主-双対近接分離法にDNNをPlug-and-Playした優決定ブラインド音源分離*

☆ 松本和樹 (早大), 矢田部浩平 (農工大)

1 はじめに

線形フィルタによる優決定ブラインド音源分離において，現実に忠実な音源モデルを実現するため，分離行列の最適化にDNNを援用する手法が提案されている [1–3]．しかし，音声同士を安定的に分離するには特別な工夫が必須である．従来法は話者に関する情報や学習時の分離行列推定を要しており，DNN自体に特殊な構造や学習法が必要であった．

本稿では，単なる音声強調器として学習したDNNで，音声同士の分離を可能にする手法を提案する．提案法は，アルゴリズム中の近接作用素をDNNで置き換える Plug-and-Play (PnP) の枠組みに基づく．具体的には，主-双対近接分離アルゴリズム (PDS) に対し，スパース性を誘導する閾値作用素とDNNを平均化した音声強調器を差し込む．平均化の導入は，従来の音源分離で安定した分離を達成する閾値作用素と，緻密な音声強調を実現するDNNとの相補的な作用をもたらし，音声同士の分離問題を解決する．実験の結果，作用素の平均化は有効に働き，スパース性のみに基づく従来手法に勝る分離性能が得られた．

2 PDSを用いた音源分離

優決定ブラインド音源分離では周波数ごとの線形分離フィルタ $\mathbf{W}[f] \in \mathbb{C}^{N\times M}$ を推定し，時間周波数領域における M チャンネルの観測信号 $\mathbf{x}[f,t] \in \mathbb{C}^M$ から N 音源の分離音 $\mathbf{y}[f,t] \in \mathbb{C}^N$ を

$$y_n[f,t] = \sum_{m=1}^{M} W_{n,m}[f]\, x_m[f,t] \tag{1}$$

として得る．多くの音源分離手法では，分離音 $\mathbf{y}[f,t]$ のスパース性を考慮した音源モデル項 $\mathcal{P}$ を導入し，分離行列を最適化問題

$$\min_{(\mathbf{W}[f])_{f=1}^{F}} \mathcal{P}(((\mathbf{y}[f,t])_{f=1}^{F})_{t=1}^{T}) - 2T\sum_{f=1}^{F}\log|\det \mathbf{W}[f]| \tag{2}$$

の解で特徴付ける．最適化にPDSを用いることで，統一的な枠組みで様々な音源モデルによる分離を実現できる．さらに，音源モデル項の近接作用素 $\mathrm{prox}_{\mathcal{P}}$ を，音源を強調するマスク推定器 $\mathcal{M}: \mathbb{C}^{NFT} \to [0,1]^{NFT}$ の適用で置き換えれば，音源の多様な性質を考慮した柔軟な音源モデル設計が可能となる [4]．

Algorithm 1 Proposed Method

Input: $\mathbf{X}$, $\mathbf{w}^{[1]}$, $\mathbf{y}^{[1]}, \mu_1, \mu_2, \alpha$
Output: $\mathbf{w}^{[K+1]}$

for $k = 1, \cdots, K$ **do**
  $\mathbf{w}^{[k+1]} = \mathrm{prox}_{-2\mu_1 \log|\det(\cdot)|}\left(\mathbf{w}^{[k]} - \mu_1\mu_2 \mathbf{X}^{\mathsf{H}}\mathbf{y}^{[k]}\right)$
  $\mathbf{z} = \mathbf{y}^{[k]} + \mathbf{X}(2\mathbf{w}^{[k+1]} - \mathbf{w}^{[k]})$
  $\mathbf{y}^{[k]} = \mathbf{z} - \mathcal{M}^{\alpha}_{\mathrm{IVA/DNN}}(\mathbf{z}) \odot \mathbf{z}$
end for

3 提案手法

緻密な音源モデルを実現するため，マスク推定器 $\mathcal{M}$ としてDNNによる音声強調器を用いることが考えられる．しかし，音声同士の音源分離では，初期の分離音において複数の話者が混在するため，その中から特定の誰かを選択的に強調するDNNでなければ，単純なPnPによる音源の分離は困難となる．

この問題を解決するため，DNNのPnPとスパース性に基づく手法を組み合わせた音源分離手法を提案する．提案法は両者の相補的な作用により，単に音声強調器として学習したDNNを用いながら安定した音源分離を実現する．

3.1 DNNとIVAの平均化マスク

従来法である独立ベクトル分析 (IVA) は分離音のグループスパース性を誘導することで音声同士の安定した分離を実現している．そこで，提案法ではPnPの安定化を目的として，DNNの生成する時間周波数マスク $\mathcal{M}_{\mathrm{DNN}}$ と，グループソフト閾値作用素と等価なマスク $\mathcal{M}_{\mathrm{IVA}}$ を平均化率 α で足し合わせたマスク

$$\mathcal{M}^{\alpha}_{\mathrm{IVA/DNN}} = (1-\alpha)\mathcal{M}_{\mathrm{IVA}} + \alpha\mathcal{M}_{\mathrm{DNN}} \tag{3}$$

を用いる．ただし，$\mathcal{M}_{\mathrm{IVA}}$ はIVAの音源モデル項に対応するマスク推定器である．

提案アルゴリズムを **Algorithm 1** に示す．ここで，$\mathbf{w} \in \mathbb{C}^{NMF}$ はベクトル化した分離行列，$\mathbf{X} \in \mathbb{C}^{NFT\times NMF}$ は白色化と作用素ノルムを1にする正規化を施した観測信号のブロック対角行列である．$\mu_1, \mu_2 \in \mathbb{R}_{++}$ はPDSのパラメータで，$\mu_1\mu_2 \le 1$ とする．K は反復回数で，$\odot$ は要素ごとの乗算を表す．

3.2 音声強調器の学習に用いる教師データ

マスク推定器 $\mathcal{M}_{\mathrm{DNN}}$ は，雑音を含む混合音の振幅スペクトログラム $\mathbf{D} \in \mathbb{R}_{+}^{F\times T}$ から目的音 $\mathbf{S} \in \mathbb{R}_{+}^{F\times T}$ を強調するマスク $\mathbf{M}$ を生成する．学習時の混合音 $\tilde{\mathbf{D}}$

*Plug-and-Play Blind Source Separation by Primal-Dual Splitting and DNN. By Kazuki Matsumoto (Waseda University) and Kohei YATABE (Tokyo University of Agriculture and Technology)

は目的音 $\mathbf{S}$，妨害音 $\mathbf{I}$，多変量標準正規分布に従う雑音 $\mathbf{Z}$ および $\mathbf{I}$ と $\mathbf{Z}$ に係る音量 $a_{\mathrm{I}}, a_{\mathrm{Z}}$ を用い，

$$\tilde{\mathbf{D}}[f,t] = ((\mathbf{S}[f,t]+a_{\mathrm{I}}\mathbf{I}[f,t])/c_1[f]+a_{\mathrm{Z}}|\mathbf{Z}[f,t]|)/c_2[f] \tag{4}$$

で得る[1]．ただし，$\mathbf{c}_1, \mathbf{c}_2 \in \mathbb{R}_{++}^{F}$ は正規化係数であり，各周波数の ℓ_2 ノルムを 1 に正規化する．学習時の目的音 $\tilde{\mathbf{S}}$ も同一の係数を用い以下の式で得る[2]．

$$\tilde{\mathbf{S}}[f,t] = \mathbf{S}[f,t]/(c_1[f]\, c_2[f]) \tag{5}$$

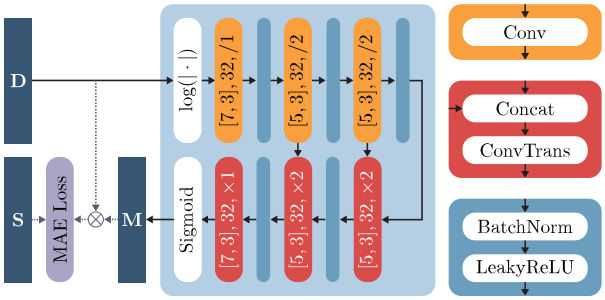
4　評価実験

DNN のアーキテクチャを図–1 に示す．目的音や妨害音は JVS [5] からランダム抽出し，計 10 000 個の教師データを作成した．標本化周波数は 16 kHz となるようダウンサンプリングした．STFT の窓長は 1024 サンプル，シフト長は 512 サンプルとし，その後ランダムな区間を切り出すことで時間フレーム数を 128 に統一した．損失には時間周波数領域での平均絶対値誤差を用いた．最適化には Adam を用い，学習率は 0.01，バッチサイズは 32，エポック数は 30 とした．

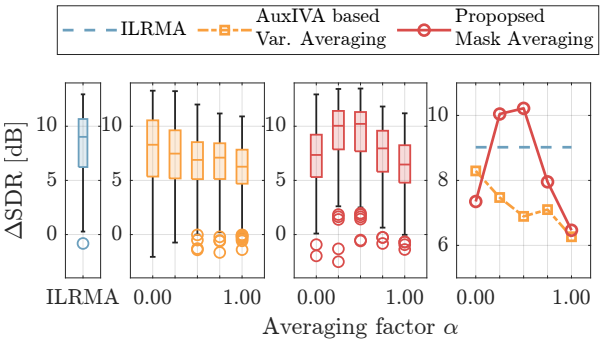
性能評価では SiSEC 2011 の dev1 に含まれる 8 話者の音声に対して室内インパルス応答を畳み込むことで作成した 2 チャンネルの 2 話者混合音を用いた．話者の組み合わせは左右の入れ替えを考慮した 56 組で，音源方向は $(-45°, 30°)$, $(-75°, 30°)$, $(-45°, 60°)$, $(-75°, 60°)$ の 4 組を用いた．音源距離は 1 m，マイク間隔は 8 cm，残響時間は 0.16 秒とした．

比較対象としては AuxILRMA-ISS および AuxIVA-ISS [6] を用いた．IVA に関しては，分散パラメータを DNN で音声強調したパワースペクトログラムと平均化することで，その効果を提案法と比較した．平均化率 α には 0.00，0.25，0.50，0.75，1.00 を用いた．反復回数は 500 回とし，PDS の最適化パラメータは $\mu_1 = \mu_2 = 1$ とした．評価指標は ΔSDR とした．

実験の結果を図–2 に示す．提案法における α が 1 のとき，すなわち単に DNN を PnP したときは IVA を下回る結果となった．しかし，$\alpha = 1/2$ 付近では性能向上が見られ，IVA および ILRMA を上回る分離性能が得られた．このことから，マスク平均化の有効性が確認された．また，分散パラメータの平均化では性能の向上が見られないことから，DNN と IVA の平均化は近接作用素の意味で行う必要があると分かった．



図–1　マスク推定 DNN のアーキテクチャ．畳み込み層や転置畳み込み層については “[カーネルサイズ], チャンネル数, ストライド” をブロック内に表記した．



図–2　評価実験の結果．1 列目は AuxILRMA の結果である．2 列目は AuxIVA に基づく分散パラメータ平均化，3 列目は提案手法の結果で，いずれも $\alpha = 0$ で従来の IVA に一致する．4 列目では各手法の中央値を同一座標軸上にプロットしている．

5　むすび

本稿では音源モデル項に DNN を活用した音声強調器を PnP する新たなブラインド音源分離手法を提案した．IVA と DNN の平均化は両者の相補的な働きを誘導し，音声同士の安定的な分離を実現した．

参考文献

[1] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari and N. Ono. “Independent deeply learned matrix analysis for determined audio source separation.” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, **27**(10), 1601–1615 (2019).

[2] L. Li, H. Kameoka and S. Makino. “Fast MVAE: joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier,” *IEEE ICASSP*, pp. 546–550 (2019).

[3] R. Scheibler and M. Togami. “Surrogate source model learning for determined source separation.” *IEEE ICASSP*, pp. 176–180 (2021).

[4] K. Yatabe and D. Kitamura. “Determined BSS based on time-frequency masking and its application to harmonic vector analysis,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, **29**. pp. 1609–1625 (2021).

[5] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari, “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoust. Sci. Technol.*, **41**(5), 761–768 (2019).

[6] R. Scheibler and N. Ono. “Fast and stable blind source separation with rank-1 updates.” *IEEE ICASSP*, pp. 236–240 (2020).

[1] DNN の学習を簡単にするためには，目的音が妨害音やノイズに対して十分に大きな音である必要がある．また，収束性の観点からは，DNN が分離音を不動点として持つことが望ましい．これらのことを考慮し，振幅係数 a_I および a_Z は 0 以上 0.5 以下の一様分布乱数とした．

[2] この正規化により，$\tilde{\mathbf{S}} = \mathbf{M} \odot \tilde{\mathbf{D}}$ を満たす真のマスク $\mathbf{M}$ で

$$0 \le \mathbf{M}[f,t] = \mathbf{S}[f,t]/\left(\mathbf{S}[f,t] + a_{\mathrm{I}}\mathbf{I}[f,t] + a_{\mathrm{Z}}c_1[f]\,|\mathbf{Z}[f,t]|\right) \le 1$$

が成り立つ．また，この式の分母は目的音と妨害音および音声整形ノイズの和であり，雑音および残響下の混合音を模している．