マルチチャネル非負値行列因子分解に対する非負二重 SVD 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 (三菱電機), 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解析可能であるため、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]。

本稿は、以前提案したバイナリマスクによる空間相関行列推定 [4] に加え、類似度による学習データから非負二重 SVD によって基底行列の計算を行い、あらかじめ計算した値を MNMF の初期値に設定することで、分離性能を向上させることを目的とする。今回は騒音環境下における音声認識実験により、提案法の有効性を示す。

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、複素数における非負性に対応する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 $\mathbf{X}$ の i、j 成分を $\mathsf{X}_{ij} \in \mathbb{C}^{M \times M}$ とし、$\mathsf{X}_{ij} = \tilde{\mathbf{x}}_{ij}\tilde{\mathbf{x}}_{ij}^H$ について

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

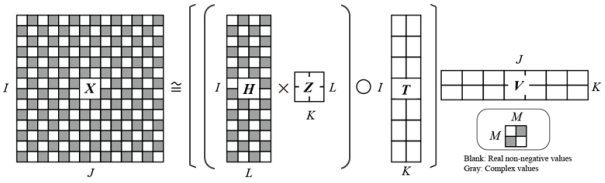

Fig. 1 MNMF で分解された行列の例

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 $\mathbf{X}$ は要素が複素行列となる階層的なエルミート半正定値行列である。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。

$$\mathbf{X} = [(\mathbf{H} \times \mathbf{Z}) \circ \mathbf{T}]\mathbf{V} \quad (2)$$

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 $\mathbf{H}$、$\mathbf{Z}$、$\mathbf{T}$、$\mathbf{V}$ 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[5] 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[6] 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

---

*Initial value setting method using non-negative double singular value decomposition for multi-channel non-negative matrix factorization. by Iori Miura (Oita University), Yuuki Tachioka, Tomohiro Narita (Mitsubishi Electric), Shingo Uenohara, and Ken'ichi Furuya (Oita University)

†2017 年 4 月 三菱電機 退職

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。IS divergence を用いた更新式は以下のようになる。

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

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

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

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 $\mathsf{H}_{il} A \mathsf{H}_{il} = B$ で求めることができる。

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

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

ただし、H'_{il} は更新前の行列 H_{il} を表しており、解き方は文献 [2] に示されている。

2.4 正規化

行列 **H** は式 (2) の一意性を保つため、行列 **Z** は確率の定義からの要請によるため、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。

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

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。

$$Y = \frac{\hat{S}}{\hat{S} + N} X \quad (11)$$

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、$Y = \tilde{y}_{ij}^{(l)}$、$\hat{S} = (\sum_{k=1}^{K} z_{lk} t_{ik} v_{kj})\ \mathsf{H}_{il}$、$\hat{S} + N = \hat{\mathsf{X}}_{ij}$、$X = \tilde{\mathbf{x}}_{ij}$ 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 [2]。

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

2.6 MNMF 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、分離性能の初期値依存性が問題となる。Fig. 2 は MNMF 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 (SDR[3]) 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。

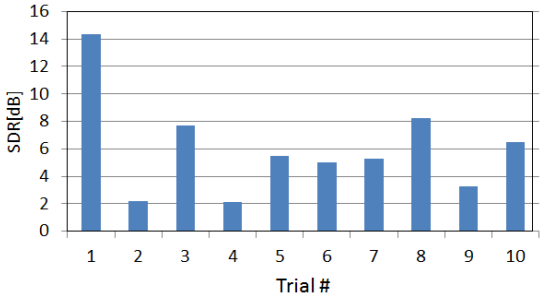

Fig. 2　音源分離性能の初期値依存性

3 提案手法

混合前の音源やインパルス応答から非負二重 SVD[7] やクロススペクトル法 [8] 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 [4]。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。以前の研究で、我々はバイナリマスクを用いた行列 **H** の初期値設定法を提案し、音声認識実験において音声認識率が向上することを確認した [9]。本稿では、行列 **H** の初期値の計算に加え、類似度による学習データから非負二重 SVD を用いて行列 **T** の初期値を計算することで、音声認識率の向上を図る。

3.1 空間相関行列 H の計算法

音源方向を既知と仮定し、バイナリマスクを使用する。バイナリマスク [10] を用いて取得したデータから、行列 **H** の初期値を求めることで、MNMF の分離性能を向上させる。

3.1.1 バイナリマスク

バイナリマスク [10] 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 M は以下のように閾値を用いて設定される。

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

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。

3.1.2 クロススペクトル法

音源データのスペクトルをフーリエ変換することで

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

M 行 1 列の 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

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

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 $\mathsf{H_i}$ を作成することで、MNMF における I 行 L 列の行列 $\mathbf{H}$ として設定出来る [8]。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行列 $\mathbf{H}$ を計算し、全フレームの行列 $\mathbf{H}$ の平均の値を MNMF の初期値とした。

3.2 基底行列 T の計算法

非負二重 SVD (Non-negative Double Singular Value Decomposition) 法 [7] と呼ばれる特異値分解を用いた初期化手法を使用して行列 $\mathbf{T}$ を作成する。

評価データが与えられたときに i-vector[11] を算出し、それに学習データの i-vector とコサイン類似度で比較し、類似度が最小となる話者のクリーン音声のスペクトログラム $\mathbf{X}_{tr}$ に対して特異値分解を行い、I 行 J 列の行列 $\mathbf{X}_{tr}$ を

$$\mathbf{X}_{tr} = U\Sigma W^{'} \quad (15)$$

のように I 行 K 列の行列 U、K 行 K 列の行列 Σ、K 行 N 行の行列 $W^{'}$ の内積で表すことが出来る。行列 $\mathbf{T}$ は

$$\mathbf{T} = U\sqrt{\Sigma} \quad (16)$$

となる。しかし行列 $\mathbf{T}$ は負の値を含むので、このまま初期値として扱うことは出来ない。そこで特異値ベクトルの負の成分を正の値に変えて、最後にゼロ成分を分離前の行列の平均値に置き換える非負二重 SVD を適用する。これにより行列 $\mathbf{T}$ は非負の行列となり、初期値として設定することが出来る。

4 音声認識実験

騒音環境下音声認識のタスクで用意された録音データ [12] を対象として、提案法により音声認識性能が改善するか評価を行う。

4.1 実験条件

用いるタスクは、多数の音源が含まれる 4 つの環境 (バス、カフェ、市街地、交差点) において、タブレットに取り付けられたマイクによって録音された音声を認識するタスクである。性能は音声認識性能を計る指標 (単語誤り率 (WER)) で評価する。また、話者が異なる学習セット、開発セット (dt)、評価セット (et) の 3 種類が用意されており、実環境での録音データと仮想環境での録音データが存在する。本稿では dt と et において以下の手法を比較する。

Table 1　音声認識の実験条件

音声認識システム	Kaldi
目的音声の言語	英語
話者	男女各 2 名
各環境毎の発話データ数	410(dt), 330(et)
発話データの総単語数	6780(dt), 5354(et)
音響モデル	GMM

Table 2　MNMF の実験条件

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
分離数	2
チャネル数	2
更新回数	500

1. 未処理のまま音声認識 (Noisy)
2. 重み付き遅延和アレーにより強調 (Baseline)
3. ランダムな初期値による MNMF(Random)
4. バイナリマスクで分離したデータから行列 $\mathbf{H}$ を計算して、MNMF の初期値に設定 (est_H)[9]
5. est_H に加え、評価データに最も近い話者のクリーン音声から行列 $\mathbf{T}$ を計算して、MNMF の初期値に設定 (est_H&T)

ただし、Random に関しては、得られた 2 つの分離音を音声認識させて、WER の良い方を選択した値である。また、Table 1 に音声認識実験の条件を、Table 2 に MNMF の条件を示す。

4.2 実験結果

Table 3 に dt の実環境における実験結果を、Table 4 に dt の仮想環境における実験結果を、Table 5 に et の実環境における実験結果を、Table 6 に et の仮想環境における実験結果を示す。また、「平均」は 4 環境の WER を平均した値を、太文字は WER が一番低かった値を表す。この結果から、Noisy や Baseline、Random と比べて、est_H や est_H&T による初期値設定法の音声認識性能が高い (WER が低い) ことが分かる。しかし、est_H と est_H&T を比較すると、WER の差が小さいことが分かる。

Table 3　dt の実環境における WER[%]

-	バス	カフェ	市街地	交差点	平均
Noisy	27.3	23.1	16.3	22.0	22.2
Baseline	20.1	16.3	12.4	16.9	16.2
Random	33.5	30.7	24.6	27.4	29.0
est_H	**17.4**	16.0	**12.0**	**16.1**	**15.4**
est_H&T	18.1	**15.9**	12.7	**16.1**	15.7

Table 4　dt の仮想環境における WER[%]

-	バス	カフェ	市街地	交差点	平均
Noisy	20.4	29.8	20.5	27.3	24.5
Baseline	16.1	23.6	15.5	21.4	19.2
Random	23.0	32.1	25.2	29.9	27.5
est_H	**13.0**	**19.0**	14.1	18.6	**16.1**
est_H&T	13.4	19.5	**14.0**	**18.0**	16.2

Table 5　et の実環境における WER[%]

-	バス	カフェ	市街地	交差点	平均
Noisy	51.9	39.7	34.0	24.5	37.5
Baseline	39.4	28.4	27.6	20.8	29.0
Random	56.8	44.6	38.6	31.1	42.8
est_H	**37.7**	**26.0**	21.2	**19.2**	**26.0**
est_H&T	38.7	27.1	**20.8**	20.4	26.8

Table 6　et の仮想環境における WER[%]

-	バス	カフェ	市街地	交差点	平均
Noisy	26.7	38.4	34.7	33.5	33.3
Baseline	20.2	31.8	30.0	28.4	27.6
Random	23.4	30.9	31.8	32.7	29.7
est_H	15.3	**23.6**	**22.8**	23.9	**21.4**
est_H&T	**15.0**	25.2	24.6	**23.0**	22.0

4.3　考察

実験結果から、雑音下の音声認識に対して、MNMF の初期値設定における行列 **H** の初期値設定法が有効であることを確認した。しかし、行列 **H** に加えて行列 **T** の初期値設定を行った結果、WER の改善が見られなかった。これは、計算した音声の基底によって、目的音成分が雑音に混ざったり、反対に雑音成分が目的音に混ざったりしてしまった可能性が考えられる。この解決策として、行列 **Z** の値を固定することで、目的音と雑音に対して正しい基底を割り当てることが出来ると考えられる。

5　まとめ

本稿では、バイナリマスクで分離したデータから空間相関行列 **H** を計算する手法と、学習データに非負二重 SVD を適用して基底行列 **T** を計算する手法を用いて、音声認識実験を行った。実験結果から、空間相関行列 **H** を計算した場合は分離性能の改善が見られたが、それに加えて基底行列 **T** を計算した結果は改善が見られなかった。潜在変数行列 **Z** の値を与えることで正しく基底を割り当てることが出来るのではないかと考えられる。

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] 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　春.

[5] C. Fvotte *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.

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

[7] C. Boutsidis, and E. Gallopoulos, “SVD based initialization: A head start for nonnegative matrix factorization,” Pattern Recognition letters, vol. 41, pp. 1350-1362, 2008.

[8] 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　秋.

[9] 三浦 伊織 他: “マルチチャネル非負値行列因子分解を用いた実環境における音声認識” 信学技報, vol. 116, no. 302, EA2016-48, pp. 1-6, 2016 年 11 月.

[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] N. Dehak *et al.*, “ Frontend factor analysis for speaker verification, ” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788?798, 5 2011.

[12] J. Barker *et al.*, “The 4th CHiME Speech Separation and Recognition Challenge”, <http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/>, (accessed 2017-07-12).