非負値テンソル二重逆畳み込みによる残響環境下の劣決定音源分離 *

◎村田直毅 [1], 亀岡弘和 [1,2], 木下慶介 [2], 荒木章子 [2], 中谷智広 [2], 小山翔一 [1], 猿渡洋 [1]
[1] 東京大学 大学院情報理工学系研究科
[2] 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

1 はじめに

複数のマイクロホンで取得した多チャネル信号を処理し，音源の空間情報を手がかりにして音源分離などを行う枠組をマイクロホンアレー信号処理という．近年，ボイスレコーダ，ノートパソコン，スマートフォン，ビデオカメラなどの身の回りにある様々な録音機器による多チャンネル録音を用いたアドホックマイクロホンアレーの研究が盛んに行われている [1]．この枠組は，特殊な装置や配線を要する従来のマイクロホンアレーに比べて手軽かつ安価にアレーシステムを構築できる点で注目されている．従来のアレーシステムでは現在商用化されているものの多くは各マイクロホンが小規模に集中配置されたものに限られているため，チャネル間の微小な時間差が音源分離のための大きな手がかりとなるのに対し，アドホックマイクロホンアレーの枠組ではマイクロホンを広範囲に分散して配置することが容易となるため，チャネル間の強度比も大きな手がかりとなる．

一般に音声信号に残響と雑音が重畳され，観測信号が得られるプロセスを順問題と捉えると，マイクロホンアレーにより目的音声のみを分離抽出する問題は逆問題と見なせる．雑音や室内伝達系の情報が未知の場合でかつマイクロホン数より音源数が多い（劣決定条件の）場合，この逆問題には解が無数に存在しうるため，解を絞り込むための何らかの仮定が必要となる．近年，劣決定条件の音源分離手法として，非負値行列因子分解 (Non-negative Matrix Factorization: NMF) の多チャンネル拡張 [2–5] が有効なアプローチとして注目されている．NMF とは非負値行列を二つの非負値行列（基底行列と係数行列）の積に分解することをいい，スペクトログラムを非負値行列と見なして NMF を適用することはスペクトログラムを低ランクな非負値行列で近似することに相当し，各時刻のスペクトルを基底行列の列数分のスペクトルテンプレートの非負結合で説明しようとしていることを意味する [6]．NMF の多チャンネル拡張は，各音源のパワースペクトログラムにこの構造を仮定した多チャンネル音源分離手法である．なお，過剰決定条件での NMF の多チャンネル拡張も提案されている [7,8]．これらの従来アプローチでは，室内伝達系に時不変性などの制約が置かれ，そのもとで当該逆問題が定式化されるが，アドホックマイクロホンアレーの枠組では手軽にアレーシステムが構築できる利点がある一方で，各マイクロホンの位置は固定されていないため音源・マイクロホンの相対位置関係が録音中に変化してしまいやすいという脆弱性を有している．このように音源・マイクロホンの相対位置関係に変化があった場合，室内伝達系に対する上述の仮定が成立しなくなり，当該仮定のもとで設計されたアルゴリズムは高い性能を発揮できなくなる．

本稿では，音源やマイクロホンの位置の小さな変化に伴う時変残響環境下での音源分離問題の解決を目指し，室内伝達系に関する時変性と時不変性の中間的な強さの仮定に対応する「半時変性」と呼ぶ特性と，音声がもつ局所的な時間周波数構造に着目した音源分離手法を提案する．

2 提案モデル

2.1 時間周波数領域における畳み込み混合モデル

この節では，時間周波数領域における畳み込み混合モデルの定式化を行う．音源からマイクロホンアレーへの伝達系が線形時不変であり，また残響成分が時間周波数解析の窓長内に収まっていると仮定できるとき，マイクロホンアレーで得られる信号は音源の瞬時混合で記述することができる．一方，窓長を超える残響成分が無視できない状況下では，マイクロホンで観測される信号は，以下のように時間周波数領域の畳み込み混合モデル

$$\boldsymbol{y}_{k,l} = \sum_{i} \sum_{n} \boldsymbol{a}_{i,k,n} s_{i,k,l-n} \qquad (1)$$

で近似することができる [9,10]．ここで，i,k はそれぞれ音源と周波数のインデックスを表し，l,n は時間フレームのインデックスを表す．$\boldsymbol{y}_{k,l} \in \mathbb{C}^J$ はマイクロホンアレーで観測される信号ベクトルである．今後，j はマイクロホンのインデックスを表し，J はマイクロホン数を表すものとする．$\boldsymbol{a}_{i,k,n}$ は各音源からマイクロホンへのステアリングベクトルであり，n フ

---

* Non-negative tensor double deconvolution for underdetermined source separation in reverberant environments. by MURATA Naoki[1], KAMEOKA Hirokazu[1,2], KINOSHITA Keisuke[2], ARAKI Shoko[2], NAKATANI Tomohiro[2], KOYAMA Shoichi[1], SARUWATARI Hiroshi[1].
[1]Graduate School of Information Science and Technology, The University of Tokyo, [2]NTT Communication Science Laboratories, NTT Corporation.

レーム遅れて到来する反響成分に対応するものとする．$s_{i,k,l}$ は各音源の時間周波数領域の複素スペクトログラムである．今，音源やマイクロホンの位置が時間変化する場合，ステアリングベクトル $\boldsymbol{a}_{i,k,n}$ は時刻 l に依存し，式 (1) の混合過程は

$$\boldsymbol{y}_{k,l} = \sum_i \sum_n \boldsymbol{a}_{i,k,n,l} s_{i,k,l-n} \tag{2}$$

のような時変系となる．ここで，各音源の複素スペクトログラムが複素ガウス分布に従う，すなわち $s_{i,k,l} \sim \mathcal{N}_{\mathbb{C}}(0, P_{i,k,l})$ を仮定すると（$P_{u,k,l}$ は音源のパワースペクトログラムとする），マイクロホンでの観測信号ベクトルは，下記の分布に従う．

$$\boldsymbol{y}_{k,l} \sim \mathcal{N}_{\mathbb{C}}\Big(\boldsymbol{0}, \sum_i \sum_n P_{i,k,l-n} \boldsymbol{a}_{i,k,n,l} \boldsymbol{a}_{i,k,n,l}^{\mathsf{H}}\Big) \tag{3}$$

ここで，時変ステアリングベクトル $\boldsymbol{a}_{i,k,n,l}$ は

$$\boldsymbol{a}_{i,k,n,l} = \begin{bmatrix} |a_{1,i,k,n,l}| & & 0 \\ & \ddots & \\ 0 & & |a_{J,i,k,n,l}| \end{bmatrix} \begin{bmatrix} e^{j\phi_{1,i,k,n,l}} \\ \vdots \\ e^{j\phi_{J,i,k,n,l}} \end{bmatrix} \tag{4}$$

のように絶対値と偏角の要素に分解することができる．今，ステアリングベクトルの振幅成分が音源やマイクロホンのわずかな移動等による環境の軽微な変化の影響を受けにくいと仮定し，$|a_{J,i,k,n,l}|$ のみが時不変であるような特殊な系を仮定する．このような混合過程を「半時変系」と呼ぶ．すなわち，$|a_{j,i,k,n,l}| = A_{j,i,k,n}$ と仮定すると，式 (4) は

$$\boldsymbol{a}_{i,k,n,l} = \underbrace{\begin{bmatrix} A_{1,i,k,n} & & 0 \\ & \ddots & \\ 0 & & A_{J,i,k,n} \end{bmatrix}}_{\boldsymbol{A}_{i,k,n}} \underbrace{\begin{bmatrix} e^{j\phi_{1,i,k,n,l}} \\ \vdots \\ e^{j\phi_{J,i,k,n,l}} \end{bmatrix}}_{\boldsymbol{\psi}_{i,k,n,l}} \tag{5}$$

と書ける．これを式 (3) に代入すると以下を得る．

$$\boldsymbol{y}_{k,l} \sim \mathcal{N}_{\mathbb{C}}\Big(\boldsymbol{0}, \sum_{i,n} P_{i,k,l-n} \boldsymbol{A}_{i,k,n} \boldsymbol{\psi}_{i,k,n,l} \boldsymbol{\psi}_{i,k,n,l}^{\mathsf{H}} \boldsymbol{A}_{i,k,n}^{\mathsf{H}}\Big) \tag{6}$$

2.2 非負値テンソル二重畳み込みモデル

アドホックマイクロホンアレーでは，アレー素子間のサンプリング周波数のわずかなずれがあったり，音源やマイクロホンのわずかな位置の変化が通常のマイクロホンアレーに比べて起こりやすいため，時変の混合系を扱う必要がある．このような時変の混合系を扱う一つの解決策は，ステアリングベクトルの時間変化量を推定し，補償した後に従来のアレー信号処理を適用することである [11]．一方で，ステアリングベクトルの位相成分を確率的に変動する確率変数と扱う方法も考えられる．[12] では音源の移動に対して頑健な残響除去法を実現するため，[13] では，アドホックマイクロホンアレーにおいてマイクロホン間のサンプリング周波数のミスマッチに対して頑健なアレー信号処理手法を実現するためにこの考え方が採用されている．提案法も後者の考え方に倣い，以下の 2 つを仮定する．

- $\phi_{j,i,k,n,l}$ と $\phi_{j',i,k,n,l'}$ ($j \neq j'$ または $l \neq l'$) は互いに独立である．
- $\phi_{j,i,k,n,l}$ は区間 $[0, 2\pi)$ で一様分布に従う．

このとき $\mathbb{E}[\boldsymbol{\psi}_{\boldsymbol{i,k,n,l}} \boldsymbol{\psi}_{\boldsymbol{i,k,n,l}}^{\mathsf{H}}]$ は単位行列となるので，式 (6) を $\phi_{j,i,k,n,l}$ に関して周辺化すると以下を得る．

$$y_{j,k,l} \sim \mathcal{N}_{\mathbb{C}}\Big(0, \sum_i \sum_n P_{i,k,l-n} A_{j,i,k,n}^2\Big) \tag{7}$$

ここまで，音源 i のパワースペクトログラム $P_{i,k,l}$ については何ら構造を仮定していなかった．NMF の多チャンネル拡張 [2–4] では，$\boldsymbol{P}_i = (P_{i,k,l})_{K \times L}$ を 2 つの非負値行列の積でモデル化している．これは，音源の各時刻におけるパワースペクトルが限られた数のスペクトルテンプレートの非負結合で表されるという仮定に相当する．しかし音声を対象とする場合はこの仮定は必ずしも正確ではない．音声には瞬時瞬時のスペクトルのみならずそのダイナミクス（局所的な時間変化パターン）に大きな特徴があるため，各フレームのスペクトルを要素単位と考えるより，数フレーム分のスペクトルを連結したものを音声を構成する要素単位と見なす方がより音声を特徴付けた表現とすることができると考えられる．そこで，本稿では，スペクトログラム素片（数フレーム分のスペクトルを連結したもの）のテンプレートとアクティベーション系列と畳み込み混合によってパワースペクトログラムをモデル化する，非負値行列因子逆畳み込み (Nonnegative Matrix Factor Deconvolution: NMFD）[14, 15] の考え方を音源のパワースペクトログラムのモデルとして採用する．これは，音源 i のパワースペクトログラムを

$$P_{i,k,l} = \sum_m \sum_\tau W_{i,k,m,\tau} H_{i,m,l-\tau} \tag{8}$$

と置くことに相当する．ここで，m, τ はそれぞれ基底，時間のインデックスを表す．$W_{i,k,m,\tau}$ と $H_{i,m,l}$ はそれぞれスペクトログラム素片テンプレートとそのアクティベーション系列である．なお，このモデルは $\tau = \{0\}$ のとき NMF のモデルに一致する．

以上の提案モデルは残響の重畳過程を表す畳み込み表現と，音源のパワースペクトログラムモデルにおけるスペクトログラム素片テンプレートの畳み込み表現の二つの畳み込みからなるモデルとなっており，残響環境下での音源分離の問題が二重逆畳み込み問題で定式化されている．このことから，提案法を「非負値テンソル二重逆畳み込み (Non-negative Tensor Double Deconvolution: NTDD)」と呼ぶ．

3　パラメータ推定アルゴリズム

式 (7) の対数尤度関数の負値は

$$C_{ML} = \sum_{j,k,l} \log \mathcal{N}_{\mathbb{C}}\Big(y_{j,k,l}|0, \sum_{i,n} P_{i,k,l-n} A^2_{i,j,k,n}\Big) \stackrel{c}{=} \sum_{j,k,l} d_{IS}\Big(|y_{j,k,l}|^2 | \sum_{i,n} P_{i,k,l-n} A^2_{i,j,k,n}\Big) \quad (9)$$

で与えられる．ただし，$d_{IS}(y|x)$ は y と x の板倉斎藤距離を表し，$\stackrel{c}{=}$ は定数項の違いを除いて等しいことを表す．すなわち，最尤推定によるパラメータ推定はマイクロホン j の観測パワースペクトログラム $Y_{j,k,l}$ と，音源 i のスペクトログラムを $P_{i,k,l} = \sum_m \sum_\tau W_{i,k,m,\tau} H_{i,m,l-\tau}$ とした下でのマイクロホン j のパワースペクトログラムモデル $X_{j,k,l} = \sum_{i,n} P_{i,k,l-n} A_{i,j,k,n}$ との板倉斎藤距離の最小化問題に帰着する．

式 (9) を最小化する未知変数 A, W, H を解析的に得ることはできないが，補助関数法 [17] により局所解探索アルゴリズムを導くことができる．また，板倉斎藤距離を内包した，より一般的な乖離度規準である β ダイバージェンス [18] を規準とした場合の局所解探索アルゴリズムも同様に導くこともできる．詳細は省略するが，[18] と同様のアイディアにより Jensen の不等式と接線不等式を用いて補助関数を設計することで以下のような乗法更新式を得ることができる．

$$A_{j,i,k,n} \leftarrow A_{j,i,k,n} \left(\frac{\sum_l Y_{j,k,l} X^{\beta-2}_{j,k,l} P_{i,k,l-n}}{\sum_l X^{\beta-1}_{j,k,l} P_{i,k,l-n}} \right)^{\varphi(\beta)}$$

$$W_{i,k,m,\tau} \leftarrow W_{i,k,m,\tau} \left(\frac{\sum_{j,l,n} Y_{j,k,l} X^{\beta-2}_{j,k,l} A_{j,i,k,l-n} H_{i,m,n-\tau}}{\sum_{j,l,n} X^{\beta-1}_{j,k,l} A_{j,i,k,l-n} H_{i,m,n-\tau}} \right)^{\varphi(\beta)}$$

$$H_{i,m,\tau} \leftarrow H_{i,m,\tau} \left(\frac{\sum_{j,k,l,n} Y_{j,k,l} X^{\beta-2}_{j,k,l} A_{j,i,k,l-n} W_{i,k,m,n-\tau}}{\sum_{j,k,l,n} X^{\beta-1}_{j,k,l} A_{j,i,k,l-n} W_{i,k,m,n-\tau}} \right)^{\varphi(\beta)}$$

ただし，$\varphi(\beta)$ は以下のように定義される．

$$\varphi(\beta) = \begin{cases} 1/(2-\beta) & (\beta < 1) \\ 1 & (1 \le \beta \le 2) \\ 1/(\beta - 1) & (2 < \beta) \end{cases} \quad (10)$$

4　残響環境下の音源分離シミュレーション

提案手法の有効性を確認するため，以下の 2 つの残響環境下の劣決定条件における音源分離実験を行った．まず，残響環境下での頑健さを確認するため，異なる残響の強さを持つ環境での音源分離実験を行った．次に，伝達系への外乱に対する頑健さを確認するため，マイクロフォンの位置が観測信号中に変化した場合における音源分離実験を行った．

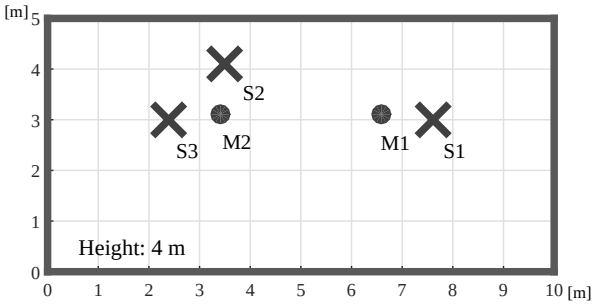

Fig. 1　Room configuration for numerical simulation

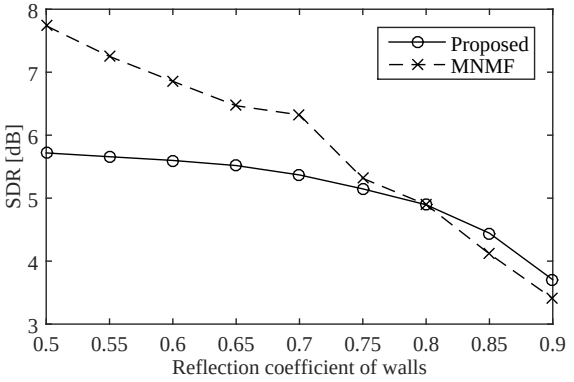

Fig. 2　Relation between SDR and reflection coefficient of walls

劣決定条件として，音源数を 3，マイクロフォン数を 2 とした．与えられた部屋の壁の反射係数，音源位置，マイクロフォン位置から，鏡像法を用いてインパルス応答を生成した．部屋の二次元形状を Fig.1 に示す．「●」(S1∼S3) は音源位置を，「×」(M1，M2) はマイクロフォン位置を表している．音源信号として，ATR 音声対話データベースの，3 話者 15 発話を用いた．音源 1,2 は女性，音源 3 は男性である．残響の強さは，壁の反射係数により変化させた．本実験の環境では，部屋の反射係数が 0.5 の時に 190 ms，0.8 の時に 450 ms の残響時間を持つ．文献 [4] で提案されている多チャンネル NMF（MNMF) と提案手法を比較した．この手法は瞬時混合モデルが仮定されているため，STFT のフレーム外に残響成分が存在する場合は性能が劣化することが予想される．

用意された 15 発話のうち，1 発話を分離用の信号とし，残りの 14 発話を事前の学習に用いた．提案手法の事前学習には NMFD を，MNMF には NMF を用いて，基底スペクトルの学習を行った．各音源に対し，それぞれ 20 個と 40 個の基底を学習した．それぞれ，距離尺度として一般化 KL ダイバージェンスを用いた．STFT のフレーム長を 32 ms とし，シフト長は 16 ms とした．評価指標として，Source-to-distortion ratio（SDR） を用いた．

壁の反射係数の変化による性能の変化を Fig.2 に示

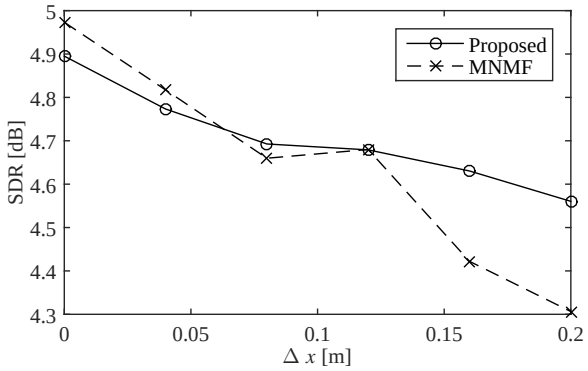

Fig. 3　Relation between SDR and microphone interval change

す．各マイクロフォンにおける各音源の SDR を平均した値を示している．反射係数の低い，瞬時混合モデルがよく成り立っている状況においては，MNMF は提案法に比べて高い分離性能を示すが，一方で，反射係数の高い環境においては，時間周波数領域上での畳込み混合をモデル化している提案法は，残響成分を推定することが可能であるため，MNMF に比べ高い分離性能を示す．

次に，伝達系への外乱に対する頑健さを確認するため，マイクロフォンの位置が観測信号中に変化した場合の音源分離実験を行った．観測信号を，Fig. 1 のマイクロフォン配置を持つ信号と，Δx m マイクロフォンの間隔を増加させた観測信号を繋げることにより生成した．この時，伝達系の振幅成分に比べて，位相成分には大きな外乱が生じることになる．この実験時の部屋の反射係数は 0.8 とした．

伝達系に外乱を加えた観測信号の分離結果を Fig. 3 に示す．MNMF は，伝達系の時不変性を仮定しており，外乱の存在により分離性能が比較的大きく落ちているが，提案法は外乱に対しての分離性能の低下があまり見られず，頑健であると言える．

5　おわりに

本稿では，音源やマイクロホンの位置の小さな変化に伴う時変残響環境下での音源分離問題の解決を目指し，室内伝達系に関する時変性と時不変性の中間的な強さの仮定に対応する「半時変性」と呼ぶ特性と，音声がもつ局所的な時間周波数構造に着目した音源分離手法「非負値テンソル二重逆畳み込み」を提案した．

謝辞 本研究は JSPS 科研費 26730100 の助成を受けたものである．

参考文献

[1] 小野ら “アドホックマイクロホンアレー–複数のモバイル録音機器で行う音響信号処理–,” 信学会 Fundamental Review, 7(4), 336–347, 2014.

[2] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” IEEE TASLP, 18(3), 550–563, 2010.

[3] A. Ozerov et al., “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” in Proc. ICASSP, 257–260, 2011.

[4] H. Sawada et al., “Multichannel extensions of nonnegative matrix factorization with complex-valued data,” IEEE TASLP, 21(5), 971–982, 2013.

[5] T. Higuchi and H. Kameoka, “Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model,” in Proc. MLSP, 2014.

[6] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in Proc. WASPAA, 177–180, 2003.

[7] H. Kameoka et al., “Statistical model of speech signals based on composite autoregressive system with application to blind source separation,” in Proc. LVA/ICA, 245–253, 2010.

[8] D. Kitamura et al., “Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model,” in Proc. ICASSP, 276–280, 2015.

[9] T. Nakatani et al., “Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation,” in Proc. ICASSP, 85–88, 2008.

[10] T. Yoshioka et al., “Blind separation and dereverberation of speech mixtures by joint optimization,” IEEE TASLP, 19(1), 69–84, 2011.

[11] S. Miyabe et al., “Blind compensation of interchannel sampling frequency mismatch with maximum likelihood estimation in STFT domain,” in Proc. ICASSP, 674–678, 2013.

[12] H. Kameoka et al., “Robust speech dereverberation Based on non-negativity and sparse nature of speech spectrograms,” in Proc. ICASSP, 45–48, 2009.

[13] H. Chiba et al., “Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording,” in Proc. IWAENC, 203–207, 2014.

[14] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in Proc. ICA, 494–499, 2004.

[15] P. D. O’grady and B. A. Perlmutter, “Convolutive non-negative matrix factorisation with a sparseness constraint,” in Proc. MLSP, 427–432, 2006.

[16] 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.

[17] D. D. Lee and H. S. Seung, “Algorithms for nonnegative matrix factorization,” in Adv. NIPS, 556–562, 2001.

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