5S-01

音源定位・分離の同時学習に基づく移動音源の深層ブラインド音源分離

宗像 北斗 [1,2]　坂東 宜昭 [2]　武田 龍 [1]　駒谷 和範 [1]　大西 正輝 [2]

1 大阪大学 産業科学研究所　2 産業技術総合研究所

1. はじめに

音源分離は互いに重なり合い観測された混合信号から音源信号を推定する技術である．近年，深層学習の発達に伴い教師あり学習に基づく音声・音楽分離手法が高い性能を達成している．一方，教師となる音源信号の収集は高コストであるため，日常生活のさまざまな音源を分離するにはいまだ課題がある．そこで本研究では，音源信号の収集を必要としない教師なし手法の確立を目指す．

教師なし音源分離法の一つであるブラインド音源分離（BSS）は混合信号を表す確率的生成モデルのパラメータを推論することで分離を行う．深層フルランク空間相関分析（Neural FCA）ではフルランク空間モデル [1] に深層ニューラルネットワーク (DNN) に基づく非線形音源モデルが導入されている [2]．Neural FCA は大量の混合信号に対して，深層生成モデルの周辺尤度を最大化するよう DNN を最適化することで音源の特徴を学習する．

多くの BSS と同様に，Neural-FCA は音源の静止を仮定しており，音源の移動により分離性能が大きく低下する．解決法の一つとして，音源定位の活用が挙げられる．音源定位で求められた到来方向（DoA）に基づく幾何的制約をステアリングベクトルに導入することで，音源の移動が表現できる [3]．さらに音源定位を生成モデルに統合し，分離と定位を同時学習する手法は，全体最適化により，分離と定位の誤差を相補的に低減できる [4]．

本稿では Neural FCA を拡張し，移動音源の分離と定位の教師なしでの同時学習を実現する（図 1）．具体的にはまず，時変な空間相関行列および DoA に基づく多チャンネル混合信号の生成過程を設計する．この生成過程に基づき，混合信号とマイクロホン配置のみから周辺事後確率を最大化するように定位，分離モデルを同時に学習する．実験の結果，提案法は従来法と比較して移動音源の分離および定位精度を大きく改善した．

2. 時変・深層フルランク空間相関分析

提案法では Neural FCA の生成・推論モデルを時変モデルに拡張する．生成モデルでは音源定位結果を時変な空間モデルに統合し，推論モデルでは混合信号から DoA を推論する．生成・推論モデルは混合信号のみから音源定位および分離を同時学習する．以降の処理は時間・周波数領域で行うとし，$f = 1, \ldots, F$ および $t = 1, \ldots, T$ はそれぞれ周波数及び時間インデックスを表す．

2.1 移動音源からなる混合信号の生成モデル

提案法では M チャンネル混合信号 $\mathbf{x}_{ft} \in \mathbb{C}^M$ を N 個の音源 $s_{nft} \in \mathbb{C}$ と雑音 $\mathbf{n}_{ft} \in \mathbb{C}^M$ の和で表す．

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{a}_{nft} s_{nft} + \mathbf{n}_{ft} \tag{1}$$

ここで $\mathbf{a}_{nft} \in \mathbb{C}^M$ は音源 n の時間 t でのステアリングベクトルとする．音源のパワースペクトル密度（PSD）

Neural Blind Source Separation for Moving Sound Sources Based on Joint Separation and Localization: *H. Munakata, Y. Bando, R. Takeda, K. Komatani, and M. Onishi*

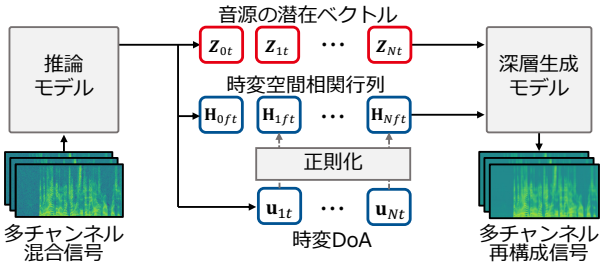

図 1: 生成・推論モデルからなる提案手法の全体像

s_{nft} は多変量標準正規分布に従う潜在ベクトル $\mathbf{z}_{nt} \in \mathbb{R}^D$ により次式のように表されると仮定する．

$$s_{nft} = g_{\theta,f}(\mathbf{z}_{nt}), \quad \mathbf{z}_{nt} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

ここで $g_{\theta,f} : \mathbb{R}^D \to \mathbb{R}$ はパラメータ θ を持つ DNN による非線形変換とする．以上より，混合信号の尤度関数は次式のような多変量複素正規分布に従う．

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=0}^{N} g_{\theta,f}(\mathbf{z}_{nt}) \mathbf{H}_{nft}\right) \tag{3}$$

ここで $\mathbf{H}_{nft} = \mathbf{a}_{nft}\mathbf{a}_{nft}^{\mathsf{H}} \in \mathbb{S}_{+}^{M\times M}$ は音源（$n = 1, \ldots, N$）とノイズ（$n = 0$）の時変な空間相関行列（SCM）である．さらに $\mathbf{a}_{nft}$ の統計的なゆらぎを許容するため，$\mathbf{H}_{nft}$ はフルランクな正定値行列へと緩和する．

音源定位結果と統合するために，各音源の SCM は DoA を表す単位ベクトル $\mathbf{u}_{nt} \in \mathbb{R}^3$ ($\|\mathbf{u}_{nt}\| = 1$) で条件づけられた以下のような事前分布を仮定する．

$$\mathbf{H}_{nft} \mid \mathbf{u}_{nt} \sim \mathcal{IW}_{\mathbb{C}}(\nu, (\nu + M)\mathbf{G}_{nft}(\mathbf{u}_{nt})) \tag{4}$$

ここで $\mathcal{IW}_{\mathbb{C}}(\nu, \mathbf{\Gamma})$ は複素逆ウィシャート分布で，ν は分布の自由度を表すハイパーパラメータである．この分布の最頻値は以下の事前 SCM で定義する．

$$\mathbf{G}_{nft}(\mathbf{u}_{nt}) = \mathbf{b}_f(\mathbf{u}_{nt})\mathbf{b}_f(\mathbf{u}_{nt})^{\mathsf{H}} + \epsilon\mathbf{I} \tag{5}$$

ここで $\epsilon > 0$ は $\mathbf{G}_{nft}(\mathbf{u}_{nt})$ を正定値にするための微小な値で，$\mathbf{b}_f(\mathbf{u}_{nt})$ は幾何計算に基づく $\mathbf{u}_{nt}$ に対するステアリングベクトルである．また $\mathbf{n}_{ft}$ は拡散性を仮定しているため，$\mathbf{G}_{nft}(\mathbf{u}_{nt})$ の代わりに単位行列を用いる．

2.2 時変パラメータに対する推論モデル

提案法の推論モデルは混合信号から 2.1 節で導入した $\mathbf{z}_{nt}$，$\mathbf{H}_{nft}$ および $\mathbf{u}_{nt}$ を予測する．$\mathbf{z}_{nt}$ は従来の Neural FCA と同様の変分事後分布を導入して求める．

$$q_\phi(\mathbf{Z} \mid \mathbf{X}) = \prod_{n,t,d} \mathcal{N}\left(z_{ntd} \mid \mu_{\phi,ntd}(\mathbf{X}), \sigma^2_{\phi,ntd}(\mathbf{X})\right) \tag{6}$$

ここで $\mu_{\phi,ntd}(\mathbf{X}) \in \mathbb{R}$ と $\sigma^2_{\phi,ntd}(\mathbf{X}) \in \mathbb{R}_+$ はパラメータ ϕ を持つ DNN の $\mathbf{X}$ を入力とした際の出力である．時変 SCM $\mathbf{H}_{nft}$ は以下のように $\mathbf{x}_{ft}$ に時間・周波数マスクをかけ，移動平均をとった値を用いる．

$$\mathbf{H}_{nft} \leftarrow \gamma_0 \mathbf{G}_{nft}(\mathbf{u}_{nt}) + \sum_{t'=0}^{T} \gamma^{|t-t'|} w_{\phi,nft'}(\mathbf{X}) \frac{\mathbf{x}_{ft'}\mathbf{x}_{ft'}^{\mathsf{H}}}{\|\mathbf{x}_{ft'}\|^2}$$

ここで $w_{\phi,nft}(\mathbf{X}) \in [0, 1]$ は DNN が出力した時間・周波数マスクで，$\gamma_0 \in \mathbb{R}_+$ および $\gamma \in (0, 1]$ はそれぞれ定位結

果の利用率，移動平均の減衰率を表すハイパーパラメータである．DoA $\mathbf{u}_{nt}$ は DNN の出力 $\tilde{\mathbf{u}}_{\phi,nt}(\mathbf{X}) \in \mathbb{R}^3$ に以下のように移動平均をとった値を用いる．

$$\mathbf{u}_{nt} \leftarrow \sum_{t'=0}^{T} \eta^{|t-t'|} \tilde{\mathbf{u}}_{\phi,nt'}(\mathbf{X}) \tag{7}$$

ここで $\eta \in (0, 1]$ は移動平均の減衰率を表すハイパーパラメータである．$\mathbf{u}_{nt}$ は単位ベクトルになるよう正規化する．本節で導入した移動平均は，$\mathbf{H}_{nft}$ および $\mathbf{u}_{nt}$ に時間的連続性を持たせる．

2.3 変分償却推論に基づく教師なし学習

提案法は変分償却推論に基づき，多チャンネル混合信号およびマイクロホン配置のみを用いて音源の生成モデル $g_{\theta,f}$ および推論モデルを同時に学習する．提案法では周辺事後確率 $p_{\theta,\phi}(\mathbf{H}_\phi \,|\, \mathbf{X}, \mathbf{U}_\phi)$ の変分下界 $\mathcal{L}'_{\theta,\phi}$ を最大化するよう学習する．

$$\mathcal{L}'_{\theta,\phi}(\mathbf{X}) = \mathbb{E}_{q_\phi}\left[\log p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{H})\right] + \log p(\mathbf{H}_\phi \mid \mathbf{U}_\phi) - \mathcal{D}_{\mathrm{KL}}\left[q_\phi(\mathbf{Z} \mid \mathbf{X}) \mid p(\mathbf{Z})\right]$$

ここで $\mathbf{H}_\phi$ と $\mathbf{U}_\phi$ はそれぞれ 2.2 節で求めた SCM および DoA の集合である．第 1 項から第 3 項はそれぞれ混合信号の周辺尤度，(4) 式に基づく正則化項，さらに $\mathbf{z}_{nt}$ の推論値と事前分布間の Kullback-Leibler ダイバージェンスを表す．DNN のパラメータ θ および ϕ は確率勾配法により更新する．一度学習した DNN は，未知の混合信号の分離および各移動音源の定位に使用できる．分離信号 $\mathbf{Y}_{nft} = g_{\theta,f}(\boldsymbol{\mu}_{\phi,nt}(\mathbf{X}))\mathbf{H}_{nft} \in \mathbb{C}^M$ は多チャンネルウィーナーフィルタから得られる．

3. 評価実験

移動音源，静止音源による 2 種類の数値混合データセット [5] を用いて提案法の分離，定位性能を評価した．

3.1 データセット

はじめに移動音源からなる多チャンネル混合信号のデータセットを作成した．音源信号は WSJ0 コーパスの音源に鏡像法により生成したインパルス応答を畳み込んで生成した．音源数を 2，マイクのチャンネル数を 6 とした．インパルス応答は 0.1 s ごとに変化させた．音源の移動速度は 0°/s から 45°/s からランダムにサンプルした．残響時間は 200 ms で，サンプリングレートは 16 kHz とした．学習，検証，評価用にそれぞれ 20000, 5000，3000 サンプルの混合信号を生成した．同様にして，静止音源からなるデータセットも生成した．

3.2 実験条件

各 DNN は 1 次元畳み込み層を用いて構成した．生成モデルは 3 層の 1×1 畳み込み層で構成した．推論モデルは 32 層の膨張畳み込み層 [6] で構成し，提案法では時間 · 周波数マスクおよび DoA の出力層を追加した．提案法は反復推定を必要とせず，オンライン処理への応用も容易なため，因果的膨張畳み込み層 [6] を用いたオンラインモデルでも実験をおこなった．オンラインモデルでは推論時のみ 2.2 節の移動平均を t' が 0 から $t + 100$ の範囲で計算した．入力特徴量には対数スペクトログラムに加え，マイクロホン配置を利用した特徴として球面上に一様に配置した 1000 点から計算した混合信号の DoA [7] を用いた．ハイパーパラメータ D，ν，ϵ，γ_0，および η はそれぞれ 50，$M + 1$，0.001，0.1，および 0.99 とした．また γ は音源クラスに対しては 0.99，ノイズクラスに対しては 1（時不変）とした．過度な正則化を防ぐため，変分下界の $\log p(\mathbf{H}_\phi \,|\, \mathbf{U}_\phi)$ は 0.001 倍して学習した．ドロップアウトは 0.1 としたが，オンラインモデルでは学習を促すため 0 とした．スペクトログラムは窓長 512 サンプル，ホップ長 128 サンプルの短時間フーリエ変換で得た．学習時には [2] と同様に KL アニーリングを行った．ベースラインとなる FastMNMF2 [1] の基底数は 8 とした．DoA-HMM [4] は MUSIC を用いて初期化した．

表 1: 各データセットに対する分離と定位性能

手法	静止音源 SDR	静止音源 DoA 誤差	移動音源 SDR	移動音源 DoA 誤差
FastMNMF2	12.82	–	4.01	–
DoA-HMM	7.83	2.81	7.96	3.89
Neural FCA	**16.06**	–	8.27	–
TV Neural FCA [5]	14.21	**2.48**	**12.53**	**3.04**
＋ オンライン拡張	11.70	3.34	10.09	7.41

3.3 実験結果

提案法 (TV Neural FCA) およびベースラインの分離，定位性能をそれぞれ信号対歪比（SDR）および DoA 誤差 [8] で評価し，表 1 に示す．移動音源データでは，時不変モデルである FastMNMF2 および従来の Neural FCA の性能は大きく低下した．一方，提案法の SDR は 12 dB 以上であり，さらに静止音源データにおいても FastMNMF2 および DoA-HMM より高い分離性能を示した．また提案法は二つのデータいずれにおいても DoA-HMM より高い定位性能を示した．提案法はオンライン化に伴い，学習初期に定位に失敗するようなパラメータに陥ることがあったため，学習が安定化するような初期値を事前に選択した．オンライン化により分離，定位性能は低下しているが，いずれのデータにおいてもオフライン法である DoA-HMM を上回る分離性能を示した．

4. おわりに

移動音源の分離，定位を教師なしで同時学習する手法を提案した．実験の結果，提案法の分離，定位精度は従来法と比較して改善した．今後はオンラインモデルの分離，定位精度の改善を目指す．

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] K. Sekiguchi, et al. Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation. *IEEE/ACM TASLP*, Vol. 28, pp. 2610–2625, 2020.

[2] Y. Bando, et al. Neural full-rank spatial covariance analysis for blind source separation. *IEEE SPL*, Vol. 28, pp. 1670–1674, 2021.

[3] D. Kounades-Bastian, et al. A variational EM algorithm for the separation of time-varying convolutive audio mixtures. *IEEE/ACM TASLP*, Vol. 24, No. 8, pp. 1408–1423, 2016.

[4] T. Higuchi, et al. Underdetermined blind separation and tracking of moving sources based ONDOA-HMM. In *Proc. ICASSP*, pp. 3191–3195, 2014.

[5] H. Munakata, et al. Joint separation and localization of moving sound sources based on neural full-rank spatial covariance analysis. *IEEE SPL (under review)*.

[6] Y. Luo, et al. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM TASLP*, Vol. 27, No. 8, pp. 1256–1266, 2019.

[7] M. Togami. Spatial constraint on multi-channel deep clustering. In *Proc. ICASSP*, pp. 531–535, 2019.

[8] S. Adavanne, et al. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE JSTSP*, Vol. 13, No. 1, pp. 34–48, 2018.