SepNet: 高速多チャンネル音源分離のための分離行列予測ネットワーク *

☆井上翔太 [1]　亀岡弘和 [2]　李莉 [1]　牧野昭二 [1]

[1] 筑波大学　[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

ブラインド音源分離（Blind Source Separation: BSS）は，音源に関する情報や音源とマイク間の伝達関数等の事前情報を用いずに観測された混合信号のみから個々の信号を推定する技術であり，ロボット聴覚，補聴器，音声認識や異常音検知等様々なアプリケーションの精度向上に貢献する．マイクロホンの数が音源数を上回る優決定条件下においては，音源信号間の独立性を最大化するように分離フィルタを推定することを目的とする独立成分分析（Independent Component Analysis: ICA）[1] が有用であることが知られており，それに基づく時間周波数領域での分離手法が数多く提案されている [2–5]．これらの手法は，音源に関する時間周波数領域で成り立つ様々な仮定やマイクロホンアレーの周波数応答に関する仮定を有効に活用できるという利点がある．例えば，独立低ランク行列分析（Independent Low-Rank Matrix Analysis: ILRMA）[3,4] は，各音源信号のパワースペクトログラムを非負値行列とみなし，二つの非負値行列の積で表現するアプローチである．これは，各時間フレームにおけるパワースペクトルを時間的に変化する振幅によってスケーリングされた基底スペクトルの線形和で近似することに相当する．これにより，ILRMA は音源のスペクトル構造を手がかりにしながら周波数ごとの音源分離とパーミュテーション整合と呼ぶ問題の同時解決を可能としている．近年，ICA をはじめとした信号処理に基づく手法に深層学習（Deep Neural Network: DNN）を導入することで，更なる分離精度の向上を実現するアプローチに対する注目が高まっている [5,7–11]．これらの手法は DNN の持つ豊かな関数表現能力を生かし，ILRMA などの手法における行列積で表すパワースペクトログラムモデルに DNN を用いることで高い分離精度を実現している．

上述の手法ではいずれも反復的な最適化アルゴリズムを用いてパラメータ推定を行う．反復射影法（Iterative Projection: IP）を用いた分離行列更新法はその代表例である [5,7,9,10]．IP 法は比較的に少ない反復回数で局所解に収束を可能とし，高速なアルゴリズムであるが，反復当たりの計算コストは音源数およびチャンネル数の 4 乗に比例して増加する．大規模なマイクロホンアレイに適用する際に膨大な計算コストとなり，深刻な問題となる．この問題を解決するために，計算コストの増加を音源数およびチャンネル数の 3 乗比例に留めた Iterative Source Steering (ISS) 法が提案されている [12]．IP 法や ISS 法の更新法則は人為的に設計した尤度関数を最大化するように導出されている．このような反復アルゴリズムの各反復計算をネットワークの一層とみなせば，全体をパラメータ学習が必要としない DNN として解釈できる．従って，適切なネットワーク構造の設計と学習法により，DNN を用いた分離行列の推論プロセスの学習も考えられる．

本稿では，データ駆動の分離行列予測手法である *SepNet* を提案する．SepNet は反復アルゴリズムの初期値に相当する分離行列を，多チャンネル観測信号に応じた適切な分離行列へ射影する非線形変換関数を学習するネットワークであり，テスト時には入力混合信号を手がかりとして分離行列を推定する．ネットワーク構造は従来の反復アルゴリズムにおける各反復の計算過程を模した DNN ブロックを複数縦列接続し，深いネットワークを構築する．提案法は各層に 2 次元や 3 次元の畳み込み層（Convolutional Neural Network: CNN）が用いられるため，フォワード計算の計算コストが音源数及びチャンネル数の 2 乗比例に留まり，従来法より計算コスト削減に大いに有利である．また，BSS におけるパーミュテーション問題を柔軟に対応するため，ネットワーク学習には Permutation Invariant Training (PIT) を用いる．提案法が高速な音源分離を実現できることを 2 チャンネルおよび 3 チャンネルの音声分離実験で示す．

2　音源分離問題の定式化

2.1　優決定条件下の BSS 問題

I 個のマイクロホンで J 個の音源から到来する信号を観測する場合を考える．i 番目のマイクで観測される信号と j 番目の音源信号の時間周波数成分をそれぞれ $x_i(f,n)$ と $s_j(f,n)$ とする．ただし，f と n は周波数と時間フレームのインデックスである．優決定条件下において $I = J$ とする．音源とマイクの間の室内インパルス応答長が時間周波数展開における窓長よりも十分短い場合には，音源信号 $\mathbf{s}(f,n) = [s_1(f,n),\ldots,s_J(f,n)]^{\mathsf{T}} \in \mathbb{C}^J$ と観測信号 $\mathbf{x}(f,n) = [x_1(f,n),\ldots,x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$ の関係性は瞬時混合系を用いて

$$\mathbf{s}(f,n) = \mathbf{W}^{\mathsf{H}}(f)\mathbf{x}(f,n) \tag{1}$$

$$\mathbf{W}^{\mathsf{H}}(f) = [\mathbf{w}_1(f),\ldots,\mathbf{w}_I(f)] \in \mathbb{C}^{I\times I} \tag{2}$$

と表せる．ここで，$\mathbf{W}^{\mathsf{H}}(f)$ は分離行列を表し，$(\cdot)^{\mathsf{T}}$ は行列の転置であり，$(\cdot)^{\mathsf{H}}$ はエルミート転置である．BSS 手法の目的は多チャンネル観測信号 $\mathcal{X} = \{x_i(f,n)\}_{i,f,n}$ から分離行列 $\mathcal{W} = \{\mathbf{w}_j(f)\}_{j,f}$ を推定することである．

2.2　独立低ランク行列分析による音源分離

次に，観測信号が生成されるプロセスを記述する．音源 j の複素スペクトログラム $s_j(f,n)$ を

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)) \tag{3}$$

のように平均が 0，分散が $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$ の

---

*SepNet: separation matrix prediction network for determined multichannel source separation. By Shota Inoue (University of Tsukuba), Hirokazu Kameoka (NTT Communication Science Science Laboratories), Li Li, Shoji Makino (University of Tsukuba)

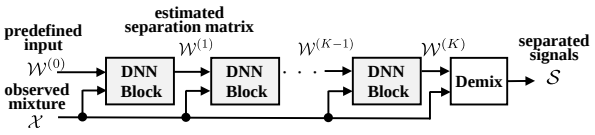


Fig. 1: 提案法のネットワーク構造を表すブロック図.

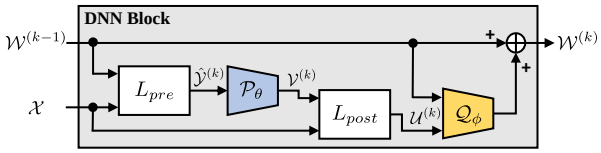


Fig. 2: DNN ブロックの内部構造．学習可能なパラメータを持たない 2 つの層（$L_{\mathrm{pre}}, L_{\mathrm{post}}$）と，学習可能なパラメータを持つ 2 つの層（$\mathcal{P}_\theta$, $\mathcal{Q}_\phi$）の合計 4 層で構成される．

複素正規分布に従う確率変数と仮定する．各音源が統計的に独立である場合，$\mathbf{s}(f,n)$ は

$$\mathbf{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|\mathbf{0}, \mathbf{V}(f,n)) \tag{4}$$

に従う．ここで，$\mathbf{V}(f,n)$ は $v_1(f,n),\ldots,v_J(f,n)$ を対角成分に持つ対角行列である．式 (1) と式 (4) より，観測信号 $\mathbf{x}(f,n)$ は

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|0, (\mathbf{W}^{\mathsf{H}}(f))^{-1}\mathbf{V}(f,n)\mathbf{W}(f)^{-1}) \tag{5}$$

に従う．従って，分離行列 $\mathcal{W}$ と各音源の音源モデルパラメータ $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$ が与えられた下での観測信号 $\mathcal{X}$ の負対数尤度関数は

$$\mathcal{L}(\mathcal{V},\mathcal{W}|\mathcal{X}) \stackrel{c}{=} -2N \log\left|\det \mathbf{W}^{\mathsf{H}}(f)\right| + \sum_{f,n,j}\left(\log v_j(f,n) + \frac{|\mathbf{w}_j^{\mathsf{H}}(f)\mathbf{x}(f,n)|^2}{v_j(f,n)}\right) \tag{6}$$

のように書ける．ここで，$\stackrel{c}{=}$ はパラメータに依存しない項を除いた等号を表す．

ILRMA では各音源のパワースペクトログラムを非負行列とみなし，非負値行列因子分解によって式 (6) の音源モデル $v_j(f,n)$ を基底スペクトル $b_{j,m}(f) \geq 0$ と時間フレームごとのゲイン $h_{j,m}(n) \geq 0$ の線形和 $v_j(f,n) = \sum_{m=1}^{M} b_{j,m}(f)h_{j,m}(n)$ で近似表現する．ただし，$m = 1,\ldots,M$ は基底のインデックスである．$\mathcal{B} = \{b_{j,m}(f)\}_{j,m,f}$，$\mathcal{H} = \{h_{j,m}(n)\}_{j,m,n}$ とすると，ILRMA では負対数尤度関数式 (6) が反復ごとに減少するように $\mathcal{B}$，$\mathcal{H}$，および $\mathcal{W}$ を更新する．$\mathcal{B}$ と $\mathcal{H}$ の更新には補助関数法に基づく更新式が用いられる．$\mathcal{W}$ の更新には IP 法

$$\mathbf{w}_j(f) \leftarrow (\mathbf{W}^{\mathsf{H}}(f)\mathbf{\Sigma}_{x/v_j(f)})^{-1}\mathbf{e}_j, \tag{7}$$

$$\mathbf{w}_j(f) \leftarrow \mathbf{w}_j(f)/\sqrt{\mathbf{w}_j^{\mathsf{H}}(f)\mathbf{\Sigma}_{x/v_j}(f)\mathbf{w}_j(f)} \tag{8}$$

を用いることができる．ここで，$\mathbf{\Sigma}_{x/v_j}(f) = (1/N)\sum_n \mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n)/v_j(f,n)$ であり，$\mathbf{e}_j$ は $I\times I$ の単位行列の j 番目の列ベクトルを表す．ILRMA は式 (6) の局所解への収束性が保証されており，比較的少ない反復回数での局所解への収束が実験的に示されている．しかしながら，式 (7) に示されているように各反復で J 個の音源ごとに $\mathcal{O}(J^3)$ の計算量の逆行列計算が必要であるため，アルゴリズムの計算量は $\mathcal{O}(J^4)$ となる．

3 提案法

より少ない計算量での高精度な分離行列の推定を目的として，DNN ベースの分離行列推定手法である SepNet を提案する．SepNet は複数の DNN ブロックにより構成される深いネットワークであり，k 番目のブロックでは分離行列 $\mathcal{W}^{(k-1)}$ と多チャンネル観測信号 $\mathcal{X}$ が入力される場合に，より正確な分離行列 $\mathcal{W}^{(k)}$ を推定するネットワークである．本節では，SepNet の DNN ブロックの設計について述べた後に，ネットワーク学習について記述する．

3.1 ネットワーク構造

図 1 と図 2 に，提案法のアーキテクチャ全体と各 DNN ブロックの内部構造を示す．図 1 に示すように，SepNet は縦列接続された K 個の同じ構造を持つ DNN ブロックで構成されており，フォーワード計算は入力された分離行列の初期値 $\mathcal{W}^{(0)}$ を K 回更新することと等しい．2.1 節に述べた ILRMA の更新則をヒントに，各 DNN ブロックを 4 つ微分可能な層で構成するように設計した．第 1 層 L_{pre} では，入力分離行列 $\mathcal{W}^{(k-1)}$ を観測信号 $\mathcal{X}$ に適用し，分離信号のパワースペクトログラムを得る．第 2 層 $\mathcal{P}_\theta$ では，L_{pre} で得られた各分離信号のパワースペクトログラムから，各音源の音源モデルパラメータに相当する中間出力を推定する．第 3 層 L_{post} では，$\mathcal{P}_\theta$ で推定された中間出力を用いて，重み付き空間共分散行列を計算する．第 4 層の $\mathcal{Q}_\phi$ では，L_{post} で計算された重み付き空間共分散行列と入力分離行列 $\mathcal{W}^{(k-1)}$ から分離行列の更新量 $\Delta\mathcal{W}^{(k)}$ を推定する．ここで，L_{pre} と L_{post} は，学習可能なパラメータを持たない層であり，$\mathcal{P}_\theta$ と $\mathcal{Q}_\phi$ は，それぞれ学習可能なパラメータ θ と ϕ を持つ非線形層である．

$(k-1)$ 番目の DNN ブロック出力を $\mathcal{W}^{(k-1)} = \{\mathbf{w}_j^{(k-1)}(f)\}_{j,f}$ とし，k 番目の DNN ブロックにおける第 1 層から第 3 層の中間出力を $\hat{\mathcal{Y}}^{(k)} = \{\hat{y}_j^{(k)}(f,n)\}_{j,f,n}$，$\mathcal{V}^{(k)} = \{v_j^{(k)}(f,n)\}_{j,f,n}$ および $\mathcal{U}^{(k)} = \{\mathbf{U}_j^{(k)}(f)\}_{j,f}$ と定義すると，k 番目の DNN ブロックの最終的な出力 $\mathcal{W}^{(k)} = \{\mathbf{w}_j^{(k)}(f)\}_{j,f}$ は以下のように表される．

$$\hat{\mathcal{Y}}^{(k)} = L_{\mathrm{pre}}(\mathcal{W}^{(k-1)}, \mathcal{X}), \tag{9}$$

$$\mathcal{V}^{(k)} = \mathcal{P}_{\theta^{(k)}}(\hat{\mathcal{Y}}^{(k)}), \tag{10}$$

$$\mathcal{U}^{(k)} = L_{\mathrm{post}}(\mathcal{V}^{(k)}), \tag{11}$$

$$\mathcal{W}^{(k)} = \mathcal{W}^{(k-1)} + \mathcal{Q}_{\phi^{(k)}}(\mathcal{W}^{(k-1)}, \mathcal{U}^{(k)}). \tag{12}$$

ここで，$\hat{\mathcal{Y}}$ と $\mathcal{U}^{(k)}$ の各要素 $\hat{y}_j^{(k)}(f,n)$ および $\mathbf{U}_j^{(k)}(f)$ はそれぞれ

$$\hat{y}_j^{(k)}(f,n) = |\mathbf{w}_j^{(k-1)\mathsf{H}}(f)\mathbf{x}(f,n)|^2, \tag{13}$$

$$\mathbf{U}_j^{(k)}(f) = \frac{1}{N}\sum_n \frac{\mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n)}{v_j^{(k)}(f,n)} \tag{14}$$

である．式 (10) は ILRMA において $v_j(f,n)$ を行列積で近似するプロセスに相当し，暫定的な分離信号の

パワースペクトログラムを精緻化する過程だとみなせる．式 (12) に残差学習を用いることで，ネットワーク全体の学習を安定させることが期待できる．また，4 章で述べるように，$\mathcal{P}_\theta$ と $\mathcal{Q}_\phi$ の全ての層を 2 次元や 3 次元 CNN のみで構成するため，各 DNN ブロックのフォワード計算における計算コストの増加を音源数およびチャンネル数の 2 乗に留めることができる．

式 (10) と式 (12) において，各 DNN ブロックのネットワークパラメータ $\theta^{(k)}$ および $\phi^{(k)}$ をブロックごとに異なるものとして学習する場合と，全ての DNN ブロックのパラメータを同一として学習する場合の二つが考えられる．本稿では，前者を *Tied* モデルと呼び，後者を *Untied* モデルと呼ぶ．

3.2 学習方法

すべての DNN ブロックがより正確な分離行列を推定するように学習するため，多チャンネル観測信号と音源信号のペア $\{\mathcal{X}, \mathcal{S}\}$ が与えられた場合，ネットワークパラメータはすべての DNN ブロックの出力によって得られた分離信号 $\mathbf{w}_j^{(k)\mathsf{H}}(f)\mathbf{x}(f,n),\ k = 1, \ldots, K$ と音源信号 $s_j(f,n)$ の誤差

$$E(\Theta) = \mathbb{E}_{\mathcal{X},\mathcal{S}}\Big[\sum_{k,j}\sum_{f,n}|\mathbf{w}_j^{(k)\mathsf{H}}(f)\mathbf{x}(f,n) - s_j(f,n)|\Big] \tag{15}$$

の最小化により学習させる．ここで，Θ は学習対象のネットワークパラメータを表し，$\mathbb{E}_{\mathcal{X},\mathcal{S}}[\cdot]$ は全ての学習ペアデータに対するサンプル平均を表す．しかしながら，前述のように BSS 問題には分離信号の出力順序に任意性がある．そこで，ネットワークパラメータの学習時においては目的とする $\mathcal{S}$ の音源の順序と $\mathcal{W}$ によって分離された信号の順序の異なりを許容することが望ましい．この問題を解消するため，Permutation Invariant Training (PIT) [13] 学習を目的関数に導入する．PIT を用いた学習では，出力の分離信号と目的の音源信号の最適な割り当てを計算し，割り当てられた分離信号と音源信号間の誤差を最小化する．j 番目の分離信号と対応する音源信号のインデックスを $\pi(j) \in \{1, \ldots, J\}$ とおくと，PIT を用いた目的関数は

$$E(\Theta) = \mathbb{E}_{\mathcal{X},\mathcal{S}}\Big[\min_{\pi}\sum_{k,j}\sum_{f,n}|\mathbf{w}_j^{(k)\mathsf{H}}(f)\mathbf{x}(f,n) - s_{\pi(j)}(f,n)|\Big] \tag{16}$$

と書ける．学習時には，各学習データにおいて $J!$ 通りの分離信号と音源信号間の誤差を評価し，誤差が最小となる組み合わせを用いる．一方で，我々は $I = J = 2$ の場合に限っては音源の到来方向順に順序を固定することで，PIT を導入せずに同等の分離結果が得られることを実験的に確認した．

4 評価実験

提案手法の有効性を評価するため，既存手法の ILRMA と提案手法について 2 チャンネル 2 音源および 3 チャンネル 3 音源の音声分離実験を実施した．ネットワーク学習のための音声信号として CMU ARCTIC データセット [14] の 18 話者の発話データを用いた．また，評価データの音声信号として Voice Conversion Challenge (VCC) 2018 [15] の男女各 2 話者，合計 4 話者の発話データを用いた．多チャンネル観測信号の生成に用いた 2 チャンネル 2 音源および 3 チャンネル 3 音源のインパルス応答は，鏡像法 [16] を用いて生成した．生成に用いた部屋の奥行，幅および高さはそれぞれ 4.0 m，5.0 m，および 3.0 m とした．また，それ以外のインパルス応答の生成に必要なパラメータは以下の手順で学習データごとに決定した．まず，残響時間（RT60）は 55 ms から 160 ms の範囲でランダムに選択する．次に，マイクロホンアレイの配置場所を壁から 0.5 m 以上離れた座標でランダムに選択する．最後に，マイクロホンアレイの中心から半径 0.5–1.0 m のいずれかの距離に，正面 0° から 180° 方向に 20° 以上の角度差で音源を配置する．2 音源および 3 音源のそれぞれの条件で，3000 発話の学習データと 100 発話の評価データを作成した．また，音声信号の標本化周波数を 8 kHz とし，短時間 Fourier 変換の条件を窓長 64 ms，シフト量 32 ms とした．$\mathcal{P}_\theta$ と $\mathcal{Q}_\phi$ のネットワーク構造を図 3 に示す．$\mathcal{P}_\theta$ にはゲート（Gated Linear Unit: GLU）付きの 2 次元の畳み込み層（Conv）と逆畳み込み層（Deconv）を用いた．$\mathcal{Q}_\phi$ にはゲート付きの 3 次元の畳み込み層および逆畳み込み層を用いた．また，$\mathcal{Q}_\phi$ において，入力である複素の行列式 (14) の実数部と虚数部をチャンネル方向に分割して 3 次元配列として扱った．全ての層を CNN で構成しているため，任意の長さの観測信号 $\mathcal{X}$ への適用が可能である．学習時における発話データの時間フレーム長は 128 とした．反復回数に相当する DNN ブロックの数 K は Tied モデルおよび Untied モデルでそれぞれ 10 および 4 とした．ILRMA の基底数 M を 2 とし，反復回数を 50 回とした．全てのアルゴリズムを Python および PyTorch で実装し，Intel (R) Core i7-7800X CPU@3.50 GHz と GeForce TITAN V GPU を用いて分離性能の評価と処理時間の計測を実施した．分離性能の客観評価尺度としては signal-to-distortions ratio (SDR)，signal-to-interferences ratio (SIR)，および signals-to-artifacts ratio (SAR) [17] を用いた．

各手法での処理時間に対する SDR，SIR および SAR の平均値の推移を図 4 に示す．ILRMA は 1 から 10 回目の反復および 10 から 50 回目まで 10 回ごとの結果を，提案法は各 DNN ブロックの結果をそれぞれプロットした．実験結果より，Tied モデルと Untied モデルの双方が最初のブロックから精度の高い分離行列を出力可能であることが確認できる．特に Untied モデルの最終的な出力が ILRMA を上回る分離性能を達成可能であり，より短い処理時間で ILRMA と同等の分離性能を達成可能であることが示された．Tied モデルは SAR に関して Untied モデルをわずかに上回っているが，SDR および SIR に関しては Untied モデルが大きく向上している．この結果より，SepNet がより優れた分離性能を達成するためにより柔軟なネットワーク構造の設計が重要であると考察する．

5 まとめ

本稿では，決定条件下での周波数領域の音源分離問題において，DNN を用いた分離行列予測手法である SepNet を提案した．提案法は従来法を参考とした DNN の設計と事前学習により，DNN による多チャンネル観測信号から分離行列を推定する過程の学習

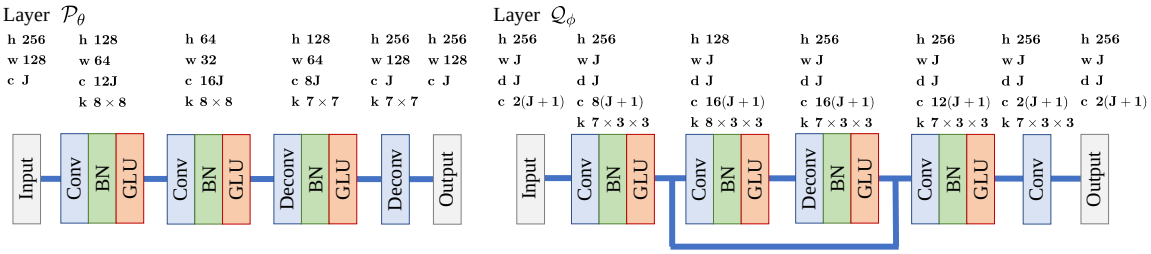
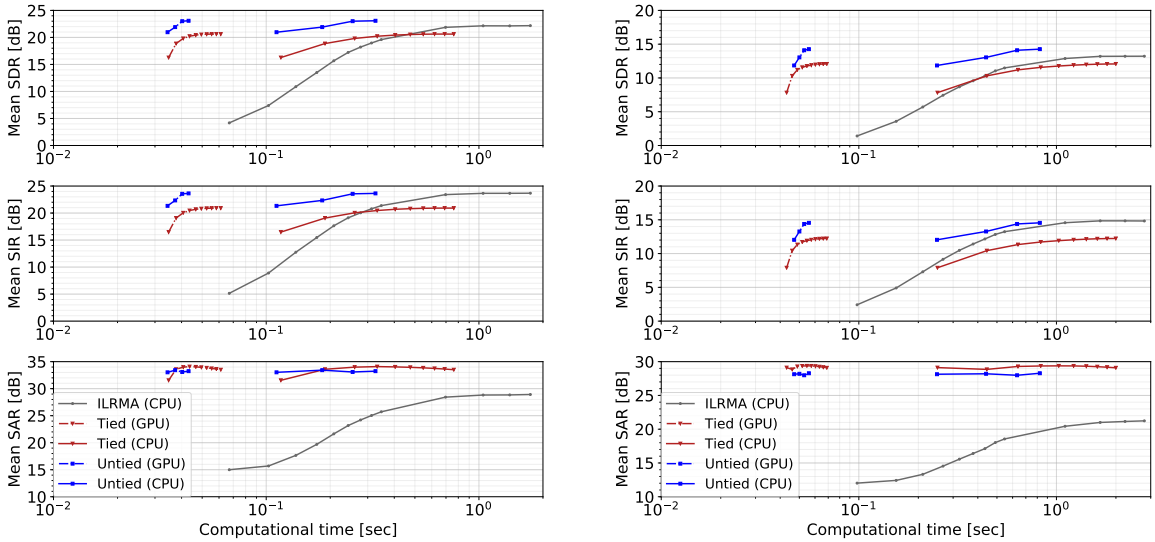


Fig. 3: $\mathcal{P}_\theta$ and $\mathcal{Q}_\phi$ のネットワーク構造．“h”，“w”，“d”，“c” と “k” はそれぞれ入力の高さ，幅，深さ，チャネル数およびフィルタのカーネルサイズを表す．“J” は観測信号のチャンネル数を表す．“Conv”，“Deconv”，“BN” と “GLU” は畳み込みと逆畳み込み，バッチ正規化，および Gated Linear Unit を表す．



(a) 2 チャンネル 2 音源の分離結果．

(b) 3 チャンネル 3 音源の分離結果．

Fig. 4: 各手法の処理時間と SDR, SIR および SAR の平均値．Tied モデルと Untied モデルのブロック数を 10 および 4 とした．ILRMA は反復回数を 50 回とし，1 から 10 回目と以降 10 回ごとの結果を示した．

を目的としている．また，ネットワーク出力の分離行列と目的とする音源信号間に生じるパーミュテーションの任意性に対応するため，パラメータ学習において PIT を導入した損失関数を用いた学習を行った．2 音源および 3 音源の音声分離実験から，SepNet は従来法の ILRMA と比較してより高速かつ精度の高い分離行列を推定可能であることが示された．SepNet は ILRMA や IP 法に対してチャンネル数および音源数の増加に対しての推論における計算コストの増加に関して優位性があるが，一方で事前学習の難化に関して課題がある．これは，PIT で評価する組み合わせの総数が音源数の 2 乗に比例して増加するためであり，より効率的な学習手順を検討する必要がある．

謝辞　本研究の一部は，JST CREST JPMJCR19A3 および JSPS 科研費 19H04131 の助成を受けたものである．

参考文献

[1] A. Hyvärinen *et al.*, *Neural networks*, 13(4), 411–430, 2000.

[2] N. Ono, *WASPAA*, 189–192, 2011.

[3] H. Kameoka *et al.*, *LVA/ICA*, 245–253, 2010.

[4] D. Kitamura *et al.*, *Audio Source Separation*, 125–155, 2018.

[5] H. Kameoka *et al.*, *Neural Computation*, 31(9), 1891–1914, 2019.

[6] D. D. Lee *et al.*, *NIPS*, 556–562, 2001.

[7] S. Mogami *et al.*, *EUSIPCO*, 1557–1561, 2018.

[8] A. A. Nugraha *et al.*, *IEEE/ACM Trans. ASLP*, 14(9), 1652–1664, 2016.

[9] L. Li *et al. IEEE Access*, 8(1), 228740–228753, 2020.

[10] S. Inoue *et al.*, *ICASSP*, 96–100, 2019.

[11] K. Sekiguchi *et al.*, *APSIPA*, 1233-–1239, 2018.

[12] R. Scheibler *et al.*, *ICASSP*, 236–240, 2020.

[13] D. Yu *et al.*, *ICASSP*, 241–245, 2017.

[14] J. Kominek *et al.*, *SSW*, 223–224, 2004.

[15] J. Lorenzo-Trueba *et al.*, arXiv preprint: 1804.04262, 2018.

[16] J. B. Allen *et al.*, *ASA*, 65(4), 943–950, 1979.

[17] E. Vincent *et al.*, *IEEE Trans. ASLP*, 14(4), 1462–1469, 2006.