非負値テンソル因子分解に基づく 分散マイクロホンアレイを用いたスポットフォーミング*

☆綾野翔馬 (香川高専), 李莉, 関翔悟 (サイバーエージェント), 北村大地 (香川高専)

1 はじめに

複数の音源の音が混ざりあった信号から特定の音源の信号を抽出する音源分離技術は広く用いられている．例えば，会議において特定の話者の音声のみを取り出すことができれば議事録の作成が容易になる．

音源分離の手法の一つに，複数の同期マイクロホンで構成されるマイクロホンアレイを用いたビームフォーミング (beamforming: BF) があり，遅延和 BF や最小分散無歪応答 (minimum variance distortionless response: MVDR) BF 等が活用される．BF はマイクロホンアレイから見て特定の方位に存在する音源を強調するため，同一方位に複数の音源がある場合はそれらを分離することができない．そこで，Fig. 1 のように複数のマイクロホンアレイを用いて，空間上の特定の領域に存在する目的音源を強調するスポットフォーミングと呼ばれる技術が提案されている [1].

スポットフォーミングはいくつかの手法が提案されており，マイクロホンアレイの配置を最適化する手法 [2] や，分散して置かれたマイクロホンアレイ内の全マイクで空間フィルタを設計する手法 [1] がある．しかしながら，前者はマイクロホンアレイが容易に移動可能でなければならない点，後者はマイクロホンアレイ間の完全な同期が必要な点において大がかりなシステムが要求される．そこで，非負値行列因子分解 (nonnegative matrix factorization: NMF) [3] を用いた，マイクロホンアレイ同士の同期の条件が緩和されたスポットフォーミング [4] が提案された．本論文では，分散マイクロホンアレイを用いたスポットフォーミングとして，非負値テンソル因子分解 (nonnegative tensor factorization: NTF) を用いた手法を提案する．

2 分散マイクロホンアレイを用いたスポットフォーミング

2.1 想定する状況

本稿では Fig. 1 のように目的音源と干渉音源が存在し，また複数のマイクロホンアレイが分散して置かれている状況を考える．ここで，各マイクロホンアレイ内では同期録音しているが，マイクロホンアレイ間では非同期である状況を想定している．Fig. 1 の場合，各マイクロホンアレイは正面方向に BF を行うことで，目的音源を強調できるが，同時に同一方位上に存在する干渉音源も強調されてしまう．Fig. 1 のような状況では，この干渉音源がマイクロホンアレイ毎に異なるため，各 BF の出力信号間で共通する音源成分を抽出できれば，目的音源のみの信号が得られると考えられる．

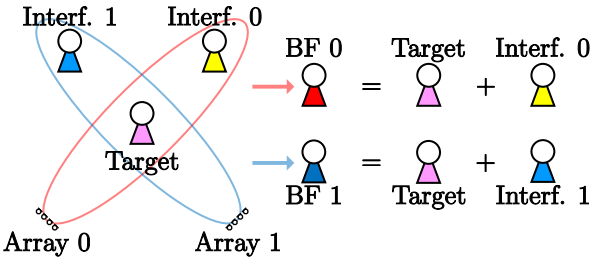

Fig. 1 Situations and signals estimated by two beamformers.

2.2 信号モデル

本稿では，文字 D について，D を 3 階テンソル，$\boldsymbol{D}$ を行列，$\mathbf{D}$ をベクトルとし，(a,b,c) 要素を $D_{a,b,c}$ で表す．また信号については，小文字と大文字でそれぞれ時間領域と時間周波数領域を表す．マイクロホンアレイのインデクスを $a = 0,1,\cdots,A-1$，マイクロホンアレイ内のマイクロホンのインデクスを $m = 0,1,\cdots,M-1$，離散時間インデクスを $t = 0,1,\cdots,T-1$ とする．a 番目の各マイクロホンアレイの M 個のマイクの観測信号 $\boldsymbol{x}^{(a)} \in \mathbb{R}^{M\times T}$ に対して，方位 θ_a を強調する任意の BF $f_{\theta_a} : \mathbb{R}^{M\times T} \to \mathbb{R}^{T}$ を適用する処理を次式で表す．

$$\mathbf{y}^{(a)} = f_{\theta_a}(\boldsymbol{x}^{(a)}) \tag{1}$$

BF の出力 $\mathbf{y}^{(a)} \in \mathbb{R}^{T}$ に対して短時間 Fourier 変換 (short-time Fourier transform: STFT) を行い，スペクトログラム $\boldsymbol{Y}^{(a)} \in \mathbb{C}^{I\times J}$ を得る．ここで，$i = 0,1,\cdots,I-1$ は周波数ビン，$j = 0,1,\cdots,J-1$ は時間フレームである．

2.3 従来法

文献 [4] の手法では，BF の出力 $Y^{(a)}_{i,j}$ のマイクロホンアレイ方向と時間フレーム方向を結合して 1 次元とし，時間周波数マスクを作成することでスポットフォーミングを行う．結合した振幅スペクトログラムを $\boldsymbol{C}^{(\mathrm{conv})} \in \mathbb{R}^{I\times N}_{\geq 0}$ とすると次式で与えられる．

$$C^{(\mathrm{conv})}_{i,n} := C^{(\mathrm{conv})}_{i,aJ+j} = \left|Y^{(a)}_{i,j}\right| \tag{2}$$

ここで，$N = AJ$ であり，$n = 0,1,\cdots,N-1$ は $\boldsymbol{C}^{(\mathrm{conv})}$ の列インデクスである．さらに，この行列 $\boldsymbol{C}^{(\mathrm{conv})}$ に対して次式のように NMF を適用する．

$$C^{(\mathrm{conv})}_{i,n} \simeq \textstyle\sum_k T_{i,k}\tilde{V}_{n,k} \quad (\boldsymbol{C}^{(\mathrm{conv})} \simeq \boldsymbol{T}\tilde{\boldsymbol{V}}^{\mathrm{T}}) \tag{3}$$

*Spotforming using distributed microphone array based on nonnegative tensor factorization. By Shoma AYANO (NIT Kagawa), Li LI, Shogo SEKI (CyberAgent, Inc.), and Daichi KITAMURA (NIT Kagawa).

ここで，$\boldsymbol{T} \in \{\boldsymbol{T} \in [0,1]^{I\times K} | \sum_i T_{i,k} = 1\}$ 及び $\tilde{\boldsymbol{V}} \in \mathbb{R}_{\geq 0}^{N\times K}$ は基底行列及びアクティベーション行列であり，$k = 0, 1, \cdots, K-1$ は NMF の基底ベクトルのインデクスである．NMF の目的関数は次式となる．

$$\begin{aligned} &\underset{\boldsymbol{T},\tilde{\boldsymbol{V}}}{\text{minimize}} \ \sum_{i,n} \mathcal{D}\left(C_{i,n}^{(\text{conv})} \middle| \sum_k T_{i,k}\tilde{V}_{n,k}\right) \\ &\text{s.t. } T_{i,k}, \tilde{V}_{n,k} \geq 0 \ \forall i,k,n \end{aligned} \tag{4}$$

文献 [4] では，距離関数 $\mathcal{D}$ に Euclid 距離を使用している．式 (4) を最小化する基底行列 $\boldsymbol{T}$ 及びアクティベーション行列 $\tilde{\boldsymbol{V}}$ を求めた後，$\tilde{\boldsymbol{V}}$ を用いてバイナリマスク行列 $\tilde{\boldsymbol{H}} \in \{0,1\}^{J\times K}$ を次式のように作成する．

$$\tilde{H}_{j,k} = \begin{cases} 1 \ (\text{if } \tilde{V}_{aJ+j,k} > \mu \ \forall a) \\ 0 \ (\text{o/w}) \end{cases} \tag{5}$$

ここで，$\mu \in \mathbb{R}_{\geq 0}$ は閾値である．すなわち，このバイナリマスクはすべてのマイクロホンアレイで μ より大きいアクティベーションを持つ成分を共通の目的音源成分とみなし 1 としている．このバイナリマスクを用いて目的音源の推定振幅スペクトログラム $\boldsymbol{E}^{(a)(\text{conv})} \in \mathbb{R}_{\geq 0}^{I\times J}$ を次式で得る．

$$E_{i,j}^{(a)(\text{conv})} = \sum_k T_{i,k}\tilde{H}_{j,k}\tilde{V}_{aJ+j,k} \tag{6}$$

これは各マイクロホンアレイの信号に対して同じ時間周波数バイナリマスク $\tilde{\boldsymbol{H}}$ を適用していることに相当する．最後に，各マイクロホンアレイの $\boldsymbol{E}^{(a)(\text{conv})}$ に対して位相を与えて逆 STFT を行い，時間ズレを補正して和をとることで目的音源 $\mathbf{e}^{(\text{conv})} \in \mathbb{R}^T$ を得る．

従来法では，NMF における基底ベクトル数 K を観測信号に応じて適切に調整しなければならない．また，式 (5) のバイナリマスクは，マイクロホンアレイ間の時間フレームを跨ぐ（STFT のシフト長を超える）ほどの時間ずれは考慮されていないため，許容できる非同期性（録音開始時刻のずれ及びサンプリング周波数のずれ）の条件がやや厳しい．そこで，次章では NTF を用いることでこれらの制約を緩和した提案法を説明する．

3 提案法

3.1 非負値テンソル表現を用いたモデル化の利点

提案法は，各 BF の出力信号の振幅スペクトログラムで 3 階テンソル $\mathsf{C}^{(\text{prop})} \in \mathbb{R}_{\geq 0}^{A\times I\times J}$ を構成する．

$$C_{a,i,j}^{(\text{prop})} := \left|Y_{i,j}^{(a)}\right| \tag{7}$$

この 3 階テンソル $\mathsf{C}^{(\text{prop})}$ は式 (2) と異なり，マイクロホンアレイ及び時間フレームの物理的な次元を維持している．提案法は，この $\mathsf{C}^{(\text{prop})}$ に対して次節の NTF を適用し，分配行列，基底行列，アクティベーション行列の 3 要素に分解する．K 個の基底ベクトルは分配行列によって自動的に各マイクロホンアレイに振り分けられるため，観測信号の性質に強く依存せず適切に共通成分を抽出できることが期待される．さらに，分配行列を用いて時間フレームに非依存なバイナリマスクを構成できるため，原理的には完全に非同期な A 個のマイクロホンアレイでもスポットフォーミングが可能となる．

3.2 NTF に基づくスポットフォーミング

提案法の処理の流れを Fig. 2 に示す．提案法では，式 (7) で定義される観測非負値テンソル $\mathsf{C}^{(\text{prop})}$ に対して NTF を適用する．このとき次式のように，分配行列 $\boldsymbol{Z} \in \{\boldsymbol{Z} \in [0,1]^{A\times K} | \sum_a Z_{a,k} = 1\}$，基底行列 $\boldsymbol{T}$，及びアクティベーション行列 $\boldsymbol{V} \in \mathbb{R}_{\geq 0}^{J\times K}$ の積に分解する．

$$C_{a,i,j}^{(\text{prop})} \simeq \sum_k Z_{a,k}T_{i,k}V_{j,k} \tag{8}$$

この NTF における $\boldsymbol{Z}$ は，K 個の基底ベクトルを A 個のマイクロホンアレイに分配する役割を持つ．この NTF の目的関数は次式となる．

$$\begin{aligned} &\underset{\boldsymbol{Z},\boldsymbol{T},\boldsymbol{V}}{\text{minimize}} \ \sum_{a,i,j} \mathcal{D}\left(C_{a,i,j}^{(\text{prop})} | \sum_k Z_{a,k}T_{i,k}V_{j,k}\right) \\ &\text{s.t. } Z_{a,k}, T_{i,k}, V_{j,k} \geq 0 \ \forall a,i,j,k \end{aligned} \tag{9}$$

次節に示す反復更新式で $\boldsymbol{Z}$，$\boldsymbol{T}$，$\boldsymbol{V}$ を求めた後，$\boldsymbol{Z}$ を用いてバイナリマスク $\mathbf{H} \in \{0,1\}^K$ を構成する．まず，$\alpha_k = \min\{Z_{0,k}, Z_{1,k}, \cdots, Z_{A-1,k}\}$ を各 k について計算し，$\alpha_0, \alpha_1, \cdots, \alpha_{K-1}$ を降順に並べた際の先頭 γ 個のインデクスの集合 $\Gamma \in \{\Gamma \subseteq \{k \in \mathbb{Z} | 0 \leq k < K\} | n(\Gamma) = \gamma\}$ を求める．ここで，$n(\cdot)$ は．集合 $\cdot$ の要素数を表す．$\mathbf{H}$ は Γ を用いて次式で表される．

$$H_k = \begin{cases} 1 \ (\text{if } k \in \Gamma) \\ 0 \ (\text{o/w}) \end{cases} \tag{10}$$

すなわち，k 番目の基底ベクトルは α_k が大きいほど全てのマイクロホンアレイで共通に使用されることを表すため，それらの基底ベクトルが目的音源成分であるとみなすバイナリマスクを構成している．式 (10) では α_k が大きい順に γ 個の基底ベクトルを目的音源成分として選択している．

このバイナリマスクを用いて目的音源を推定する Wiener フィルタを構成する．この処理は，提案法による目的音源の推定複素スペクトログラムを $\boldsymbol{E}^{(a)(\text{prop})} \in \mathbb{C}^{I\times J}$ とすると次式となる．

$$E_{i,j}^{(a)(\text{prop})} = \frac{\sum_k (H_k Z_{a,k}T_{i,k}V_{j,k})^2}{\sum_k (Z_{a,k}T_{i,k}V_{j,k})^2} Y_{i,j}^{(a)} \tag{11}$$

最後に，各マイクロホンアレイの $\boldsymbol{E}^{(a)(\text{prop})}$ に逆 STFT を適用し，時間ズレを補正して和をとることで目的音源 $\mathbf{e}^{(\text{prop})} \in \mathbb{R}^T$ を得る．

3.3 NTF における各変数の反復更新式の導出

本節では，文献 [3] の補助関数法を用いて式 (9) の最適化問題を解く反復更新式を導出する．目的関数 $\mathcal{D}$

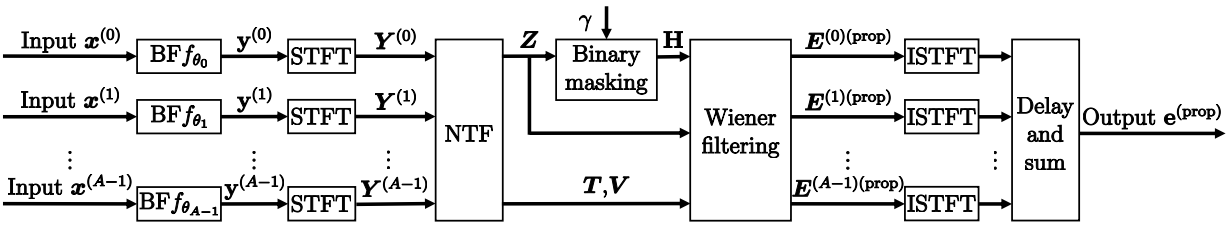

Fig. 2　Process flow of the proposed spotforming method.

には次式で表される定数項を除いた一般化 Kullback–Leibler 擬距離を用いる．

$$\mathcal{D}(y|x) = -y \ln x + x \tag{12}$$

全体の目的関数を新たに $\mathcal{G}$ と定義する．

$$\mathcal{G} = \textstyle\sum_{a,i,j} \mathcal{D}(C^{(\mathrm{prop})}_{a,i,j} | \sum_k Z_{a,k} T_{i,k} V_{j,k}) \tag{13}$$

補助関数法を適用するために補助変数 $\mathsf{U} \in \{U_{a,i,j,k} \in [0,1]^{A\times I\times J\times K} | \sum_{a,i,j,k} U_{a,i,j,k} = 1\}$ を導入する．Jensen の不等式を適用して得られる補助関数を $\mathcal{G}$ とすると次式となる．

$$\begin{aligned}\mathcal{G} &\leq \mathcal{G}^{+} \\ &= \textstyle\sum_{a,i,j} \left[-C^{(\mathrm{prop})}_{a,i,j} \sum_k U_{a,i,j,k} \log \left(\frac{Z_{a,k} T_{i,k} V_{j,k}}{U_{a,i,j,k}} \right) \right. \\ &\quad \left. + \textstyle\sum_k Z_{a,k} T_{i,k} V_{j,k} \right]\end{aligned} \tag{14}$$

$\partial \mathcal{G}^{+} / \partial Z_{a,k} = 0$ より次式を得る．

$$\textstyle\sum_{i,j} \left\{ -C^{(\mathrm{prop})}_{a,i,j} U_{a,i,j,k} \frac{1}{Z_{a,k}} + T_{i,k} V_{j,k} \right\} = 0 \tag{15}$$

これを $Z_{a,k}$ について解くと，次式となる．

$$Z_{a,k} = \frac{\sum_{i,j} C^{(\mathrm{prop})}_{a,i,j} U_{a,i,j,k}}{\sum_{i,j} T_{i,k} V_{j,k}} \tag{16}$$

補助変数 U に $\mathcal{G}^{+} = \mathcal{G}$ となる値（補助変数の等号成立条件）を代入することで，次の反復更新式を得る．

$$Z_{a,k} \leftarrow Z_{a,k} \frac{\sum_{i,j} C^{(\mathrm{prop})}_{a,i,j} \frac{T_{i,k} V_{j,k}}{\sum_l Z_{a,l} T_{i,l} V_{j,l}}}{\sum_{i,j} T_{i,k} V_{j,k}} \tag{17}$$

同様にして，$T_{i,k}$ 及び $V_{j,k}$ の反復更新式を得る．

$$T_{i,k} \leftarrow T_{i,k} \frac{\sum_{a,j} C^{(\mathrm{prop})}_{a,i,j} \frac{Z_{a,k} V_{j,k}}{\sum_l Z_{a,l} T_{i,l} V_{j,l}}}{\sum_{a,j} Z_{a,k} V_{j,k}} \tag{18}$$

$$V_{j,k} \leftarrow V_{j,k} \frac{\sum_{a,i} C^{(\mathrm{prop})}_{a,i,j} \frac{Z_{a,k} T_{i,k}}{\sum_l Z_{a,l} T_{i,l} V_{j,l}}}{\sum_{a,i} Z_{a,k} T_{i,k}} \tag{19}$$

従って，$\boldsymbol{Z}$，$\boldsymbol{T}$，$\boldsymbol{V}$ を初期化し，式 (17)–(19) を反復更新することで，$\boldsymbol{Z}$，$\boldsymbol{T}$，$\boldsymbol{V}$ が推定される．

4　実験

4.1　実験条件

本稿では，Pyroomacoustics [5] を用いたシミュレーションにより従来法及び提案法のスポットフォーミングの性能を比較する．部屋の形状と音源及びマイクロホンアレイの配置条件を Fig. 3 に示す．Target は目的音源，各 Interf. は干渉音源を表す．ここでは，壁面反射のみを考慮する 2 次元の鏡像法を用いている．壁面の反射回数と反射係数を調節し，$T_{60} = 0$ ms 及び $T_{60} = 512$ ms の 2 種類の部屋を作成した．各音源のドライソースには Table 1 に示す LibriTTS [6] の開発データセット (クリーン) の一部を利用した．サンプリング周波数は 16 kHz とした．

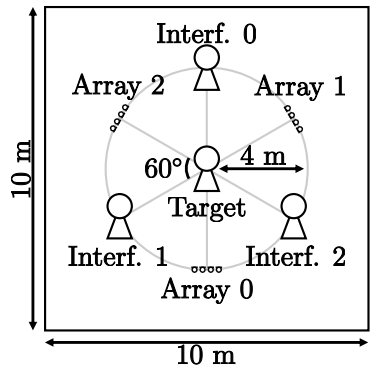

Fig. 3　Condition of room, source, and microphone array locations.

Table 1　Dry sources

File name	Source	Duration [s]
84_121123_000008_000002.wav	Target	3.38
652_130737_000012_000000.wav	Interf. 0	4.05
3000_15664_000020_000005.wav	Interf. 1	4.07
1272_141231_000024_000005.wav	Interf. 2	3.47

各マイクロホンアレイは 4 個のマイクロホンを間隔 2.83 cm で等間隔に配置して構成した．すべてのマイクロホンアレイは完全同期とし，BF には MVDR を用いた．ただし，理想的な条件として MVDR の目的音源のステアリングベクトルには Target の直接音のインパルス応答を用い，雑音分散共分散行列の計算には全 Interf. からの信号の和を用いた．STFT の窓長及びシフト長はそれぞれ 32 ms 及び 16 ms とし，窓関数には Hann 窓を用いた．NMF 及び NTF の基底ベクトル数は $K = 30$，反復更新回数は 100 回とした．なお本稿では，条件を揃えるために，従来法では NMF の距離関数に式 (12) を用いた反復更新式を採用し，さらにバイナリマスクを作成した後は提案法と同様の Wiener フィルタを適用した．従来法の $\boldsymbol{T}$ 及び $\tilde{\boldsymbol{V}}$ と提案法の $\boldsymbol{T}$ 及び $\boldsymbol{V}$ の初期値は区間 (0, 1) の一様分布乱数を用い，提案法の $\boldsymbol{Z}$ の初期値は全て 0.5 とした．客観評価値には，10 種類の乱数初期

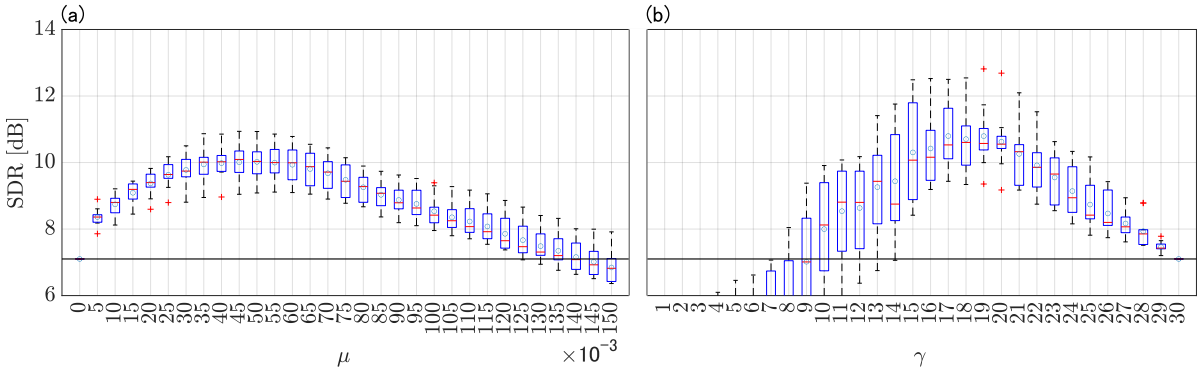

Fig. 4　Box plots of SDR values with 10 random initializations: (a) conventional and (b) proposed methods when $T_{60} = 0$ ms.

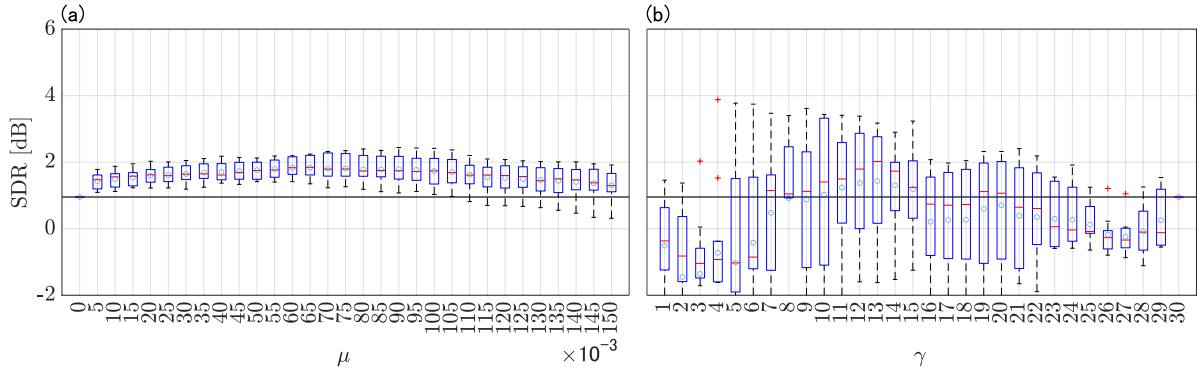

Fig. 5　Box plots of SDR values with 10 random initializations: (a) conventional and (b) proposed methods when $T_{60} = 512$ ms.

値に対する信号対歪み比 (source-to-distortion ratio: SDR) [7] を用いた．

4.2　実験結果

SDR の箱ひげ図を Figs. 4 及び 5 に示す．前者は $T_{60} = 0$ ms，後者は $T_{60} = 512$ ms の結果である．青い箱は四分位範囲，箱の内部の赤線及び青丸はそれぞれ中央値及び平均値，破線は外れ値を除く値の範囲，赤い十字は外れ値を表す．また，黒の水平線は BF 出力の SDR を示す．

両方の残響条件で提案法の SDR の中央値が従来法を上回っていることが分かる．一方で，従来法は μ による SDR のばらつきが小さいのに対し，提案法は同じ γ を用いても SDR のばらつきが大きくなっている．特に $T_{60} = 512$ ms の場合では，提案法のばらつきが顕著である．従って，提案法は NTF の各変数に適切な初期値を与えること，適切な γ を選択することができれば，高い性能が得られると考えられる．

5　おわりに

本稿では，分散マイクロホンアレイと NTF を用いた新しいスポットフォーミング手法を提案した．提案法は NTF の初期値によって性能のばらつきが大きいものの，中央値では従来法を上回ることを確認した．今後の課題として NTF に適切な初期値を与える方法の検討，最良の γ を自動決定する手法等が挙げられる．また，非同期の実録音信号を用いて，実際の応用における性能についても調査する．

参考文献

[1] M. Taseska and E. A. P. Habets, “Spotforming: spatial filtering with distributed arrays for position-selective sound acquisition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1291–1304, 2016.

[2] K. Sekiguchi, Y. Bando, K. Itoyama, and K. Yoshii, “Layout optimization of cooperative distributed microphone arrays based on estimation of source separation performance,” *J. Robotics and Mechatronics*, vol. 29, no. 1, pp.83–93, 2017.

[3] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Proc. NeurIPS*, pp. 556–562, 2000.

[4] Y. Kagimoto, K. Itoyama, K. Nishida, and K. Nakadai, “Spotforming by NMF using multiple microphone arrays,” *Proc. IROS*, pp. 9253–9258, 2022.

[5] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: a Python package for audio room simulation and array processing algorithms,” *Proc. ICASSP*, pp. 351–355, 2018.

[6] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: a corpus derived from librispeech for text-to-speech,” *Proc. Interspeech*, pp. 1526–1530, 2019.

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