5S-02

深層ブラインド音源分離を用いた転移学習による環境音分離

合澤 隆拓[1,2]　坂東 宜昭[2]　糸山 克寿[1,3]　西田 健次[1]　中臺 一博[1]
[1]東京工業大学　[2]産業技術総合研究所　[3](株) ホンダ・リサーチ・インスティチュート・ジャパン

1. はじめに

複数の音が混合された観測信号から個別の音源を抽出する音源分離は，音環境理解の基盤技術の一つである．近年，深層学習に基づく手法が，従来の統計的信号処理に基づく手法と比較して音楽や音声の分離において高い性能を達成している [1]．一方，人間の耳のように音源の種類に関係なく聞き分ける技術が実現できれば，人間とロボットの自然なコミュニケーションが可能になる．

このような技術を実現するために，FUSS データセット [2] など，多様な種類の音で構成されたデータセットが提案されている．あらゆる環境で収録した混合音を分離するためには，音声や音楽の分離と比較して多様な種類のスペクトル構造を学習する必要がある．学習した混合音で推論すれば高い性能が得られるが，未知混合音では得られにくい．目的のタスクごとにモデルを用意するのはその都度大量のデータで学習させる必要があり非効率なため，事前に音源分離を学習させたモデルを目的に合わせた混合音で転移学習できれば，学習コストを抑えることができる．

本研究では，非線形ブラインド音源分離法の一つである深層フルランク空間相関分析 [3]（Neural FCA）を用いた目的環境への教師なし転移学習を提案する．本手法ではまず，様々な種類の環境音が収録されたデータセットを用いて音源分離 DNN を教師あり事前学習し，汎用的な音源分離を学習したモデルを得る．この事前学習モデルを Neural FCA を用いて目的環境の混合音のみから教師なし転移学習することで，事前学習データに含まれない未知の音源信号でも高性能な分離ができるモデルが構築できる．FUSS データセットを多チャネル拡張した混合音を用いて，提案法の有効性を確認した．

2. 深層フルランク空間相関分析

提案法の基盤となる Neural FCA について述べる．

2.1 生成モデル

本研究では，弱ラベルとして各音源のアクティベーション変数を導入した生成モデルを定義する [4]．アクティベーション変数は，音源 n の時間フレーム t でのアクティブ状態 $u_{nt} \in \{0,1\}$ を表す．M チャネルの観測混合音 $\mathbf{x}_{ft}$ を以下のように N_{src} 個の音源信号と 1 個の雑音信号の像の和 $\mathbf{s}_{nft} \in \mathbb{C}^M$ で表す．

$$\mathbf{x}_{ft} = \sum_{n \in \mathfrak{N}_t} \mathbf{s}_{nft} \tag{1}$$

ただし，$\mathfrak{N}_t = \{n | u_{nt} = 1\} \cup \{N_{\mathrm{src}} + 1\}$ は時間 t に存在する音源の集合であり．$n = N_{\mathrm{src}} + 1$ は，雑音クラスである．音源信号の像 $\mathbf{s}_{nft} \in \mathbb{C}^M$ は，音源 n の特徴を表す低次元の潜在ベクトル $\mathbf{z}_{nt} \in \mathbb{R}^D$ を用いて以下のような多変量複素ガウス分布で表現する．

$$\mathbf{s}_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, g_{\theta,f}(\mathbf{z}_{nt})\mathbf{H}_{nf}\right), z_{ntd} \sim \mathcal{N}(0,1) \tag{2}$$

ここで，$g_{\theta,f} : \mathbb{R}^D \to \mathbb{R}_+$ は，$\mathbf{z}_{nt}$ からパワースペクトル密度（PSD）を出力するパラメータ θ を持つ DNN である．$\mathbf{H}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}} \in \mathbb{S}_+^{M \times M}$ は周波数 f における音源 n の空間相関行列 (SCM) であり，$\mathbf{a}_{nf} \in \mathbb{C}^M$ は音源 n のステアリングベクトルである．音源の小さな変動を許容するため，SCM をフルランクに緩和する．以上より，観測混合音 $\mathbf{x}_{ft}$ は，以下の多変量複素ガウス分布に従う．

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n \in \mathfrak{N}_t} \mathbf{s}_{nft}\right) \tag{3}$$

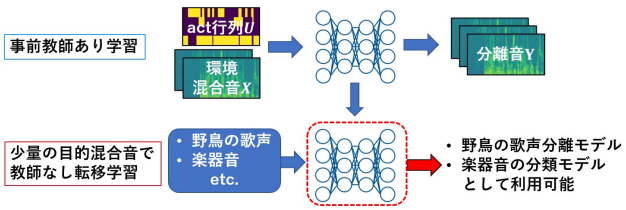

図 1: 目的環境音での教師なし転移学習

2.2 償却変分推論を用いた教師なし学習

多チャネル混合音 $\mathbf{X}$ とアクティベーション変数 $\mathbf{U}$ から，音源モデル $g_{\theta,f}$ と潜在変数 $\mathbf{Z}$ 及び空間相関行列 $\mathbf{H}$ の推論モデルを教師なし学習する．混合音 $\mathbf{X}$ に対する対数周辺尤度を最大化するように学習されるが，直接計算困難なため，推論モデル $q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{U})$ を導入した償却変分推論 [5] を用いる．

$$q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{U}) = \prod_{n,t,d} \mathcal{N}_{\mathbb{C}}\left(z_{ntd} \mid \mu_{\phi,ntd}(\mathbf{X},\mathbf{U}), \sigma^2_{\phi,ntd}(\mathbf{X},\mathbf{U})\right) \tag{4}$$

ただし，$\mu_{\phi,ntd}(\mathbf{X},\mathbf{U}) \in \mathbb{R}$ と $\sigma^2_{\phi,ntd}(\mathbf{X},\mathbf{U}) \in \mathbb{R}_+$ は，特徴量 $\mathbf{X},\mathbf{U}$ を入力とするパラメータ ϕ を持つ DNN の出力である．空間相関行列 $\mathbf{H}_{nf}$ は以下のように出力する．

$$\mathbf{H}_{nf} = \sum_{t \in \{t | u_{nt}=1\}} m_{ntf}(\mathbf{X},\mathbf{U})\mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathsf{H}} \tag{5}$$

$m_{ntf}(\mathbf{X},\mathbf{U}) \in [0,1]$ は，時間周波数マスクであり，$\mathbf{X},\mathbf{U}$ を入力とする DNN の出力である．$\mathrm{tr}(\mathbf{H}_{nf}) = M$ となるように正規化する．変分償却推論では，学習データに対する以下の変分下限 $\mathcal{L}_x$ を最大化するように，DNN の

Environmental sound separation by transfer learning based on deep neural source separation: *T. Aizawa, Y. Bando, K. Itoyama, K. Nishida, K. Nakadai*

パラメータ θ と ϕ，SCM $\mathbf{H}_{nf}$ を同時に最適化する．

$$\mathcal{L}_x = \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{X}|\mathbf{Z},\mathbf{H},\mathbf{U})] - \mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{U})|p(\mathbf{Z})] \quad (6)$$

$\mathcal{L}_x$ の第一項の最大化は混合音と分離音の和が近くなること，第二項の最小化は，潜在変数 $\mathbf{Z}$ が事前分布から離れないことを促す．

3. 目的環境音での教師なし転移学習

本手法では，最大 N_{src} 個の音源が含まれている状況で，以下の問題設定により音源分離を行う．

入力: Mch 混合音 $\mathbf{x}_{ft} \in \mathbb{C}^M$ と音源 $n = 1, \ldots, N_{\mathrm{src}}$ の時間フレーム t でのアクティベーション $u_{nt} \in \{0, 1\}$
出力: 音源 n の分離音 $\hat{s}_{nft} \in \mathbb{C}$

3.1 学習方法

汎用的な音源分離を学習するために，様々な環境混合音と教師音のペアを用いて事前に教師あり学習を行う．ここでは，2. 節の手法を教師あり学習に拡張する．教師あり学習では以下の変分下限 $\mathcal{L}$ を最小化するように，音源モデル $g_{\theta,f}$，潜在変数の推論モデル q_ϕ と空間相関行列の推論モデル m_{ntf} を学習させる．

$$\mathcal{L} = \mathcal{L}_x + \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{S}|\mathbf{Z},\mathbf{H},\mathbf{U})] - \mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{U})|p(\mathbf{Z})] \quad (7)$$

$\mathcal{L}$ の第二項は，教師音と分離音の板倉斎藤距離 [6] と等価である．

学習したモデルを目的環境の混合音のみを用いて教師なし転移学習する．(6) 式を最大化するように，モデルパラメータをファインチューニングする．事前に汎用的な音源分離を学習しているため，ランダム初期値から学習する場合と比較してデータ数が少量でも性能が良い分離モデルを獲得できる．

3.2 推論方法

2 回の学習で得た音源分離 DNN を使い，未知の混合音を分離する．実環境音を用いる際は，音源定位手法や音響イベント検出などを用いてアクティベーション変数 $\mathbf{U}$ を取得し，混合音 $\mathbf{X}$ とともに入力する．分離音 $\hat{s}_{nft}$ は，DNN が推定した PSD と SCM から，マルチチャネルウィーナフィルタを用いて得る．

4. 評価実験

FUSS データセット [2] を多チャネル拡張し，提案法を評価した．

4.1 データセット

FUSS データセットは，多様な環境音が収録され，6 チャネル混合音の中に 1-4 個の音源が含まれている．また，音声信号における性能との比較のため，Spatialized WSJ0-mix データセットから音源数が 2，3 の 4 チャネル混合音でも実験を行った．各録音は 16kHz で収録され，学習，検証および評価セットに分割されている．

4.2 実験設定

提案法のネットワークアーキテクチャは [4] と類似の構成とした．スペクトログラムは短時間フーリエ変換によって求め，窓長 512 サンプル，ホップ長 128 サンプルとした．潜在変数 $\mathbf{z}_{nt}$ の次元は，音声では 50，環境音では 100 とした．音声は 0.1，環境音は 0.25 のドロップアウトを行った．また，(6) (7) 式の KL 項の重みを周期的に変化させる KL アニーリング [3] を行った．まず，学習データで教師あり学習させ，その後，未知混合音として評価データを使い転移学習を行った．評価データ数は，学習データ数の 6-15%と比較的少量である．環境音の教師あり学習は 50 エポック，それ以外の学習は 200 エポックで行った．教師あり学習（転移学習前）と未知混合音（本実験では評価データ）に対する転移学習の性能を評価した．

表 1: 評価データによる平均 SDR(dB)

	転移学習前	転移学習後
音声 ($N_{\mathrm{src}} = 2$)	11.0 ±4.20	14.2 ±4.79
音声 ($N_{\mathrm{src}} = 3$)	2.64 ±4.30	5.00 ±4.92
環境音 ($N_{\mathrm{src}} = 1$)	10.6 ±4.44	10.8 ±4.50
環境音 ($N_{\mathrm{src}} = 2$)	5.24 ±7.54	5.65 ±7.83
環境音 ($N_{\mathrm{src}} = 3$)	1.65 ±8.81	2.29 ±9.06
環境音 ($N_{\mathrm{src}} = 4$)	−0.46 ±9.65	−0.04 ±10.2

4.3 実験結果

表 1 に評価データでの実験結果を SDR で示す．音声では，転移学習により未知混合音に対する SDR が 2 から 3 dB 程度，環境音では 0.2 から 0.6 dB 程度向上した．音の種類により性能向上量に差が生じた要因は，未知混合音でも音声はスペクトル構造が類似しているため，転移学習によりモデルを適応させやすかった一方，環境音の未知混合音にはデータ数は少ないものの，様々な構造が含まれており，モデルの表現力に限界があるからだと考えられる．音源数と性能向上量に傾向はみられなかった．

5. おわりに

本稿では，事前学習モデルを用いた目的環境音での教師なし転移学習について述べた．音声では 2 から 3 dB 程度性能向上が見られたが，様々なスペクトル構造が含まれる環境音ではわずかな向上にとどまった．今後は，実環境音や限定されたタスク（例えば，野鳥の鳴き声の分離）での実験により，提案法の有効性をさらに検証する．

謝辞: 本研究の一部は，JST ACT-X JPMJAX200N および NEDO の支援を受けた．

参考文献

[1] Ilya et al Kavalerov. Universal sound separation. In *2019 IEEE WASPAA*, 175–179, 2019.
[2] Scott Wisdom *et al.* What's all the FUSS about free universal sound separation data? *CoRR*, 2020.
[3] Y. Bando et al. Neural full-rank spatial covariance analysis for blind source separation. *IEEE SPL*, 28:1670–1674, 2021.
[4] Y.Bando *et al.* Weakly-Supervised Neural Full-Rank Spatial Covariance Analysis for a Front-End System of Distant Speech Recognition. In *Proc. Interspeech 2022*, 3824–3828, 2022.
[5] D. P. Kingma et al. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
[6] F. Itakura. Analysis synthesis telephony based on the maximum likelihood method. *ICA*, 1968.