機械学習とアレイ信号処理を用いたロボット聴覚に関する研究

The Research on robot audition using machine learning and array signal processing

電子・機械技術部 ロボット・制御科　清野若菜

本研究では、アレイ信号処理を用いた少数チャンネルマイクロホンアレイによる音源定位性能の評価及び、機械学習を用いた環境中の音の「方向」と「対象」を検出するシステムの実装と評価を行った。水平方向１音源を対象とした音源定位では、4ch のマイクロホンアレイでも 8ch と同等の精度であることが分かった。環境音の方向及び対象検出では、既存の学習済みモデルを用いて独自の検出対象を追加学習することで、各現場に適用するためのシステムを比較的容易に構築できることが分かった。

Key words: 機械学習、アレイ信号処理、環境音分析、音源定位

１．緒言

2050 年までに、AI ロボットが社会インフラとなり、工業、介護・福祉、災害現場等の様々な場面で人の暮らしをより豊かにするツールとして広く浸透する状況が想定される[1)]。ロボットが自律的に考え、判断するためには、センサによって環境を計測し、周囲の状況を認識する技術が必要である。ロボットの「目」となる画像認識技術は広く社会実装が進んでいるが、ロボットの「耳」となるロボット聴覚技術は研究開発段階の技術も多く、実用化の例は少ない。そこで本研究は、①小型 AI サービスロボットにも搭載可能な少数チャンネルマイクロホンアレイの試作及び性能評価、②機械学習を用いた環境中の音の「方向」と「対象」を検出するロボット聴覚システムの実装と性能評価を目的とした。

２．実験

２．１．小型音源定位システムの開発

２．１．１．4ch マイクロホンアレイの試作

少数チャンネル入力信号を用いた音源定位の性能を評価するため、4ch のマイクロホンアレイを試作した。試作したマイクロホンアレイを図１に示す。本研究では、水平方向の音源定位性能を評価することとし、半径 36.5[mm]の円周上に 90[°]間隔でマイクロホンを配置する構造とした。マイクロホンはカスケード接続可能な MEMS マイクロホン(システムインフロンティア CSMIC-S1R1）を使用し、多チャンネルオーディオインターフェース（システムインフロンティア RASP-ZX）を PC と接続してアレイ信号処理を行うシステムとした。マイクロホンアレイのフレームは、3D プリンタ（Formlabs Form3L）で作製した。

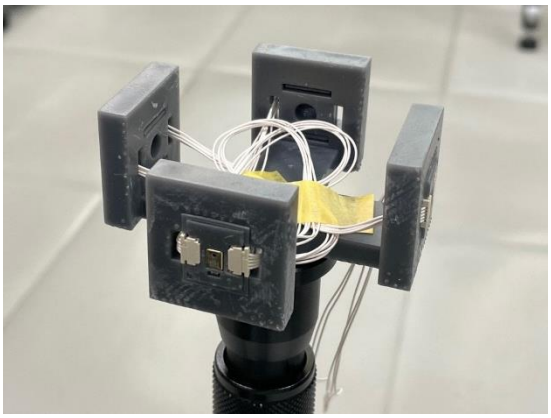

図１　4ch マイクロホンアレイ

２．１．２　音源定位の実装

試作した 4ch マイクロホンアレイ及び市販の 8ch マイクロホンアレイ(システムインフロンティア TAMAGO)を用いて、水平方向の音源定位を実装し、精度を比較した。実装には、ロボット聴覚オープンソースソフトウェア HARK[2)] が提供する MUSIC（multiple signal classification）法によるオンライン音源定位を用いた。MUSIC 法とは、部分空間法による音源定位手法の１つである[3)]。実装条件を表１に示す。マイクロホンアレイから入力される多チャンネル信号をサンプリング周波数 16000[Hz]で取得し、512 ポイントごとに短時間フーリエ変換（STFT）を行うことで、周波数領域に変換し観測ベクトルを生成する。観測ベクトルから空間相関行列を求め、一般化固有値分解により空間相関行列を固有値及び固有ベクトルに分解する。MUSIC 法の空間スペクトルは、得られた固有ベクトル及び、各方向の音源からマイクロホンまでの伝達関数であるアレイ・マニフォールド・ベクトルから、以下の式[3)]によって求めた。なお、アレイ・マニフォールド・ベクトルは、マイクロホンアレイの形状から幾何計算によって 1[°]の分解能で求めた。

事業名「 「そだてる研究室」事業 」

$$P(\theta) = \frac{|\boldsymbol{a}^H(\theta)\boldsymbol{a}(\theta)|}{\sum_{i=N+1}^{M}|\boldsymbol{a}^H(\theta)\boldsymbol{e}_i|^2}$$

$P(\theta)$：空間スペクトル

$\boldsymbol{a}(\theta)$：アレイ・マニフォールド・ベクトル

$\boldsymbol{e}_i$：固有値番号i番目の固有ベクトル

M：マイクロホン数

N：音源数

本実験は当所実験室内で実施した。実験時のマイクロホンアレイ及び音源の配置を図２に示す。マイクロホンアレイを部屋の中央に設置し、マイクロホンアレイから半径 1[m]の円周上 12 か所に音源を順次配置した。音源は直径 70mm の小型スピーカー（Anker Sound Core mini）を用いた。各音源位置で化学プラントの配管から発生するスチーム音を再生し、音源定位した際の空間スペクトルを記録した。

表１　音源定位の実装条件

パラメータ	値
MUSIC アルゴリズム	一般化固有値分解(SEVD)
STFT 点数	512 ポイント
サンプリング周波数	16000[Hz]
フレーム長	32[ms](512 ポイント)
フレームシフト	10[ms](160 ポイント)
使用周波数	500～6400[Hz]
アレイ・マニフォールド・ベクトルの分解能	1[°]
音源配置[°]	0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330

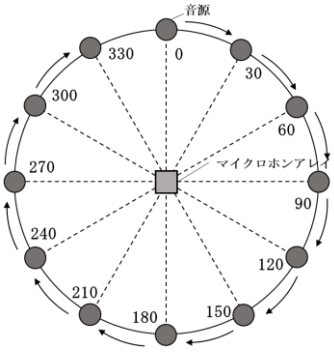

図２　実験配置図

２．２．ロボット聴覚システムの実装

２．２．１．機械学習モデル及びデータセット

音の到来方向と対象を検出するロボット聴覚システムを構築した。学習モデルには、国際的な環境音分析の競技会である DCASE 2023 Challenge Task3「Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes（音響イベントの位置推定と検出)」のベースラインモデル[4) 5) 6) 7)]を用いた。データセットには、同じく DCASE 2023 Task3 で提供されている１次アンビソニックス形式のデータセットを用いた。データセットのクラス分類を表２に示す。クラス数は 13 種類であり、いずれも家庭等から発生する環境音である。データセットは、学習用 90 個、テスト用 78 個の wav 形式の録音データ及びメタデータで構成される。メタデータには、各タイムフレームに検出される環境音のクラス No、音源番号、音源の方位角及び仰角が記録されている。モデルの学習は、学習繰り返し回数を示すエポック数を 200 とした。

表２　データセットのクラス分類

クラス No.	環境音の種類
0	Female speech, woman speaking
1	Male speech, man speaking
2	Clapping
3	Telephone
4	Laughter
5	Domestic sounds
6	Walk, footsteps
7	Door, open or close
8	Music
9	Musical instrument
10	Water tap, faucet
11	Bell
12	Knock

２．２．２．水流音の位置推定及び検出

前節のデータセットに対し、独自の音源を追加してモデルの追加学習を行い、学習に用いていないテスト用データの推論性能を評価した。音源は、配管内の水流音を想定し、本実験では水道の水流音を IC レコーダーで録音した。DAW ソフト及びアンビソニックス制作用 VST プラグインを用いた信号処理により、任意の到来方向情報を持つ１次アンビソニックス形式の wav ファイルに変換し、クラス No. を 10(Water tap, faucet)としてデータセットを作成した。1 データ当たりの信号長は 60[s]とし、学習用データ数を 12、テスト用データ数を 7 とした。追加学習は、前節で学習したベースラインモデルをファインチューニングに用い、エポック数は 50 とした。

３．結果及び考察

３．１．4ch マイクロホンアレイの試作

３．１．１　音源定位の性能比較

各音源配置での 4ch マイクロホンアレイ及び 8ch マイクロホンアレイの音源定位誤差を図３に示す。全 12

か所の平均音源定位誤差は、4ch マイクロホンアレイで 4.4[°]、8ch マイクロホンアレイで 2.8[°]であった。本実験では 1[°]の分解能のアレイ・マニフォールド・ベクトルを用いたが、両マイクロホンアレイともに平均 5[°]以内の誤差で定位できていた。このことから、水平方向に音源数が 1 つの場合の音源定位性能は、4ch マイクロホンアレイでも 8ch と同等であることが分かった。

図 4 に方位角 180[°]の際の両マイクロホンアレイの空間スペクトルを示す。空間スペクトルは、オンライン音源定位により出力された 10 回分のデータを平均し、縦軸を正規化した。実測値から求めた音源方向の固有ベクトルが、正解音源方向のアレイ・マニフォールド・ベクトルと完全に一致する場合、空間スペクトルは本来音源方向にパルス状の鋭いピークを持つ。図 4 の空間スペクトルは、両マイクロホンともに音源方向をピークとしてなだらかな形状となっている。これは、アレイ・マニフォールド・ベクトルをマイクロホン形状から幾何計算により求めたため、部屋の反射や残響の影響により実測値の固有ベクトルとは完全に一致しなかったためと考えられる。また、4ch マイクロホンアレイの空間スペクトルは、8ch に比べて音源方向以外に小さなピークが見られる。これは、マイクロホンアレイの半径を 8ch と同じにし、マイクロホン間隔を広くしたことにより、高周波域で空間折り返しひずみが発生したためと考えられる。

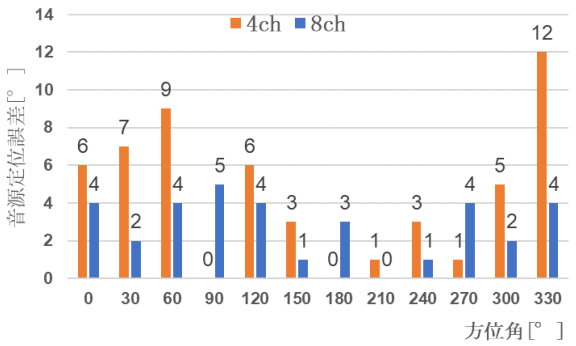

図３　各音源方向の音源定位誤差

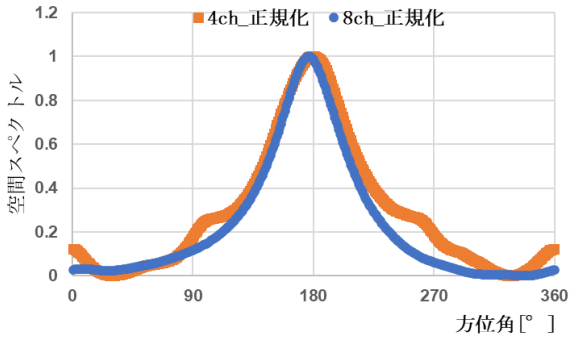

図４　MUSIC 法による空間スペクトル
（音源方向：180[°]）

３．２．ロボット聴覚システムの実装

３．２．１．ベースラインモデルの性能評価

学習したベースラインモデルの性能を表３に示す。Error rate は、対象検出に関する指標であり、1[s]のセグメント間で予測クラスが正解クラスと同じ場合を正解とした誤答率を表す。Localization error は、位置推定に関する指標であり、それぞれ音源方向の定位誤差を表す。いずれの指標も公式ドキュメント[3)]に記載の記録よりも劣る結果となった。これは、公式記録が DCASE 2022[8)]で提供されている外部データセット 1200 個を含めているのに対し、本研究では DCASE 2023 の開発データセットのみを用いているため、学習データ数が少ない事に起因するものと考えられる。

表３　ベースラインモデルの性能

性能評価指標	値　()内は公式記録
Error rate	0.75 (0.57)
Localization error[°]	65.8 (21.6)

３．２．２．水流音の位置推定及び検出

表４に水流音の推論結果、図５に対象検出及び位置推定の可視化結果を示す。図５上段は対象音源のスペクトログラム、下段左列は正解値、右列は予測値を示す。対象検出は各フレームでクラス No.10 の水流音を示しており、正しく検出できていた。位置推定について、方位角は正解値 120[°]に対し平均 1[°]以内の誤差で定位できていた。仰角は、正解値 30[°]に対し、平均 14[°]の定位誤差があった。このことから、既存の学習モデルをファインチューニングに用いて独自のデータを追加学習することで、少ないデータ数とエポック数で対象音源の検出及び位置推定が実現できることが分かった。

表４　水流音の推論結果

性能評価指標	値
Error rate	0.67
Localization error[°]	59.5

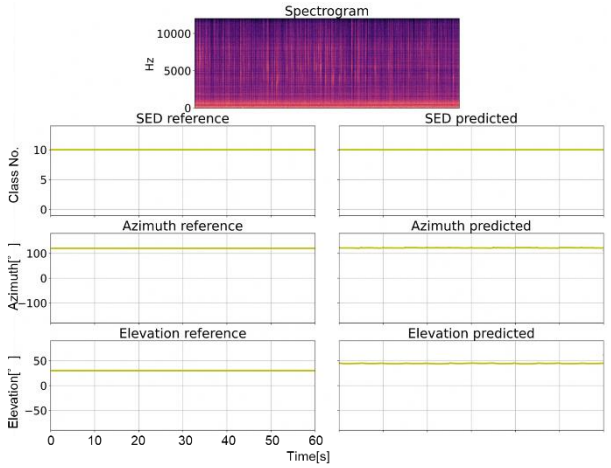

図５　対象検出及び位置推定結果の可視化
（正解値は方位角:120[°]、仰角 30:[°]）

４．結言

本研究では、アレイ信号処理を用いた少数チャンネルマイクロホンアレイによる音源定位性能の評価及び、機械学習を用いた環境中の音の「方向」と「対象」を検出するシステムの実装と評価を行った。水平方向１音源を対象とした音源定位では、4ch のマイクロホンアレイでも 8ch と同等の精度であることが分かった。環境音の方向及び対象検出では、水流音の位置推定及び検出を行い、結果を可視化した。既存の学習済みモデルを用いて独自の検出対象を追加学習することで、各現場に適用するためのシステムを比較的容易に構築できることが分かった。今後は AI ロボットへの搭載に向けて、オンライン推論によるリアルタイム音源位置推定及び対象検出の実現を目指す。

参考文献

１）国立研究開発法人科学技術振興機構．ムーンショット型研究開発事業．ムーンショット目標３．https://www.jst.go.jp/moonshot/program/goal3/index.html（参照 2024-02-21）

２）Kazuhiro Nakadai, Hiroshi G. Okuno, and Takeshi Mizumoto.” Development, Deployment and Applications of Robot Audition Open Source Software HARK”. Journal of Robotics and Mechatronics.2017, vol.29, No.1, p.16-25.

３）浅野太．“部分空間法”.音のアレイ信号処理 -音源の定位・追跡と分離-．コロナ社, 2011, p.107-146.

４）GitHub.“sharathadavanne/seld-dcase2023”. https://github.com/sharathadavanne/seld-dcase2023（参照 2024-2-22).

５）Sharath Adavanne, Archontis Politis, Joonas Nikunen and Tuomas Virtanen. "Sound event localization and detection of overlapping sources using convolutional recurrent neural network". IEEE Journal of Selected Topics in Signal Processing .2019, vo.13, Issue.1, p.34-48.

６）Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Naoya Takahashi, Emiru Tsunoo, and Yuki Mitsufuji, " Multi-ACCDOA: localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training". ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).2022.

７）Parthasaarathy Sudarsanam, Archontis Politis, Konstantinos Drossos."Assessment of Self-Attention on Learned Features For Sound Event Localization and Detection". in proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021.2021, p.100-104.

８）DCASE2022 Challenge.“Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes”. https://dcase.community/challenge2022/task-sound-event-localization-and-detection-evaluated-in-real-spatial-sound-scenes（参照 2024-2-22).