチャネル拡張したものであり、観測行列を 4 つの行列 **H**、**Z**、**T**、**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

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

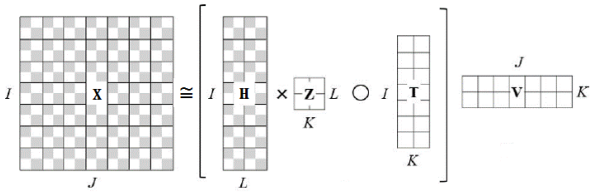

Fig. 1 MNMF で分解された行列の例

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。

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

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

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

と表すことができ、理想的には行列 X と $\hat{\mathsf{X}}_{ij}$ 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 **T**、**V**、**H**、**Z** 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[6] を用いて以下のように定義する。

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

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

Multiplicative update rule[7] と呼ばれる反復アリゴリズムを、ランダムな非負の値で初期化した行列

*Multi-channel Non-negative Matrix Factorization with Binary Mask Initialization. by Iori Miura (Oita University), Yuuki Tachioka, Tomohiro Narita, Jun Ishii (Mitsubishi Electric), Fuminori Yoshiyama, Shingo Uenohara, and Ken'ichi Furuya (Oita University)

T、**V**、**Z** ならびに各要素へ単位行列を持たせた行列 **H** に繰り返し適用することで、$D_{IS}(\mathsf{X},\hat{\mathsf{X}})$ を最小化するような各行列を得る。IS divergence を用いた更新式は以下のようになる。

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

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

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

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 **H** と行列 **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。

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

2.6 従来法の課題

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。Fig. 2 は MNMF アルゴリズムにランダムな初期値を 10 回与えて音源分離を行った際の分離性能を示している [4]。この図から、分離性能は初期値ごとに大きく異なっていることがわかる。

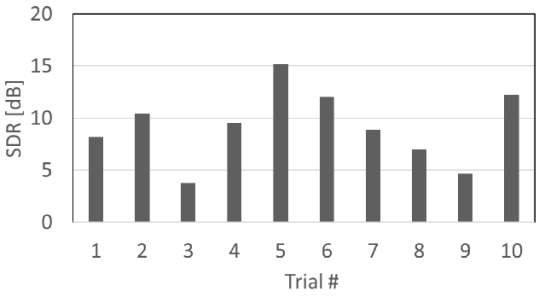

Fig. 2　音源分離性能の初期値依存性

3　バイナリマスクを用いた初期値設定

混合前の音源やインパルス応答から k-means 法 [8] やクロススペクトル法 [9] を用いて、基底行列 **T** および空間相関行列 **H** の初期値を計算することで、分離性能が向上することが分かっている [5]。しかし、多くの応用において、事前にそれらの情報を取得することは困難である。そこで本稿では他の音源分離を用いて取得したデータから、基底行列 **T** および空間相関行列 **H** の初期値を求める方法を提案する。今回は音源方向が既知であると仮定してバイナリマスク [10] を用いる。

3.1 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。

3.2 k-means 法 [8]

k-means 法とはデータ行列 X を任意数のクラスタに分割し、クラスタごとの平均を算出するアルゴリズムである。スペクトログラムに適用する場合、任意の基底数だけスペクトルパターンをクラスタリングし、各クラスタの平均の値を基底として使用することができる。この手法では混合前の 3 音源からそれぞれ 10 個ずつ基底を作成 (計 30 個) し、基底行列 **T** の初期値として設定する。

Table 1　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

3.3　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

$$A_i = \begin{bmatrix} a_{i,1} & \dots & a_{i,M} \end{bmatrix}^\top \tag{13}$$

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

$$\mathsf{H}_i = A_i A_i^H = \begin{bmatrix} |a_{1,1}|^2 & \cdots & a_{i,1}a_{i,M}^* \\ \vdots & \ddots & \vdots \\ a_{i,M}a_{i,1}^* & \cdots & |a_{i,M}|^2 \end{bmatrix} \tag{14}$$

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。

3.4　提案手法

混合信号にバイナリマスクを適用して分離した 3 つの音源データから、基底行列 $\mathbf{T}$ および空間相関行列 $\mathbf{H}$ の初期値を求める。基底行列 $\mathbf{T}$ を求める方法は k-means 法を用いる。また、空間相関行列 $\mathbf{H}$ を求める方法はクロススペクトル法を用いる。

4　実験

4.1　実験条件

実験に用いた混合信号は Table 1[11] の音楽データに Fig. 3 の環境で測定したインパルス応答 ($M = 2$) を畳み込み作成した。実験条件は Table 2 に示す。ランダムに生成した 10 個の初期値パターンを用意し、各提案手法によって作成した行列と置き換えて、音源

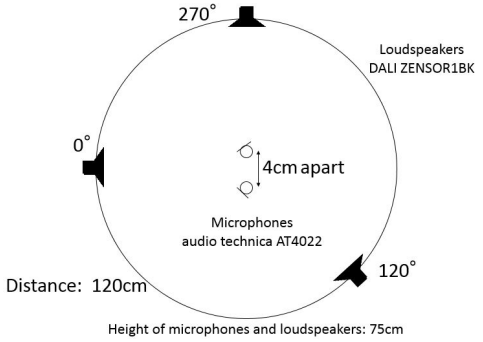

Fig. 3　音源の配置図

Table 2　実験条件

インパルス応答長	300
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

分離を実行する。分離性能の評価基準は音声と歪みの比を表す SDR[3] を用いている。ここでは以下のパターンで SDR の比較を行う。

1. バイナリマスク (Fig. 4 で “bin-mask” と示す)
2. 初期値をランダムとした従来法の MNMF (“random”)
3. k-means 法で基底行列 $\mathbf{T}$ を計算し、MNMF の初期値に設定 (“k-means”)
4. クロススペクトル法で空間相関行列 $\mathbf{H}$ を計算し、MNMF の初期値に設定 (“cross”)
5. “k-means” と “cross” を使用し、MNMF の初期値に設定 (“k-means & cross”)

手法 3 ～ 5 ではバイナリマスクを適用して分離した 3 つの音データから初期値を計算する。ただしマイク間隔が 4cm であることから、空間的エイリアシングの影響でおおむね 4250Hz 以上の信号は分離できていないと考えられる。

4.2　実験結果

Fig. 4 は各音楽データでの、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。バイナリマスクであらかじめ分離したデータから基底行列 $\mathbf{T}$ の初期値を計算することで (k-means)、ID3 を除いて、分離性能が向上することが分かった。また、クロススペクトル法で空間相関行列 $\mathbf{H}$ を計算し、初期値とした場合 (cross) は、ID3 を除き標準偏差が小さくなっている。しかし、基底行列 $\mathbf{T}$ と空間相関行列 $\mathbf{H}$ を計算し、初期値とした場合 (k-means & cross) は、空間相関行列 $\mathbf{H}$ のみを計算した場合 (cross) と比べると分離性能に大きな差が見られなかった。cross, k-means& cross は bin-mask

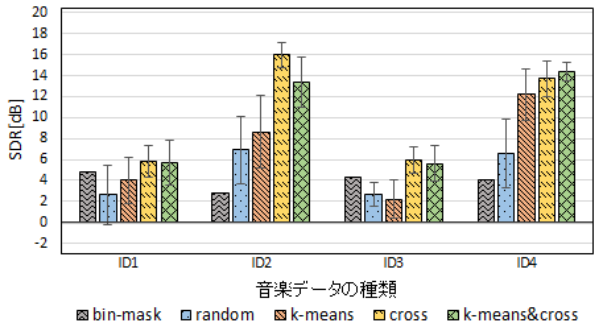

Fig. 4　バイナリマスクを用いた初期値設定法による実験結果

の分離性能を上回り、MNMF と bin-mask を組み合わせる提案法の有効性が示された。

4.3　考察

Fig. 4 を見ると、バイナリマスクを用いて分離したデータから基底行列 **T** を計算した場合は、混合前のデータから基底行列 **T** を計算する場合 [5] に比べて、分離性能の上がり具合が低い場合が多い。これは、バイナリマスクで完全に分離できていないために、各音源ごとの基底がうまく推定できなかったためと考えられる。また、音源ごとの SDR を見ると drum など広帯域な音源の分離性能が低かったので、他音源の音域と被る音源の分離が難しいことも原因の 1 つと考えられる。

バイナリマスクを用いて分離したデータから空間相関行列 **H** を計算した場合は、音楽データごとで分離性能の上がり具合が異なる。ランダムな初期値である従来法や各音源のインパルス応答から空間相関行列を計算した場合 [5] と比較してみると、もともと分離性能が低い楽曲に対して、提案手法の効果が薄いことが分かる。このことから、楽曲によって正しく空間相関を計算できるかどうかに差があると考えられる。また、本実験でクロススペクトル法を用いる時は、全フレームの平均の値を空間相関行列 **H** の初期値としているので、出現頻度の低い音域の重要度が低くなると考えられる。

5　まとめ

本稿では、バイナリマスクによって分離したデータから、MNMF の各行列の初期値を計算する手法を提案し、実験を行った。実験の結果から、あらかじめ分離したデータを使って空間相関行列 **H** を計算することで、従来法に比べて分離性能が上がることを確認した。しかし、楽曲データによって分離性能の上がり具合が異なるので、推定が難しいデータにおいても空間相関が推定できるような改良が必要となる。

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” 日本音響学会講演論文集, 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” 日本音響学会講演論文集, pp. 669-672, 2016

[6] C. Fvotte, 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.

[7] M. Nakano *et al.*, “Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence,” In Proc.MLSP 2010, pp. 283-288, 2010.

[8] D. Arthur, S. Vassilvitskii, “k-means++: The Advantages of Careful Seeding,” Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithm, pp. 1027-1035, 2007.

[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” 日本音響学会講演論文集, 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.