音源定位・到来方向推定

島田一希（ソニー AI），高橋秀介（ソニーグループ株式会社），光藤祐基（ソニー AI）
kazuki.shimada@sony.com, shusuke.takahashi@sony.com, yuhki.mitsufuji@sony.com

1. 音源定位・到来方向推定とは

複数のマイクロホンを空間的に配置したものをマイクロホンアレーと呼ぶ．音源定位とは，マイクロホンアレーで観測されたマルチチャネル音響信号に基づき，マイクロホンアレーからの音源位置を推定するタスクである．実際には，マイクロホンアレーからの距離は推定せず，マイクロホンアレーから見て音源がどの方向から到来したかを推定する，すなわち到来方向推定として解かれることが多い．音源定位には多くの応用があり，音源分離や音声強調，あるいはヒューマンロボットインタラクションなどで活用される．例えばロボットが話者の位置を推定できることは，インタラクションにおいて重要である．

以前の音源定位の記事では，Multiple Signal Classification（MUSIC）法など信号処理に基づく手法が解説された．信号処理に基づく手法はこの分野で長く使われているが，雑音や残響，複数音源の状況では性能が落ちるという課題がある．そのような難しい状況に対応できる手法を目指し，近年深層学習に基づく手法の取組みが増えている[1]．多くの論文で深層学習に基づく手法が従来の信号処理に基づく手法よりも高い性能を示しており，例えば畳込みリカレントニューラルネットワークが MUSIC 法よりも残響下で方向推定誤差を減らすことが報告されている[2]．また深層学習に基づく音源定位の発展に伴い，音源定位と音響イベント検出を同時解決する，すなわち，音響イベントの種類，区間，到来方向を同時に推定する，音響イベント定位及び検出（SELD：Sound Event Localization and Detection）というタスクも広く取り組まれている[3]．本稿では深層学習に基づく音源定位，そして音響イベント定位及び検出（SELD）を紹介する．

2. 深層学習に基づく音源定位

深層学習に基づく音源定位システムの基本構成を図 1 に示した[1]．まずマイクロホンアレーで収録したマルチチャネル音響信号を特徴抽出モジュールで処理して入力特徴量を得る．次に入力特徴量をニューラルネットワークで処理して，音源位置または到来方向を推定する．

この基本構成には二つの理由がある[1]．一つは音源位置の情報はマイクロホンアレーで収録したマルチチャネル音響信号に含まれることである．ある音源から各マイクロホンまでの伝搬の違いから，収録信号の各チャネル間では位相や振幅の差が生じる．具体的にあるマイクロホン i での収録信号 $x_i(t)$ は，ある音源 j の信号 $s_j(t)$ をその音源 j からマイクロホン i までのインパルス応答 $a_{i,j}(t)$ で畳み込むことで，次のように得られる．

$$x_i(t) = a_{i,j}(t) * s_j(t) + n_i(t) = \sum_{\tau=0}^{T-1} a_{i,j}(\tau) s_j(t-\tau) + n_i(t) \quad (1)$$

ここで，$n_i(t)$ はマイクロホン i での雑音である．$*$ は畳込みを示し，ディジタル信号の場合 t，τ は離散時間インデックス，T はインパルス応答の有効範囲である．音源が複数，ここでは J 個とする，の場合は次が得られる．

$$x_i(t) = \sum_{j=1}^{J} a_{i,j}(t) * s_j(t) + n_i(t) \quad (2)$$

この構成のもう一つの理由は，ニューラルネットワークが大量の訓練データから入出力の関係を学習できることである．マルチチャネル音響信号と音源位置の関係は一般に複雑であり，式(2)のように複数音源で雑音や残響がある場合は特に複雑だが，ニューラルネットワークを使うことで訓練データからその関係を学習できる．

入力特徴量，ニューラルネットワーク構造，出力方式などで様々な取組みがある[1]．入力特徴量は，信号波形を直接入力とする取組みもあるが他分野と比べ少なく，Generalized Cross-Correlation Phase Transform（GCC-PHAT）など信号処理に基づく特徴量も広く使われている．ニューラルネットワーク構造は畳込みリカレントニューラルネットワークを用いることが多く，近年は Transformer の導入も進む．音源位置または到来方向の推定結果を得るための出力方式は大

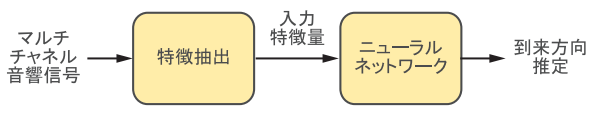

図 1　深層学習に基づく音源定位システムの基本構成　“A Survey of Sound Source Localization with Deep Learning Methods”[1] の図 1 を参考に著者が作成．マルチチャネル音響信号から抽出した入力特徴量をニューラルネットワークで処理して到来方向を推定する．

本会ハンドブック「知識の森」
https://www.ieice-hbkb.org/portal/doc_index.html

きく分けて分類方式と回帰方式の二つがある[4),5)]．分類方式では到来方向を幾つかの部分に分けてそれぞれを各クラスとしてどのクラスに音源が位置するかの分類問題を解き，回帰方式では直交座標系（x, y, z）あるいは極座標系（θ, φ, r）での音源位置を回帰問題として解いている．

3. 音響イベント定位及び検出（SELD）

音響イベント定位及び検出（SELD）は，音源定位と音響イベント検出を同時解決するタスクである．SELD は環境音分析のタスクでもあり，「いつ」「何の音が発生したか」の推定に加え，「どこで音が発生したか」も同時に推定するタスクとなる．SELD システムは音源定位システムと音響イベント検出システムの組合せでも実現できるが，複数音源時にそれぞれの音響イベントへの到来方向の割当が容易でない，また各システムの組合せでは SELD として最適化されない，という課題がある．そこで SELD システムはニューラルネットワークで定位と検出を統合して同時に推定することが一般的であり，代表的なものに SELDnet がある[3)]．SELDnet の基本構成を図 2 に示した．マルチチャネル音響信号から入力特徴量を得るところまでは深層学習に基づく音源定位と同様である．入力特徴量をニューラルネットワークで処理して，各音響イベントクラスごとのアクティビティ（イベントが発生していれば 1 を，そうでなければ 0 をとる）及び到来方向ベクトルを分岐して出力する．アクティビティはバイナリークロスエントロピーを，到来方向ベクトルは平均二乗誤差を損失関数として同時に学習する．

SELDnet を改良した手法として，Activity-coupled Cartesian Direction of Arrival（ACCDOA）[6)] や Event Independent Network（EIN）[7)] が知られている．ACCDOA の基本構成を図 3 に示した．SELDnet は異なる損失関数を使用して学習するため，損失関数間のハイパパラメータ調整が必要になる．ACCDOA ベクトルは直交座標系における到来方向ベクトルの長さをアクティビティに割り当てるものであり，この ACCDOA ベクトルに関する平均二乗誤差のみを損失関数として SELD システムを訓練できる[6)]．そのため損失関数間のハイパパラメータ調整がいらず，そして分岐もないためモデルサイズも小さくできる．EIN の基本構成を図 4 に示した．SELDnet は音響イベントクラスごとに出力するため，同じクラスの重なりに対応できない．EIN はトラックごとの出力を導入し，同じ音響イベントクラスが重なっても異なるトラックに出力できる[7)]．各トラックには音源がそれぞれ割り当てられ，各トラックには対応する音源の音響イベントクラスと到来方向ベクトルを出力するよう学習する．

そのほかに，入力特徴量[8)]，ニューラルネットワーク構造，データオーギュメンテーション[9),10)] などの取組みがある．入力特徴量は音源定位に有効な特徴量に加え，音響イベント検出に有効な特徴量である対数メルスペクトログラムや振幅スペクトログラムが使われる．また SELD に特化した Spatial Cue-Augmented Log-Spectrogram（SALSA）という特徴量も提案されている[8)]．ニューラルネットワーク構造は深層学習に基づく音源定位と同様，畳込みリカレントニューラルネットワークや Transformer を用いることが多い．SELD の教師あり学習では，音響イベントの種類，区間，到来方向のラベルが必要になるが，そのようなデータは少ないため，データオーギュメンテーションも活発に研究されている[9),10)]．

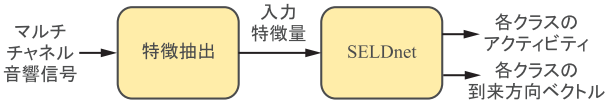

図 2　SELDnet[3)] の基本構成　　入力特徴量をニューラルネットワークで処理してクラスごとのアクティビティと到来方向ベクトルを出力する．

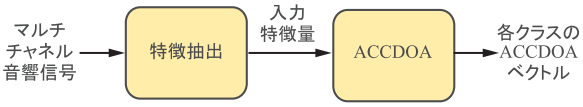

図 3　Activity-coupled Cartesian Direction of Arrival（ACCDOA）[6)] の基本構成　　ニューラルネットワークからクラスごとの ACCDOA ベクトルを出力する．

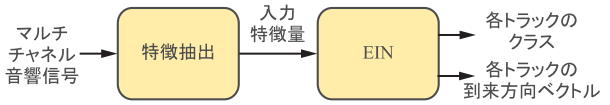

図 4　Event Independent Network（EIN）[7)] の基本構成　　ニューラルネットワークから各トラックに対応する音源の音響イベントクラスと到来方向ベクトルを出力する．

SELD は環境音分析のコンペティション Detection and Classification of Acoustics Scenes and Events（DCASE）Challenge[注 1)] において継続して取り組まれている．このコンペティションで使われたデータセット Sony-TAU Realistic Spatial Soundscapes 2023（STARSS23）は，SELD システムを実収録で評価できるため，広く使われている[11)]．

4. おわりに

本稿では音源定位，特に近年多くの取組みがある深層学習に基づく音源定位，また音響イベント定位及び検出（SELD）を解説した．今回紹介しなかった信号処理に基づく音源定位については以前の記事を参照してほしい．深層学習に基づく音源定位についてより詳しく知りたい場合はサーベイ論文“A Survey of Sound Source Localization with Deep Learning Methods”[1)] をお薦めする．SELD については前述した DCASE Challenge の該当タスクや本稿で紹介した論文を参照されるとよいだろう．現在も，様々なマイクロホンアレーに対応するよう学習する音源定位[12)]，マイクロホンアレーの移動を考慮する SELD[13)]，音響イベントクラスをテキストで指定する SELD[14)] など研究が盛んに行われており，今後も注目の領域となっている．

文　　献

（1）　P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, “A survey of sound source localization with deep learning methods,” J. Acoust. Soc. America, vol. 152, no. 1, pp. 107-151, 2022.

（注 1）　https://dcase.community/

（2） S. Adavanne, A. Politis, and T. Virtanen, “Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network,” Proc. IEEE EUSIPCO, 2018.
（3） S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” IEEE JSTSP, vol. 13, no. 1, pp. 34-48, 2018.
（4） Z. Tang, J.D. Kanu, K. Hogan, and D. Manocha, “Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks,” Proc. Interspeech, 2019.
（5） L. Perotin, A. Défossez, E. Vincent, R. Serizel, and A. Guérin, “Regression versus classification for neural network based audio source localization,” Proc. IEEE WASPAA, 2019.
（6） K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, “ACCDOA : Activity-coupled cartesian direction of arrival representation for sound event localization and detection,” Proc. IEEE ICASSP, 2021.
（7） Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M.D. Plumbley, “An improved event-independent network for polyphonic sound event localization and detection,” Proc. IEEE ICASSP, 2021.
（8） T.N.T. Nguyen, D.L. Jones, K.N. Watcharasupat, H. Phan, and W.-S. Gan, “SALSA-Lite : A fast and effective feature for polyphonic sound event localization and detection with microphone arrays,” Proc. IEEE ICASSP, 2022.
（9） L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, “First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation,” Proc. DCASE Workshop, 2019.
（10） Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, “A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection,” IEEE/ACM Trans. ASLP, vol. 31, pp. 1251-1264, 2023.
（11） K. Shimada, A. Politis, P. Sudarsanam, D. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, et al., “STARSS23 : An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events,” Proc. NeurIPS, 2023.
（12） Y. Wang, B. Yang, and X. Li, “IPDnet : A universal direct-path ipd estimation network for sound source localization,” IEEE/ACM Trans. ASLP, 2024.
https://ieeexplore.ieee.org/abstract/document/10771699
（13） M. Yasuda, S. Saito, A. Nakayama, and N. Harada, “6DoF SELD : Sound event localization and detection using microphones and motion tracking sensors on self-motioning human,” Proc. IEEE ICASSP, 2024.
（14） K. Shimada, K. Uchida, Y. Koyama, T. Shibuya, S. Takahashi, Y. Mitsufuji, and T. Kawahara, “Zero- and few-shot sound event localization and detection,” Proc. IEEE ICASSP, 2024.

（2024 年 9 月 5 日受付）