多チャンネル変分自己符号化器を用いた劣決定音源分離*

◎関翔悟[1]，亀岡弘和[2]，李莉[3]，戸田智基[1]，武田一哉[1]
[1]名古屋大学，[2]NTTコミュニケーション科学基礎研究所，[3]筑波大学

1 はじめに

本稿では劣決定音源分離問題を扱う．多チャンネル非負値行列因子分解（MNMF）[1, 2] は劣決定音源分離に有効な手法であり，NMF を音源のパワースペクトログラムモデリングへと適用する．これは優決定条件下での MNMF でもある独立低ランク行列分析（ILRMA）[3, 4] にも取り入れられている．これらの手法は特定の音源に対して有効である一方で，NMF によってモデル化が困難な音源に対してはその分離性能が制限される．この問題に対して，NMF の代わりに条件付き変分自己符号化器（CVAE）を音源モデルとして利用する多チャンネル変分自己符号化器（MVAE）が最近提案され，ネットワークの柔軟な表現能力による高い分離性能が確認されている [5]．

本稿では優決定音源分離で定式化されている MVAE を劣決定音源分離へと適用し，停留点への収束が保証された最適化アルゴリズムを導く．2 チャネルのマイクロフォンアレイを用いた 3 話者の分離による実験的評価を行い提案法の有効性を示す．

2 定式化

観測チャネル数および音源数をそれぞれ I，J とし，観測信号および音源信号の時間周波数表現をそれぞれ

$$\mathbf{s}(f,n) = [s_1(f,n), \cdots, s_J(f,n)]^{\mathsf{T}} \in \mathbb{C}^J \quad (1)$$

$$\mathbf{x}(f,n) = [x_1(f,n), \cdots, x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I \quad (2)$$

とする．ここで $(\cdot)^{\mathsf{T}}$ は転置を表し，f，n はそれぞれ周波数，時間を表すインデックスである．本稿では，周波数領域における瞬時混合表現で表される混合系

$$\mathbf{x}(f,n) = \mathbf{A}(f)\mathbf{s}(f,n), \quad (3)$$

$$\mathbf{A}(f) = [\mathbf{a}_1(f), \cdots, \mathbf{a}_J(f)] \in \mathbb{C}^{I\times J} \quad (4)$$

を用いる．ここで $\mathbf{A}(f)$ は混合行列である．

いま，$s_j(f,n)$ が平均 0，分散 $v_j(f,n)$ の複素ガウス分布 $\mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n))$ に従う Local Gaussian Model (LGM) を仮定する．$j \neq j'$ において $s_j(f,n)$，$s_{j'}(f,n)$ が独立であるとき，観測信号 $\mathbf{s}(f,n)$ は

$$\mathbf{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|\mathbf{0}, \mathbf{V}(f,n)) \quad (5)$$

に従う．ここで $\mathbf{V}(f,n)$ は $v_1(f,n)$, $\cdots$, $v_J(f,n)$ を要素にもつ対角行列である．(3)，(5) より観測信号 $\mathbf{x}(f,n)$ は

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, \mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f)) \quad (6)$$

に従う．ここで $(\cdot)^{\mathsf{H}}$ は共役転置を表す．

したがって，混合行列 $\mathcal{A} = \{\mathbf{A}(f)\}_f$ および音源の分散 $\mathcal{V} = \{v_j(f,n)\}_{f,n}$ を用いて観測信号 $\mathcal{X} = \{\mathbf{x}(f,n)\}_{f,n}$ に対する対数尤度は以下で与えられる．

$$\log p(\mathcal{X}|\mathcal{A},\mathcal{V}) \stackrel{c}{=} -\sum\nolimits_{f,n} \big[\mathrm{tr}(\mathbf{x}^{\mathsf{H}}(f,n)(\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f))^{-1}\mathbf{x}(f,n)) + \mathrm{logdet}(\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f))\big] \quad (7)$$

ここで $\stackrel{c}{=}$ はパラメータに関する等号を表す．

3 関連研究

3.1 MNMF[2]

観測信号の空間共分散はステアリングベクトル $\mathbf{a}_j(f)$ および分散 $v_j(f,n)$ を用いて以下で表される．

$$\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f) = \sum\nolimits_j \mathbf{a}_j(f)v_j(f,n)\mathbf{a}_j^{\mathsf{H}}(f) = \sum\nolimits_j v_j(f,n)\mathbf{R}_j(f) \quad (8)$$

ここで $\mathbf{R}_j(f)$ は音源 j の空間共分散を表す．MNMF では，各音源信号の分散 $v_j(f,n)$ を K_j のスペクトルテンプレート $h_{j,1}(f), \cdots, h_{j,K_j}(f) \geq 0$ とその時変な励起パターン $u_{j,1}(n), \cdots, u_{j,K_j}(n) \geq 0$ を用いた

$$v_j(f,n) = \sum\nolimits_{k=1}^{K_j} h_{j,k}(f)u_{j,k}(n), \quad (9)$$

もしくは，$\sum_k b_{j,k} = 1$ を満たす指示変数 $b_{j,k} \in [0,1]$ を加え音源間で共有した K のスペクトルテンプレートおよびその励起パターンを用いた

$$v_j(f,n) = \sum\nolimits_{k=1}^{K} b_{j,k}h_k(f)u_k(n) \quad (10)$$

によってモデル化する．

MNMF の最適化アルゴリズムは空間共分散 $\mathcal{R} = \{\mathbf{R}_j(f)\}_{j,f}$ と音源モデル（$\mathcal{H}_1 = \{h_{j,k}(f)\}_{j,k,f}$，$\mathcal{U}_1 = \{u_{j,k}(n)\}_{j,k,n}$ もしくは $\mathcal{B} = \{b_{j,k}\}_{j,k}$，$\mathcal{H}_2 = \{h_k(f)\}_{k,f}$，$\mathcal{U}_2 = \{u_k(n)\}_{k,n}$）の反復更新によって構成され，それぞれの更新則は Majorization-Minimization (MM) アルゴリズム [6] により得られる．

3.2 MVAE [5]

MVAE では，CVAE のデコーダを音源のパワースペクトログラムの生成モデルとして利用する．CVAE は複素スペクトログラム $\tilde{\mathbf{S}}$ と音源ラベル c で構成される学習データを用いて事前に学習される．エンコーダの分布 $q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c)$ はガウス分布

$$q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c) = \prod\nolimits_k \mathcal{N}(z(k)|\mu_\phi(k;\tilde{\mathbf{S}}, c), \sigma_\phi^2(k;\tilde{\mathbf{S}}, c)), \quad (11)$$

デコーダの分布 $p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g)$ は平均 0 の複素ガウス分布

$$p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g) = \prod\nolimits_{f,n} \mathcal{N}_{\mathbb{C}}(s(f,n)|0, v(f,n)) \quad (12)$$

$$v(f,n) = g \cdot \sigma_\theta^2(f,n;\mathbf{z},c) \quad (13)$$

を表すようにネットワークを学習する．ここで $\mathbf{z}$ は潜在変数，$\mu_\phi(k;\tilde{\mathbf{S}}, c)$，$\sigma_\phi^2(k;\tilde{\mathbf{S}}, c)$ は要素 k のデコーダ出力 $\mu_\phi(\tilde{\mathbf{S}}, c)$，$\sigma_\phi^2(\tilde{\mathbf{S}}, c)$，$\sigma_\theta^2(f,n;\mathbf{z},c)$ は要素 (f,n) のデコーダ出力 $\sigma_\theta^2(\mathbf{z}, c)$ であり，g は生成されたスペクトログラムに対するグローバルスケールである．

---

*Underdetermined Source Separation Using Multichannel Variational Autoencoder.
by SEKI, Shogo (Nagoya University), KAMEOKA, Hirokazu (NTT), LI, Li (University of Tsukuba), TODA, Tomoki (Nagoya University), TAKEDA, Kazuya (Nagoya University)

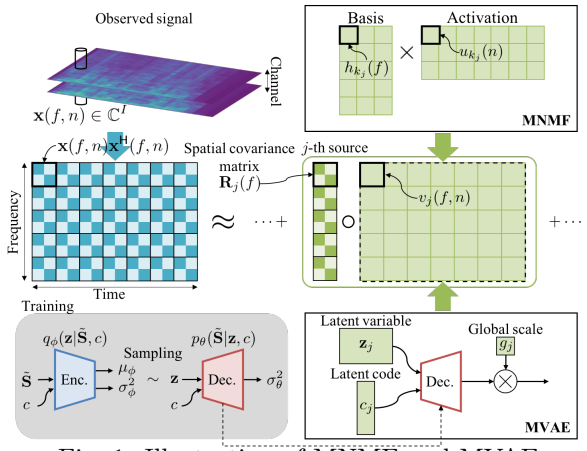

Fig. 1: Illustration of MNMF and MVAE

4 提案法：劣決定 MVAE

Fig. 1 に MVAE と (9) で与えられる音源モデルをもつ MNMF の概要を示す．MVAE においてデコーダの分布が LGM と同様の形式で表されるため，音源 j について $p_\theta(\tilde{\mathbf{S}}_j|\mathbf{z}_j,c_j,g_j)$ を用いて (7) の対数尤度に対する最適化が可能となる．したがって，MNMF と同様に MM アルゴリズムに基づいた空間共分散 $\mathcal{R}$ と音源モデル $(\mathcal{G}=\{g_j\}_j,\ \Psi=\{\mathbf{z}_j,c_j\}_j)$ の反復更新による最適化アルゴリズムが得られる．[6] より，

$$-\log p(\mathcal{X}|\mathcal{A},\mathcal{V}) \overset{c}{\leq} \sum_j \sum_{f,n} \left[\frac{\mathrm{tr}(\mathsf{X}(f,n)\mathbf{P}_j(f,n)\mathbf{R}_j^{-1}(f,n)\mathbf{P}_j(f,n))}{v_j(f,n)} + v_j(f,n)\mathrm{tr}(\mathbf{K}^{-1}(f,n)\mathbf{R}_j(f,n)) \right] \quad (14)$$

が成立する．ただし，等号成立は以下である．

$$\mathbf{P}_j(f,n) = v_j(f,n)\mathbf{R}_j(f,n)\left(\sum\nolimits_j v_j(f,n)\mathbf{R}_j(f,n)\right)^{-1} \quad (15)$$

$$\mathbf{K}(f,n) = \mathsf{X}(f,n) \quad (16)$$

ここで $\mathsf{X}(f,n) = \mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n)$ である．したがって，(14) の右辺を majorizer として利用することが可能であり，停留点への収束が保証される最適化アルゴリズムが得られる．$\mathcal{R}$ および $\mathcal{G}$ の更新則は MNMF の場合と同様に得られる [2]．また，(14) より majorizer は音源ごとの項へ分割されることから，潜在表現 Ψ は音源ごとに並行した最適化が可能である．

5 実験的評価

2 チャネルのマイクロフォンアレイを用いた 3 話者の音源分離性能を評価した．音源データとして Voice Conversion Challenge (VCC) 2018 dataset [7] から，男女計 4 話者の発話を利用した．計 4 パターンの 3 話者の組み合わせについて混合信号を生成した．[5] と同様に CVAE のエンコーダ，デコーダにはそれぞれ 3 層のゲート付き畳み込みネットワーク，3 層のゲート付き逆畳み込みネットワークを用いた．また，CVAE におけるラベル c は各話者に対応する 4 次元の one-hod 表現を用いた．(9) または (10) の音源モデルをもつ MNMF をベースライン（MNMF1，MNMF2）として提案法と比較した．ベースラインを 200 回反復した結果を各手法の初期値とし，ベースライン，提案法のイテレーションはいずれも 100 とした．提案法については異なる 2 つの初期値を用いて評価を行った（MVAE1, MVAE2）．評価尺度として，Signal-to-Distortion Ratio (SDR)，source Image-to-Spatial distortion Ratio (ISR)，Signal-to-Inference Ratio (SIR)，Signal-to-Artifact Ratio (SAR) を用いた [8]．残響時間 T_{60} が 78 ms，351 ms の場合においてそれぞれ評価を行った．実験結果を Fig. 2，3 に示す．いずれの残響環境においても，提案法において性能改善が確認できる．

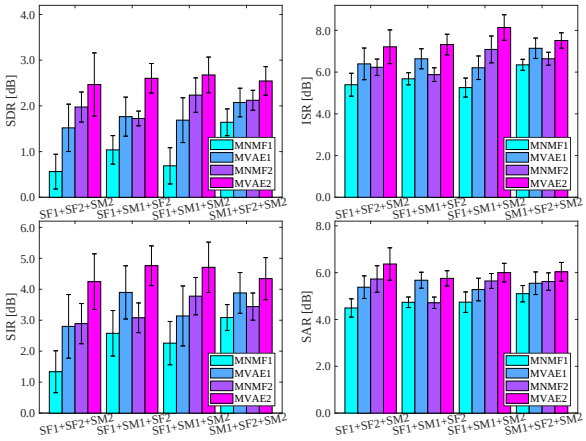

Fig. 2: Separation performances for $T_{60} = 78$ ms

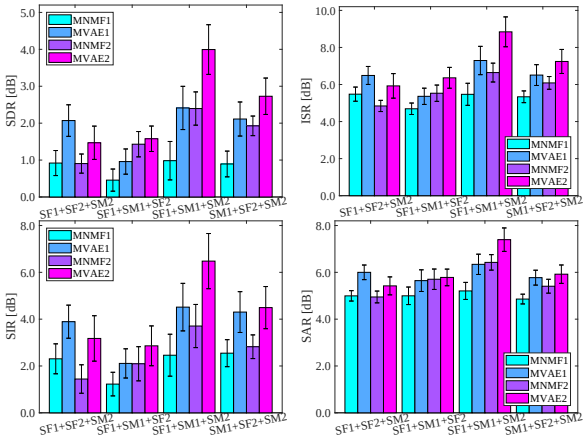

Fig. 3: Separation performances for $T_{60} = 351$ ms

6 おわりに

本稿では，優決定音源分離で高い分離性能が確認されている MVAE について一般化し，劣決定音源分離への適用を行った．提案法の最適化アルゴリズムは従来法である MNMF と同様に停留点への収束性が保証されている．2 チャネルのマイクロフォンアレイを用いた 3 話者の分離実験による実験的評価より，提案法による分離性能の改善を確認した．

謝辞 本研究の一部は，JSPS 科研費 17H01763 の助成を受け実施したものである．

参考文献

[1] Ozerov & Févotte, *IEEE TASLP*, 18(3), 550–563, 2010.
[2] Sawada, *et. al.*, *IEEE TASLP*, 21(5), 971–982, 2013.
[3] Kameoka, *et. al.*, *LVA/ICA*, 245–253, 2010.
[4] Kitamura, *et. al.*, *IEEE/ACM TASLP*, 24(9), 1622–1637, 2016.
[5] Kameoka, *et. al.*, *arXiv preprint 1808.00892*, 2018.
[6] Kameoka, *et. al.*, *Audio Souce Separation*, Springer, 95–124, 2018.
[7] Lorenzo-Trueba, *et. al.*, *arXiv preprint 1804.04262*, 2018.
[8] Vincent, *et. al.*, *IEEE TASLP*, 14(4), 1462–1469, 2006.