時間チャネル非負値行列因子分解を用いた被り音抑圧における初期値頑健性の比較 *

☆溝渕悠朔, 北村大地 (香川高専), 中村友彦, 猿渡洋 (東大), 高橋祐, 近藤多伸 (ヤマハ)

1 はじめに

音楽の生演奏を録音する場合，各音源に近接するようにマイクロホンを配置することが一般的である．この配置は，各音源のみの音響信号（以後，目的音源と呼ぶ）を得ることを目的としているが，実際には他の音源からの音も少なからず混入してしまう．これは一般に被り音と呼ばれており，ミキシングの品質や演奏の質を下げる原因の 1 つである．

複数のマイクロホンで観測された信号に対する被り音の抑圧を行う手法として，Togami らは，多チャネル観測信号の周波数毎の時間チャネル行列に非負値行列因子分解（nonnegative matrix factorization: NMF）[1] を適用する時間チャネル NMF（time-channel NMF: TCNMF）を提案している [2]．TCNMF では，周波数ビン毎に非負混合行列と各音源のアクティベーションを推定しており，自明解を避ける正則化の導入（以後，従来 TCNMF と呼ぶ）によって，マイクロホン間隔が空間的に離れている（空間エイリアシング問題が生じる）観測信号に対しても音源を分離できることが確認されている [3]．この利点が被り音抑圧においても有効であることに着目し，我々は被り音の相対的なゲインに事前分布を導入する新しい手法（以後，提案 TCNMF と呼ぶ）を提案した [5]．

提案 TCNMF はその定式化の都合上，従来 TCNMF よりも最適化変数の初期値に対して被り音抑圧の性能が頑健となることが予想される．本稿では，従来 TCNMF と提案 TCNMF のそれぞれにおいて，最適化変数の初期値に対する被り音抑圧性能の頑健性を実験的に調査する．

2 TCNMF の定式化と最適化問題

2.1 従来 TCNMF

TCNMF は時間周波数の振幅値のみを用いる音源分離手法であり，文献 [2–4] では音声強調や音声分離に適用されている．TCNMF は周波数毎の時間チャネル行列 $\boldsymbol{X}_i$ を次のように分解する．

$$\boldsymbol{X}_i \approx \boldsymbol{A}_i \boldsymbol{S}_i \quad \forall i \tag{1}$$

ここで，$\boldsymbol{X}_i = [\boldsymbol{x}_{i1} \ \cdots \ \boldsymbol{x}_{ij} \ \cdots \ \boldsymbol{x}_{iJ}] \in \mathbb{R}_{\geq 0}^{M \times J}$ は観測信号の周波数 i における全チャネル全時間の振幅値をまとめた行列，$\boldsymbol{A}_i \in \mathbb{R}_{\geq 0}^{M \times N}$ は周波数 i における各音源から各マイクへのゲインをまとめたゲイン行列，及び $\boldsymbol{S}_i = [\boldsymbol{s}_{i1} \ \cdots \ \boldsymbol{s}_{ij} \ \cdots \ \boldsymbol{s}_{iJ}] \in \mathbb{R}_{\geq 0}^{N \times J}$ は周波数 i における音源毎の全時間の振幅値をまとめた行列である．なお，i，j，m，及び n をそれぞれ周波数，時間，チャネル，及び音源の整数インデクスとし，これらの最大値を I，J，M，及び N と表す．従って，式 (1) のように観測信号 $\{\boldsymbol{X}_i\}_{i=1}^{I}$ から音源信号 $\{\boldsymbol{S}_i\}_{i=1}^{I}$ を推定できれば，音源分離や被り音抑圧が可能となる．

$\boldsymbol{A}_i$ 及び $\boldsymbol{S}_i$ の推定は，次の最適化問題となる．

$$\begin{aligned} \min_{\mathcal{A},\mathcal{S}} \sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i) + \mu \sum_{i,j} \|\boldsymbol{s}_{ij}\|_{0.5} \\ \text{s.t. } a_{imn}, s_{inj} \geq 0 \quad \forall i,m,n,j \end{aligned} \tag{2}$$

ここで，$\mathcal{D}_{\mathrm{KL}}(\cdot|\cdot)$ は次式のように定義される．

$$\begin{aligned} \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i) = \sum_{m,j} \Bigg(x_{imj} \log \frac{x_{imj}}{\sum_n a_{imn} s_{inj}} \\ - x_{imj} + \sum_n a_{imn} s_{inj} \Bigg) \end{aligned} \tag{3}$$

また，$\mathcal{A}$ 及び $\mathcal{S}$ はそれぞれ $\{\boldsymbol{A}_i\}_{i=1}^{I}$ 及び $\{\boldsymbol{S}_i\}_{i=1}^{I}$ の集合，x_{imj}，a_{imn}，及び s_{inj} はそれぞれ $\boldsymbol{X}_i$，$\boldsymbol{A}_i$，及び $\boldsymbol{S}_i$ の要素，μ は正則化のための重み係数，$\|\cdot\|_{0.5}$ は $L_{0.5}$ ノルムである．

式 (2) の第 2 項はスパース正則化項であり，決定系（$M = N$）における自明解（全ての i に対して $\boldsymbol{A}_i$ が単位行列）の回避のために導入される [2]．これは，時間周波数領域における W-disjoint-orthogonality [7] 仮定に基づいており，音声の混合信号には妥当だが，時間周波数領域で複数の音源成分が重なり合う音楽信号では成立しづらい．従って，従来 TCNMF は，音楽信号の場合には音質を低下させる可能性がある．

2.2 提案 TCNMF

提案 TCNMF は，従来 TCNMF を音楽信号の被り音抑圧のために改良した手法である [5]．本手法では，ゲイン行列 $\boldsymbol{A}_i$ に次の事前分布生成モデルを導入する．

$$a_{imn} \sim \begin{cases} \delta(a_{imn} - 1) & (m = n) \\ \mathcal{G}(a_{imn}; k, \theta) & (m \neq n) \end{cases} \tag{4}$$

$$\mathcal{G}(a; k, \theta) = \frac{1}{\Gamma(k)\theta^k} a^{k-1} e^{-a/\theta} \tag{5}$$

ここで，$\delta(a)$ は Dirac のデルタ関数であり，$\boldsymbol{A}_i$ の対角成分を 1 に固定する役割を持つ．また $\mathcal{G}(a; k, \theta)$ は確率変数 $a \geq 0$，形状母数 $k > 0$，及び尺度母数 $\theta > 0$ から成るガンマ分布であり，形状母数を $k > 1$ とすると，$\boldsymbol{A}_i$ の非対角要素が 0 になることを防ぐことができ，$\boldsymbol{A}_i$ の自明解を回避できる．一方，$\boldsymbol{S}_i$ には

* Robustness comparison of initialization for bleeding sound reduction using time-channel nonnegative matrix factorization. By Yusaku MIZOBUCHI, Daichi KITAMURA (NIT Kagawa), Tomohiko NAKAMURA, Hiroshi SARUWATARI (The University of Tokyo), Yu TAKAHASHI, and Kazunobu KONDO (Yamaha Corporation).

明示的な構造を仮定せず，非負制約事前分布のみを導入する．このガンマ事前分布の導入は，Cemgil の Bayesian NMF [6] に基づいている．

これらの事前分布を用いて $\boldsymbol{A}_i$ と $\boldsymbol{S}_i$ を最大事後確率推定で求めることは，次の問題と等価である．

$$\min_{\mathcal{A},\mathcal{S}} \sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i) + \sum_{i,m,n\neq m} \mathcal{R}(a_{imn};k,\theta)$$
$$\text{s.t. } a_{imn}, s_{inj} \geq 0 \ \forall i,m,n,j \text{ and } a_{imn} = 1 \ \forall m = n \tag{6}$$

ここで，正則化項 $\mathcal{R}(a_{imn};k,\theta)$ は次式となる．

$$\mathcal{R}(a_{imn};k,\theta) = \left[-(k-1)\log a_{imn} + \frac{1}{\theta}a_{imn}\right] \tag{7}$$

また，本手法では，観測信号に次の前処理を施す．

$$\tilde{\boldsymbol{x}}[t] \leftarrow \frac{\alpha}{v}\tilde{\boldsymbol{x}}[t] \quad \forall t \tag{8}$$
$$v = \max\left(\{\mathrm{abs}(\tilde{\boldsymbol{x}}[t])\}_{t=1}^{T}\right) \tag{9}$$

ここで，$\tilde{\boldsymbol{x}}[t] \in \mathbb{R}^M$ は時間領域の多チャネル観測信号，t は離散時間インデクス（最大値は T），max(·) は入力集合の最大値，ベクトルに対する abs(·) は要素毎の絶対値である．信号のダイナミックレンジに相当する $\alpha \geq 0$ は正則加項の重み係数と等価となる [5]．

2.3 最適化アルゴリズム

従来 TCNMF と提案 TCNMF は，補助関数法 [1] による反復最適化アルゴリズムが提案されている [2,5]．これは，$\boldsymbol{A}_i$ 及び $\boldsymbol{S}_i$ を非負乱数で初期化したうえで反復することで，最適化コスト関数が単調に減少（又は非増加）するアルゴリズムである．式 (2) 及び式 (6) はいずれも非凸最適化問題であるため，アルゴリズムの収束点として求まる解は通常，最適化変数の初期値に依存して異なる結果になることが一般的である．

3 最適化変数の初期値頑健性の調査

3.1 動機

TCNMF のコスト関数には，式 (3) の一般化 Kullback-Leibler（KL）ダイバージェンスが用いられている．KL ダイバージェンスに基づく NMF（KL-NMF）では，2 変数 $\boldsymbol{A}_i$ 及び $\boldsymbol{S}_i$ のいずれか片方を定数とみなすと凸最適化となりユニークな解が存在するという良い性質がある [1]．しかし，従来 TCNMF は式 (2) のように $L_{0.5}$ ノルムの正則化項を追加したことで，この性質は失われている．一方提案 TCNMF は，式 (4) に示すように，KLNMF の生成モデル（ポアソン分布）の共役事前分布（ガンマ分布）を仮定しているため，前述の性質が引き継がれており，式 (6) の最適化問題も 2 変数のいずれかが定数であれば，凸最適化となると考えられる．さらに，提案 TCNMF では式 (4) の事前分布により，変数 $\boldsymbol{A}_i$ の対角成分を 1 に固定していることから，初期値頑健性が大幅に改善されることが予想できる．本章では，上記の予想を検証するために，従来 TCNMF と提案 TCNMF のそれぞれにおいて，最適化変数の初期値に対する被り音抑圧性能の頑健性を実験的に調査する．

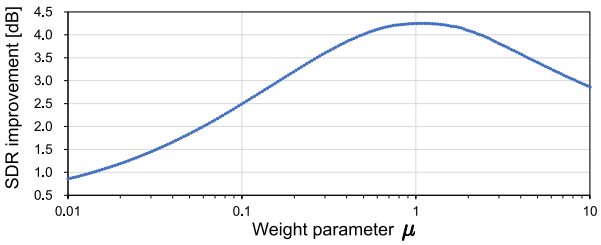

Fig. 1 SDR improvements for simulated training data obtained by conventional TCNMF.

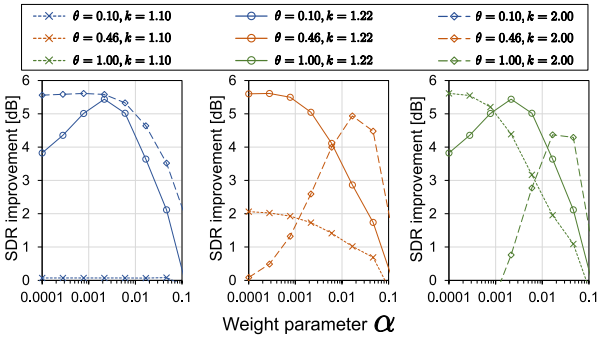

Fig. 2 SDR improvements for simulated training data obtained by proposed TCNMF.

3.2 シミュレーション信号の被り音抑圧実験

3.2.1 実験条件

ドライソースには，実楽器の音楽データセット *DSD100* [8] からランダムに 20 曲を選出し，開発データ 10 曲とテストデータ 10 曲に分割して，60〜90 秒の範囲でトリミングした信号を用いた．混合前の音源信号は，vocals，bass，drums，other の 4 種類とし，$M = N = 4$ となるように観測信号の複素スペクトログラム $\mathbf{X} \in \mathbb{C}^{I\times J\times M}$ を作成した．このとき，被り音を含む観測信号の模擬のため，周波数毎の非負混合行列 $\overline{\boldsymbol{A}_i} \in \mathbb{R}_{\geq 0}^{M\times N}$ を用いて，音源信号の複素スペクトログラム $\mathbf{S} \in \mathbb{C}^{I\times N\times J}$ を次式のように混合した．

$$\mathbf{x}_{ij} = \overline{\boldsymbol{A}_i}\mathbf{s}_{ij} \tag{10}$$

ここで，$\mathbf{x}_{ij} \in \mathbb{C}^M$ 及び $\mathbf{s}_{ij} \in \mathbb{C}^N$ はそれぞれ，$\mathbf{X}$ 及び $\mathbf{S}$ の周波数 i，時間 j におけるチャネルベクトル及び音源ベクトルである．$\overline{\boldsymbol{A}_i}$ の対角要素は 1 とし，非対角要素は (0, 0.2) の範囲の一様分布から生成される乱数に設定した．但し，開発データには 10 種類の異なる擬似乱数を用い，10 種類の観測信号を作成した．最適化変数 $\boldsymbol{A}_i$ 及び $\boldsymbol{S}_i$ の初期値は別の 10 種類の擬似乱数を用いて 10 種類作成した．信号の標本化周波数は 44.1 kHz，短時間 Fourier 変換の窓関数は Blackman 窓，窓長は 92.9 ms，シフト長は 46.5 ms とした．最適化アルゴリズムの反復回数は 200 回とした．被り音抑圧性能の客観評価には source-to-distortion ratio（SDR）[9] を用いた．本稿の実験では，観測信号の SDR からの改善量を音源毎に求めた．

3.2.2 最適なハイパーパラメータの調査

従来 TCNMF 及び提案 TCNMF のハイパーパラメータの最適値は開発データを用いて実験的に求めた．Figs. 1 及び 2 は，従来 TCNMF 及び提案 TCNMF の

Table 1　Mean and SD values [dB] for simulated test data over 100 parameter initializations

Method	Music no.	4	5	19	20	34	70	71	77	79	99
Simple	Mean	1.37	1.31	1.31	1.49	1.36	1.72	1.84	2.01	1.43	1.34
TCNMF	SD	0.02	0.02	0.07	0.03	0.11	0.06	0.05	0.23	0.05	0.06
Conventional	Mean	3.49	4.53	6.88	4.35	4.23	3.33	5.74	6.63	5.72	4.99
TCNMF	SD	0.09	0.20	0.39	0.14	0.17	0.29	0.28	0.44	0.27	0.24
Proposed	Mean	**3.80**	**5.60**	**8.17**	**4.64**	**5.12**	**4.72**	**5.88**	**9.30**	**6.51**	**6.02**
TCNMF	SD	$\mathbf{2.05\times10^{-8}}$	$\mathbf{1.08\times10^{-7}}$	$\mathbf{1.14\times10^{-7}}$	$\mathbf{5.49\times10^{-8}}$	$\mathbf{5.22\times10^{-7}}$	$\mathbf{1.18\times10^{-7}}$	$\mathbf{1.02\times10^{-7}}$	$\mathbf{1.16\times10^{-7}}$	$\mathbf{1.17\times10^{-7}}$	$\mathbf{3.77\times10^{-7}}$

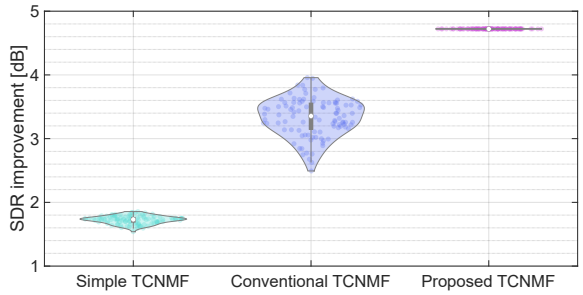

Fig. 3　Violin plots of SDR improvements for simulated test data with music no. 70.

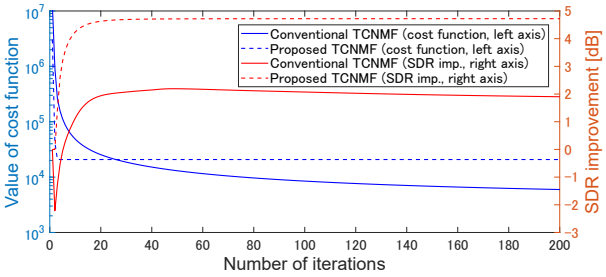

Fig. 4　Example behaviors of cost function values and SDR improvements for simulated test data with music no. 70.

各ハイパーパラメータにおける被り音抑圧性能の変化をそれぞれ示す．但し，各改善量の値は，10 種類の曲，10 種類の混合行列 $\overline{\boldsymbol{A}}_i$，及び 4 種類の音源全ての平均である．これらより，従来 TCNMF は $\mu = 1.047$，提案 TCNMF は $k = 1.22$，$\theta = 0.46$，及び $\alpha = 0.00027$ が学習データに対する最適値となった．

3.2.3　被り音抑圧の性能及び初期値頑健性の比較

前述の最適なハイパーパラメータを用いて，テストデータの被り音抑圧性能を比較した．テストデータを生成するための非負混合行列 $\overline{\boldsymbol{A}}_i$ は，開発データとは異なる擬似乱数 1 種類を用いた．初期値頑健性の確認のため，$\boldsymbol{A}_i$ 及び $\boldsymbol{S}_i$ の乱数初期値を 100 種類作成し反復最適化アルゴリズムを計算した．

Table 1 は，100 種類の初期値に対する曲毎の音源平均 SDR 改善量の平均と標準偏差（standard deviation: SD）を求めた結果である．また，Fig. 3 に曲番号 70 のデータにおける手法毎のバイオリン図を示す．但し，図中の Simple TCNMF は従来 TCNMF における $\mu = 0$ に対応する．バイオリン図中の色のついた点はひとつの初期値に対する結果（1 標本），中央の白い点は中央値，グレーの縦棒は四分位範囲，曲線はカーネル密度推定分布を表す．Table 1 及び Fig. 3 より，提案 TCNMF の被り音抑圧性能は従来 TCNMF と比較して，最適化変数の初期値の影響をほぼ受けていないことがわかる．3.1 節で述べた通りの理由により，提案 TCNMF の頑健性が確認できる．

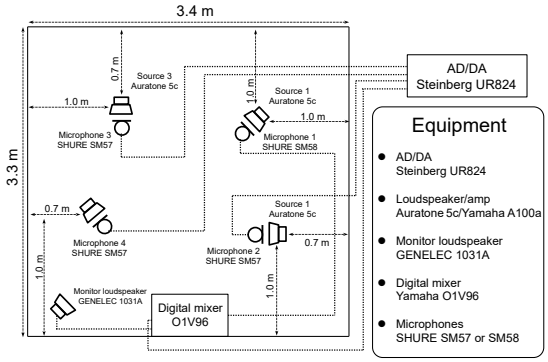

Fig. 5　Recording environment for producing real data.

最後に，反復回数とコスト関数の値及び SDR 改善量の関係を Fig. 4 に示す．いずれの TCNMF も SDR 改善量の収束速度は同程度である．

3.3　実収録混合信号の被り音抑圧実験

3.3.1　実験条件

より現実的な観測信号を用いた実験について示す．この実験では，被り音を含む観測信号を再現するために，各音源の時間信号 $\tilde{s}_n[t]$ と，各音源位置から各マイクロホンまでのインパルス応答 $\tilde{r}_{mn}[t]$ を用いて次式のように混合した．

$$\tilde{x}_m[t] = \sum_n \tilde{s}_n[t] * \tilde{r}_{mn}[t] \tag{11}$$

ここで，$\tilde{x}_m[t]$ は $\tilde{\boldsymbol{x}}[t]$ の要素，$*$ は時間領域における畳込み演算を表す．このインパルス応答 $\tilde{r}_{mn}[t]$ は Fig. 5 に示す環境下で録音した．この録音環境では，マイクロホン 1 で観測された音がミキサに入力され，モニタースピーカで再生される．従って，マイクロホン 1 の目的音源は，他のマイクロホンに強く被り音として混入する．その他の条件は 3.2.1 項と同様である．

3.3.2　最適なハイパーパラメータの調査

3.2.2 項と同様に開発データを用いて手法毎のハイパーパラメータの最適値を求めた．Figs. 6 及び 7 は，従来 TCNMF 及び提案 TCNMF の各ハイパーパラメータにおける被り音抑圧性能の変化をそれぞれ示す．これらより，従来 TCNMF は，$\mu = 0.0749$，提案 TCNMF は，$k = 1.02$，$\theta = 2.15$，及び $\alpha = 0.0008$ が学習データに対する最適値となった．

3.3.3　被り音抑圧の性能及び初期値頑健性の比較

3.2.3 項と同様に各手法の初期値頑健性を調査した．各手法のハイパーパラメータは 3.3.2 項にて決定した値を用いた．Table 2 は，100 種類の初期値に対する曲

Table 2　Mean and SD values [dB] for real test data over 100 parameter initializations

Method	Music no.	4	5	19	20	34	70	71	77	79	99
Simple	Mean	0.15	−0.58	−2.46	−0.29	−3.03	0.004	−0.30	−4.34	−2.16	−5.29
TCNMF	SD	0.08	0.50	0.97	0.54	0.73	0.26	0.82	1.22	0.65	0.66
Conventional	Mean	−0.15	−0.95	−1.93	−0.18	−3.19	−0.32	−0.19	−5.02	−2.52	−4.15
TCNMF	SD	0.23	0.51	1.16	0.61	0.78	0.49	0.81	1.36	0.84	0.56
Proposed	Mean	**1.23**	**2.33**	**1.61**	**2.25**	**0.52**	**2.34**	**2.98**	**1.97**	**1.69**	**0.72**
TCNMF	SD	$\mathbf{5.33\times10^{-7}}$	$\mathbf{1.83\times10^{-6}}$	$\mathbf{1.42\times10^{-4}}$	$\mathbf{7.66\times10^{-6}}$	$\mathbf{6.82\times10^{-6}}$	$\mathbf{3.57\times10^{-5}}$	$\mathbf{8.71\times10^{-7}}$	$\mathbf{1.09\times10^{-3}}$	$\mathbf{1.22\times10^{-6}}$	$\mathbf{2.94\times10^{-3}}$

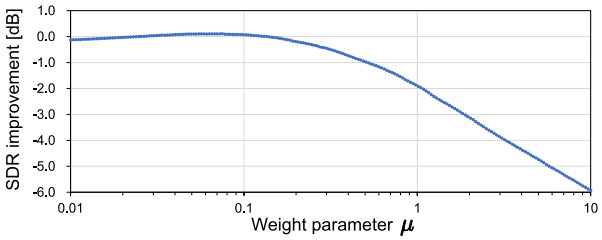

Fig. 6　SDR improvements for real training data obtained by conventional TCNMF.

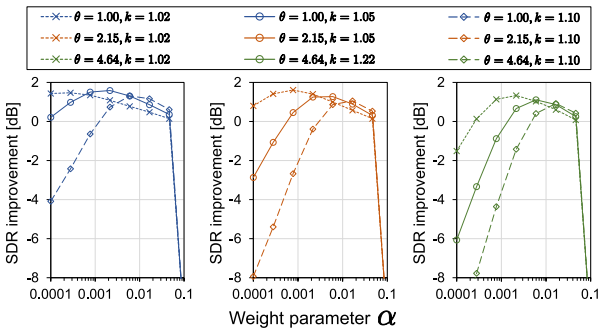

Fig. 7　SDR improvements for real training data obtained by proposed TCNMF.

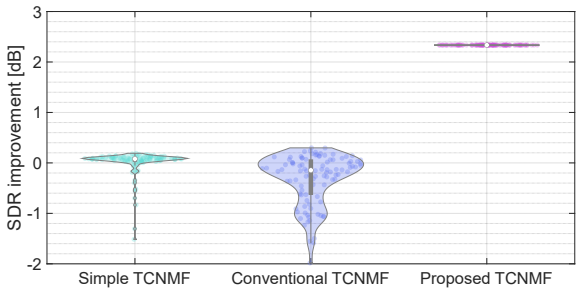

Fig. 8　Violin plots of SDR improvements for real test data with music no. 70.

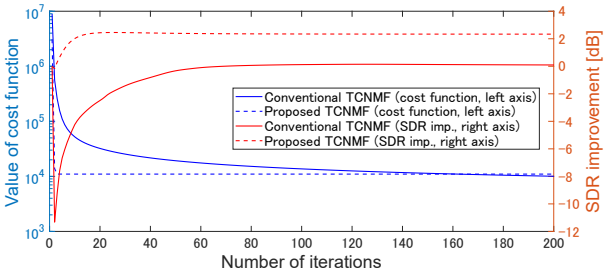

Fig. 9　Example behaviors of cost function values and SDR improvements for real test data with music no. 70.

毎の音源平均 SDR 改善量の平均と標準偏差を求めた結果である．また，Fig. 8 に曲番号 70 のデータにおける手法毎のバイオリン図を示す．Table 2 及び Fig. 8 より，シミュレーション信号の被り音抑圧実験と同様に，提案 TCNMF の被り音抑圧性能の初期値頑健性が確認できる．実収録混合信号においても，3.1 節で述べた理由により，提案 TCNMF の初期値頑健性が示されている．

最後に，反復回数とコスト関数の値及び SDR 改善量の関係を Fig. 9 に示す．Fig. 4 と比較すると，提案 TCNMF の SDR 改善量は従来 TCNMF よりも早く収束しており，優位性が確認できる．

4　まとめ

本稿では，従来 TCNMF 及び提案 TCNMF の被り音抑圧性能の最適化変数初期値に対する頑健性を調査した．実験結果より，提案 TCNMF の被り音抑圧性能は従来 TCNMF と比較して，最適化変数の初期値の影響をほぼ受けないことが確認された．これは，KLNMF の良い性質と，提案 TCNMF の推定パラメータの少なさに起因していると推測される．

謝辞

本研究の一部は，科研費 19H01116 及び 22H03652 の助成を受けた．

参考文献

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

[2] M. Togami, Y. Kawaguch, H. Kokubo, and Y. Obuchi, “Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization,” *Proc. APSIPA ASC*, pp. 522–525, 2010.

[3] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, “Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording,” *Proc. IWAENC*, pp. 203–207, 2014.

[4] Y. Murase, H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, “On microphone arrangement for multichannel speech enhancement based on nonnegative matrix factorization in time-channel domain,” *Proc. APSIPA ASC*, 2014.

[5] Y. Mizobuchi, D. Kitamura, T. Nakamura, H. Saruwatari, Y. Takahashi, and K. Kondo, “Prior distribution design for music bleeding-sound reduction based on nonnegative matrix factorization,” *Proc. APSIPA ASC*, 2021.

[6] A. T. Cemgil, “Bayesian inference for nonnegative matrix factorisation models,” *Comput. Intell. Neurosci.*, vol. 2009, no. 785152, 2009.

[7] O. Yılmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[8] A. Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, “The 2016 signal separation evaluation campaign,” *Proc. LVA/ICA*, pp. 323–332, 2017.

[9] 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.