人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-068-03

ニューラル音場推定による 仮想音場でのマイクロフォンアレイ音源定位評価

Source Localization Evaluation of Microphone Arrays in Virtual Sound Fields Estimated by Neural Networks

加藤 雅大 [1,3]　小島 諒介 [2,3] *
Masahiro Kato [1,3]　Ryosuke Kojima [2,3]

[1] 京都大学大学院 工学研究科
[1] Graduate School of Engineering, Kyoto University
[2] 京都大学大学院 医学研究科
[2] Graduate School of Medicine, Kyoto University
[3] 理化学研究所 BDR
[3] RIKEN BDR

Abstract: Achieving sim-to-real transfer in robot audition requires accurately reproducing multichannel sound fields captured by microphone arrays, enabling tasks such as sound source localization to be performed within simulation environments. Recent sound field estimation methods, including Acoustic Volume Rendering (AVR) and Neural Acoustic Fields (NAF), have been proposed; however, most of them focus on single-channel reconstruction, and few studies evaluate sound field estimation in terms of sound source localization performance. In this work, we extend AVR and NAF to multichannel settings and evaluate them using multichannel impulse responses obtained from both real and simulated environments. The results show that the multichannel extension of NAF achieves the lowest localization error, outperforming the other methods.

1　はじめに

近年，ロボット分野においてはシミュレーション環境を活用した強化学習や模倣学習が急速に発展しており，学習した知識を実機へ転移する sim-to-real（sim2real）技術が注目されている．ロボット聴覚における sim2real を実現するためには，ロボットの物理的シミュレーションに加えて，ロボットが置かれる音環境そのものを精密に再現することが不可欠である．特に，ロボットが周囲の音源位置を把握する 音源定位（sound source localization）は，ロボット聴覚における音源追跡 [平塚 24]，行動決定 [平塚 25]，話者識別 [Mošner 24]，環境センシングにおける生態音の観測 [Heath 24] など，多様なタスクの基盤となる重要な機能である [Nakadai 20]．これらの機能は，複数のマイクから得られる信号の時空間構造を解析する マイクロフォンアレイ処理 に依存しており，現実的な音環境を再現できないシミュレーションではその性能が大きく損なわれてしまう．したがって，ロボット聴覚における sim2real の実現には，マイクロフォンアレイ処理や音源定位に対して，シミュレーション段階で現実に近い精緻な音響を再現することが重要となる．

音環境をシミュレーションするための鍵となるのが，任意の位置における音を合成できるインパルス応答（Impulse Response: IR） の利用である．ある音源信号（音声，環境音，音楽など）と IR を畳み込むことで，任意の位置で観測される音響信号を高い精度で再現できる．ニューラルネットワークを用いてインパルス応答データセットから空間全体の音場推定を行う手法として Acoustic Volume Rendering（AVR）[Lan 24] や Neural Acoustic Fields（NAF）[Luo 22] といった手法が注目されている．しかし，これらの手法は単一チャンネルの音響データを対象にしていることが多く，多チャンネルデータに注目した手法は限られている．

そこで，本研究では，実環境及びシミュレーション環境において多チャンネルのインパルス応答データセットを構築し，AVR や NAF を多チャンネルデータに拡張した手法を用いて音場推定を行う．また，各手法で

*連絡先：京都大学
京都府京都市左京区聖護院川原町 54
E-mail:kojima.ryosuke.8e@kyoto-u.ac.jp

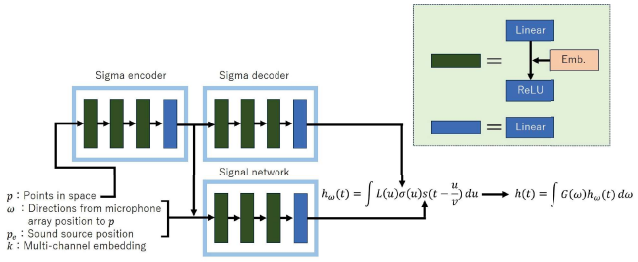


図 1: 多チャンネル AVR アーキテクチャ

生成したインパルス応答を用いて音源定位を行い，その誤差により各音場推定手法の評価を行う．本実験のコードは GitHub[1] より利用可能である．

2　ニューラル音場推定手法

本研究では，ニューラルネットワークを用いた音場推定手法として，音源位置とマイク位置の座標を入力とし，それに対応したインパルス応答を推定する手法である AVR 及び NAF をベースとし，それらを多チャンネル拡張した手法を提案する．

2.1　AVR ベースの手法

Acoustic Volume Rendering（AVR）[Lan 24] は，画像処理における複数の 2 次元画像から 3 次元シーンを再構成する手法の NeRF [Mildenhall 20] における Volume Rendering の考え方を音響信号に拡張し，位相整合性を保ちながらインパルス応答を生成する手法である．AVR では，マイク位置を中心とした球面上に放射される方向 ω を考え，その方向上の空間点で観測される局所的な音響信号をニューラルネットワーク $F : (p, \omega, p_e) \mapsto (\sigma, s(t))$ により表現する．

ここで，p は方向 ω に沿った空間中の点，p_e は音源位置，σ はその点における音響密度，$s(t)$ は t を時刻として，点 p からマイク方向へ伝搬する局所的な音響信号である．

次に，マイク位置から方向 ω に u だけ離れた点 $p(u) = p + u \cdot \omega$ について，密度 σ や伝搬距離に応じた遅延，減衰を考慮しながら $s(t)$ を u_n から u_f の範囲で積分することで，方向 ω からマイクに到達する信号 $h_\omega(t)$ は以下のように表現できる．

$$h_\omega(t) = \frac{1}{t_v} \int_{u_n}^{u_f} L(u)\, \sigma(p(u))\; s\Big(t - \frac{u}{v}; p(u), \omega\Big)\; du,$$

[1] https://github.com/KMASAHIRO/multichannel-soundfields

$$\text{where} \quad L(u) = \exp\Big(- \int_{u_n}^{u} \sigma(p(x))\, dx \Big).$$

ただし，v は音速，t_v は距離によるエネルギー減衰，$L(u)$ は透過率を表す．

最後に，マイクの指向性 $G(\omega)$ を考慮して $h_\omega(t)$ を各方向 ω について積分することで，マイク位置でのインパルス応答 $h(t)$ を得る．

$$h(t) = \int G(\omega)\, h_\omega(t)\, d\omega.$$

本稿では，AVR の多チャンネル拡張を AVR+と呼び，以下のように拡張する．図 1 に AVR+の全体のアーキテクチャを示した．AVR+では，上述のニューラルネット F の代わりに，チャンネル番号 k を入力に含む次のニューラルネット $F' : (p, \omega, p_e, k) \mapsto (\sigma_k, s_k(t))$ を使用する．ここで，チャンネル番号 k は埋め込みベクトルとして，ニューラルネットに入力する．

さらに，追加の拡張をした手法を AVR++と呼び，音源定位に関する誤差項を損失関数に加える．具体的には，微分可能性と計算量を考慮して比較的単純な delay-and-sum beamforming を用いた損失関数を定義する．ここで，マイクロフォンアレイ中心から k 番目のマイク位置へのベクトルを $\mathbf{p}_k$ として，方向 $\mathbf{u}(\theta_\ell)$ に対するパワースペクトル P_ℓ は，次のように表される．

$$P_\ell = \sum_f \left(|A_{\ell,f}|^2 \Big/ \sum_{\ell'=0}^{L-1} |A_{\ell',f}|^2 \right),$$

$$A_{\ell,f} = \frac{1}{K} \sum_{k=1}^{K} X_{k,f}\, \exp\left(-j2\pi f\, \frac{\mathbf{p}_k^{\mathsf{T}} \mathbf{u}(\theta_\ell)}{v} \right),$$

$$\mathbf{u}(\theta_\ell) = \begin{bmatrix} \cos\theta_\ell \\ \sin\theta_\ell \end{bmatrix}, \qquad \theta_\ell = \frac{2\pi\ell}{L}.$$

$X_{k,f}$ は k 番目のマイクで観測された信号をフーリエ変換により周波数領域へ変換したときの，周波数ビン f の複素スペクトルである．計算したパワースペクトル P_ℓ を用いて，微分可能となるように soft-argmax を用いて音源定位方向 $\hat{\theta}$ を計算する．

$$\hat{\theta} = \sum_{\ell=0}^{L-1} w_\ell\, \theta_\ell, \qquad w_\ell = \frac{\exp(\beta P_\ell)}{\sum_{\ell'=0}^{L-1} \exp(\beta P_{\ell'})}.$$

正解波形及び AVR 推定波形による音源定位方向をそれぞれ $\hat{\theta}_{\text{true}}$，$\hat{\theta}_{\text{pred}}$ として，次のように計算した誤差 L_{DS} を従来の AVR の損失関数に加えて，ニューラルネットワークの学習を行う．

$$\mathcal{L}_{\text{DS}} = \left| \sin\hat{\theta}_{\text{pred}} - \sin\hat{\theta}_{\text{true}} \right| + \left| \cos\hat{\theta}_{\text{pred}} - \cos\hat{\theta}_{\text{true}} \right|.$$

ここでは，音源候補方向の分解能 L は 360 とした．

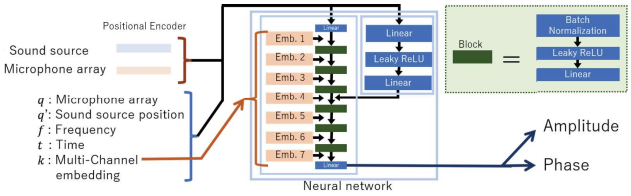


図 2: 多チャンネル NAF アーキテクチャ

2.2 NAF ベースの手法

Neural Acoustic Fields（NAF）[Luo 22] は，時間-周波数領域のインパルス応答を予測対象として，マイク位置 q ，音源位置 q' ，時刻 t ，周波数 f に対応するスペクトルの振幅 v_{mag} 及び位相 v_{IF} を，ニューラルネットワーク $\Omega : (q, q', t, f) \mapsto \big(v_{\mathrm{mag}}(t, f), v_{\mathrm{IF}}(t, f)\big)$ により推定する．

本稿では，NAF の多チャンネルデータ拡張を NAF+ と呼び，ニューラルネットワーク Ω の代わりに，チャンネル番号 k の埋め込みベクトルを利用するニューラルネットワーク $\Omega' : (q, q', t, f, k) \mapsto \big(v_{\mathrm{mag},k}(t, f), v_{\mathrm{IF},k}(t, f)\big)$ を使用する．

NAF+では従来の 2 チャンネルに対する NAF のアーキテクチャをベースに，図 2 に示すアーキテクチャを構成した．ここで，各チャンネルの埋め込みベクトルは中間層の次元と同じ次元数を持つ埋め込みを設定して，各層の間に足し合わせる構造となっている．

3 データセット

ここでは，実環境での測定により構築したデータセット，AcoustiX[Lan 24] 及び Pyroomacoustics[Scheibler 18] によるシミュレーションで構築したデータセットの計 3 つの評価データセットについて述べる．

3.1 実データセット

実環境における TSP (Time Stretched Pulse) 信号を用いた計測によって，実データの多チャンネルインパルス応答データセットを構築した [加藤 24]．データセットは，同一環境でのマイク位置，音源位置，多チャンネルインパルス応答をデータの一組として，合計 184 個の組から構成される．

次に，具体的な測定条件を示す．図 3 および図 4 に示す環境で，部屋の寸法は 6.110m（幅方向）×8.807m（奥行方向）であった．スピーカーと円形マイクロフォンアレイを 1 つずつ用意し，幅方向に 1m 間隔で 4 箇所，奥行方向に 1m 間隔で 6 箇所の計 $4 \times 6 = 24$ 箇所をマイクロフォンアレイの配置候補箇所とし，角の 4 箇所及び中央の 4 箇所の計 8 箇所をスピーカーの配置候補箇所とした．スピーカー及びマイクロフォンアレイの高さは 1.5m に固定した．円形マイクロフォンアレイにはチャンネル数は 8，半径は 0.0365m の TAMAGO (SYSTEM IN FRONTIER Inc.) を使用した．

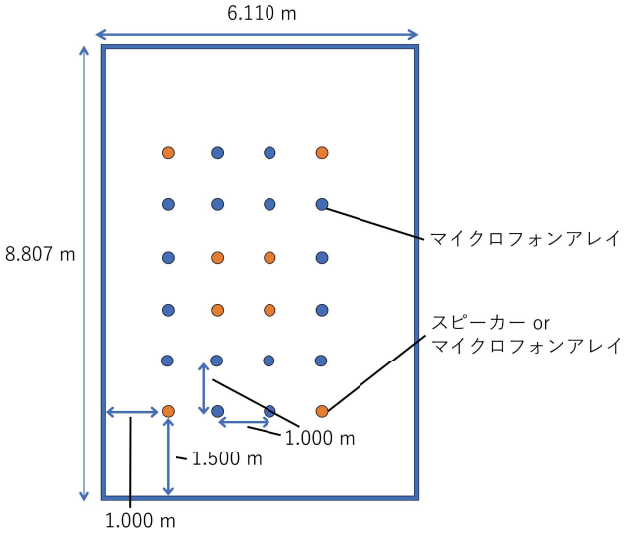


図 3: データ測定環境

図 4: 実データ測定時の部屋の写真

計測時のサンプリング周波数は 16kHz とし，TSP 信号の再生及び記録には Audacity[2]を用いた．スピーカー及び円形マイクロフォンアレイの位置を変えて測定を繰り返し，計 $8 \times 23 = 184$ 回の測定を行った．1 回の測定に付き，スピーカーから TSP 信号を連続で 10 回再生させ，円形マイクロフォンアレイで記録を行った．

測定終了後，記録した 10 個の TSP 信号をそれぞれ ITSP 信号と畳み込んでインパルス応答を生成し，10 個のインパルス応答の平均を取った．その後，測定間の記録開始タイミングのずれを調整するため，1 番目チャンネルのインパルス応答における振幅のピーク時刻が全データで一致するように加工した．また，振幅

[2]https://www.audacityteam.org/

のピーク値の全データ間での最大値により各インパルス応答を正規化した．さらに，AVR 提案論文の状況に合わせて，各インパルス応答の長さを 0.1s に切り詰めた．このようにして，計 184 個の多チャンネル室内インパルス応答からなる実データのデータセットを得た．

実データセットのクオリティチェックとして，幾何計算から求めた伝達関数による周波数正規化 MUSIC 法 [Salvati 14] によって，収録信号を用いた音源定位を行ったところ，実際の音源方向から $6.50° \pm 4.99°$ の絶対誤差があった．

3.2 シミュレーション 1 (AcoustiX)

AcoustiX[Lan 24] は，レイトレーシングを用いた音響シミュレータである．電波用のレイトレーシングエンジンである Sionna ray tracing（Sionna RT）[Hoydis 22] を音響向けに拡張したシミュレータであり，Sionna RT が計算した各レイの反射経路に対して遅延時間や減衰を考慮することで，位相や到達時刻の正確性に注意したインパルス応答の計算が可能である．

AcoustiX を用いて，3.1 節と同様の設定（部屋寸法，音源及びマイクロフォンアレイ配置，マイクロフォンアレイ形状）でシミュレーションを行った．シミュレーションの設定値はデフォルトの値を用い，レイの数は 50000，反響回数 10 回まで追跡し，サンプリング周波数は 16kHz，0.1s のインパルス応答を計算した．このようにして，計 184 個の多チャンネル室内インパルス応答からなるシミュレーションデータセット 1 を得た．

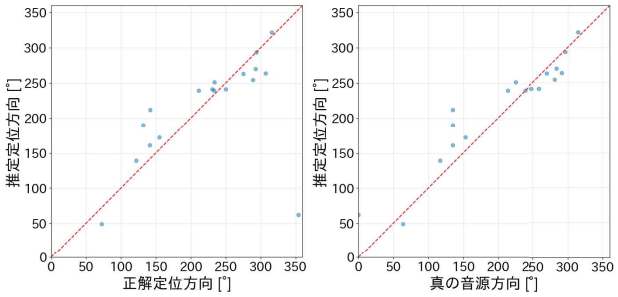
クオリティチェックとして，周波数正規化 MUSIC 法を用いてシミュレーション 1 の信号から音源定位を行ったところ $3.11° \pm 3.67°$ の絶対誤差があった．

3.3 シミュレーション 2 (Pyroomacoustics)

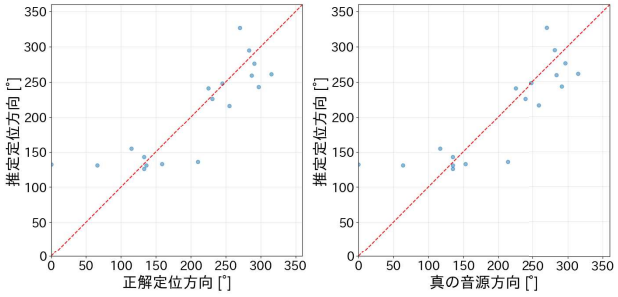
Pyroomacoustics[Scheibler 18] は広く利用される音響シミュレータであり，鏡像法を用いて効率的にインパルス応答の計算が可能である．

Pyroomacoustics を用いて，3.1 節及び 3.2 節と同様の設定（同一部屋寸法，音源及びマイクロフォンアレイ配置，マイクロフォンアレイ形状）でシミュレーションを行った．反響回数は 10 回まで追跡し，サンプリング周波数は 16kHz，0.1s のインパルス応答を計算した．このようにして，計 184 個の多チャンネル室内インパルス応答からなるシミュレーションデータセット 2 を得た．

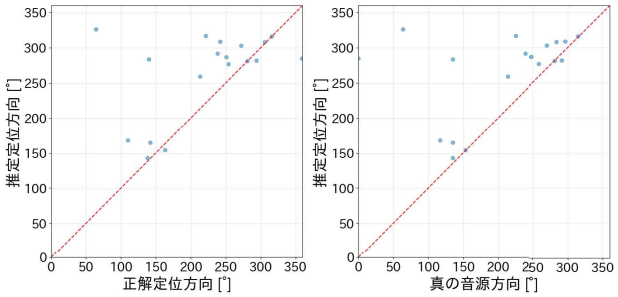
クオリティチェックとして，周波数正規化 MUSIC 法を用いてこのシミュレーションデータセット 2 の信号から音源定位を行ったところ $5.83° \pm 6.46°$ の絶対誤差があった．



(a) 実データ



(b) シミュレーション 1



(c) シミュレーション 2

図 5: 実験 1 の結果，左図：DoA-A 誤差に対応（横軸：正解定位方向，縦軸：推定定位方向），右側：DoA-B 誤差に対応（横軸：真の音源方向，縦軸：推定定位方向）

4 実験

4.1 実験 1: ニューラル音場推定手法の比較

2 節で提案した各手法の比較をするため，3 節で構築した各データセットに対し，90%を訓練データ，10%を検証データとして分割し，推定インパルス応答を入力として周波数正規化 MUSIC 法を用いた音源定位誤差による各手法の評価を行った．各手法は Optuna を用いて，それぞれ表 3 と表 2 の範囲でハイパーパラメータチューニングを行った．

評価指標として，推定した信号から音源定位により求められた音源方向（予測定位方向）と，正解の信号から音源定位により推定された音源方向（正解定位方向）の絶対誤差を DoA-A 誤差として導入した．また，推定した信号から音源定位により求められた音源方向

表 1: 各手法の IR 評価指標および音源定位誤差

(a) 実データ

モデル	Phase	Amp.	Env. [%]	T60 [%]	C50 [dB]	EDT [ms]	DoA-A [°]	DoA-B [°]
AVR	1.60±0.81	18.21±25.29	115.23±134.31	28.47±25.36	16.85±11.01	11.42±7.80	31.75±28.89	32.35±25.06
AVR+	1.58±0.82	0.77±1.39	7.51±11.83	26.63±24.99	12.81±4.75	11.89±8.01	36.31±26.81	36.86±26.01
AVR++	1.61±0.80	1.29±2.49	14.26±16.48	43.65±34.78	21.90±12.32	18.63±13.63	37.75±22.59	32.83±23.12
NAF	1.62±0.80	0.98±0.02	9.39±14.69	26.56±15.63	14.34±6.71	65.76±14.99	33.06±26.65	28.99±21.16
NAF+	1.62±0.81	0.80±0.16	8.04±13.09	30.46±17.71	13.35±6.99	31.69±14.15	**20.25±14.54**	**18.65±12.98**

(b) シミュレーション 1

モデル	Phase	Amp.	Env. [%]	T60 [%]	C50 [dB]	EDT [ms]	DoA-A [°]	DoA-B [°]
AVR	1.61±0.81	0.46±0.29	8.81±14.71	22.09±10.74	5.88±2.86	54.18±37.60	43.38±32.56	42.24±33.52
AVR+	1.62±0.80	1.61±1.90	11.22±15.60	44.83±9.73	16.55±3.64	202.46±61.66	30.13±13.43	29.14±13.77
AVR++	1.62±0.80	0.52±0.57	9.44±12.84	13.01±9.62	2.30±1.63	52.91±38.97	34.94±19.89	34.81±19.44
NAF	1.60±0.80	0.56±0.09	5.15±8.64	14.21±9.00	1.39±0.93	44.90±32.93	29.69±27.52	28.98±26.75
NAF+	1.62±0.81	0.58±0.10	4.59±8.13	9.42±7.21	1.49±1.11	36.78±29.92	**27.25±20.67**	**27.40±20.37**

(c) シミュレーション 2

モデル	Phase	Amp.	Env. [%]	T60 [%]	C50 [dB]	EDT [ms]	DoA-A [°]	DoA-B [°]
AVR	1.58±0.81	0.29±0.14	9.34±10.39	14.18±8.30	3.00±1.82	19.77±13.99	43.06±26.09	48.09±26.13
AVR+	1.52±0.82	0.46±0.11	8.30±9.79	15.54±6.36	2.92±1.60	20.15±11.02	32.31±21.46	**29.88±20.42**
AVR++	1.62±0.80	0.64±0.18	9.92±11.79	16.87±8.07	8.01±2.61	25.51±14.83	32.63±25.03	30.26±25.27
NAF	1.62±0.81	0.85±0.07	8.14±11.05	31.60±10.14	8.13±1.42	36.62±21.75	**30.19±17.09**	31.83±17.85
NAF+	1.62±0.81	0.61±0.08	8.28±11.34	18.00±11.94	1.26±0.95	26.97±17.51	33.50±28.71	30.81±26.37

表 2: ハイパーパラメータ（NAF, NAF+）

項目	範囲
層数	4 〜 12
残差層の層数	0 〜 3
中間層の特徴量数	64 〜 1024
座標に加えるノイズ量	0.0 〜 0.2
初期学習率	1×10^{-5} 〜 1×10^{-2}
学習率減衰率	0.05 〜 0.5
位相損失の重み	0.1 〜 10
1 step のサンプリング点数（時刻×周波数）	200 〜 4000

表 3: ハイパーパラメータ（AVR, AVR+，AVR++）

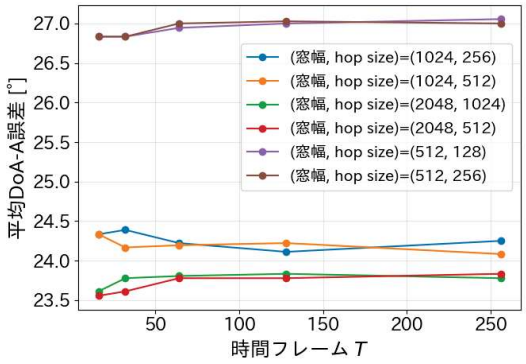
項目	範囲
初期学習率	1×10^{-6} 〜 1×10^{-4}
最小学習率	初期学習率の 1% 〜 50%
半径方向のサンプル数	40 〜 80
水平角方向の分割数	48 〜 80
重み減衰	0 〜 0.001
スペクトル誤差の重み	0 〜 100
角度誤差の重み	0 〜 100
時間誤差の重み	0 〜 100
エネルギー誤差の重み	0 〜 100
Multi-STFT 誤差の重み	0 〜 100
DAS 誤差の重み	0 〜 100
σ エンコーダの幅	32 〜 512
σ デコーダの幅	32 〜 512
signal ネットワークの幅	128 〜 1024
σ エンコーダに埋め込み	True / False
σ デコーダに埋め込み	True / False
signal ネットワークに埋め込み	True / False

（予測定位方向）と，実際の音源方向（真の音源方向）の絶対誤差を DoA-B 誤差と定めた．

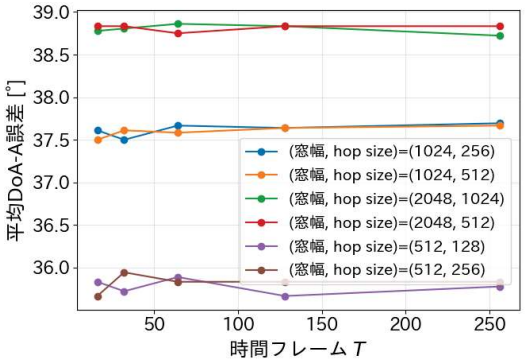
ニューラルネットではモデルを教師信号に近づけるよう学習するため，検証の評価指標としては DoA-A 誤差の平均を最小化するようにチューニングを行った．音源定位の手法には MUSIC 法を用い，STFT により得られた時間-周波数領域のインパルス応答を利用して音源定位を行った．STFT のパラメータは窓幅を 512，hop size を 128，窓関数をハニング窓とした．各試行において 50 epochs 学習を行い，各 epoch 終了時に検証データに対する DoA-A 誤差の計算を行って，その最小値を記録した．

2 節で提案した AVR, AVR+, AVR++, NAF, NAF+ に対して 100 回の試行を行った際の最良の結果を表 1a, 表 1b, 表 1c に示す．表 1a が実データセット，表 1b がシミュレーション 1，表 1c がシミュレーション 2 に対する結果を表す．

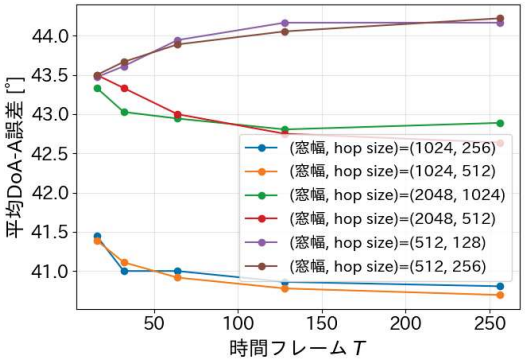
DoA-A 誤差及び DoA-B 誤差は外れ値による分散への影響が大きかったため，誤差最大値の 90 パーセンタイルを基準として外れ値除去を行った結果となっている．さらに，IR 評価指標として，位相 Phase，振幅 Amp.，包絡線 Env.，残響時間 T60，明瞭度 C50，初期残響時間 EDT に対する誤差を表に示している．表 1a，表 1b より，実データ及びシミュレーションデータ



(a) 実データ



(b) シミュレーション 1



(c) シミュレーション 2

図 6: ホワイトノイズ音源での音源定位誤差 DoA-A

1 に対しては NAF+の音源定位誤差が最小となっていることがわかる．表 1c では，DoA-A で NAF，DoA-B で AVR+が最小の誤差を示しているが，NAF+も小さい誤差となっていることがわかる．これらの結果から，音源定位誤差で比較すると，提案手法の中で NAF+が最良のモデルであるといえる．

次に，検証データに対して NAF+が推定した各インパルス応答について，音源定位による評価を行った結果を図 5a，図 5b，図 5c に示す．図 5a が実データセットについての結果，図 5b がシミュレーション 1 についての結果，図 5c がシミュレーション 2 についての結果を表す．左側の図は横軸が正解インパルス応答による音源定位方向，縦軸が推定インパルス応答による音源定位方向を表す．すなわち，誤差 DoA-A を表す．右側の図は横軸が実際の音源方向，縦軸が推定インパルス応答による音源定位方向を表す．すなわち，誤差 DoA-B を表す．いずれの図も，斜めに引かれた赤の破線上に点が近いほど誤差が小さいことを表す．なお，縦軸横軸ともに範囲が 0° から 360° であり，実際には上下と左右の端は連続していることに注意する．いずれのグラフでも，数点の外れ値を除いて，破線上に近い位置に音源定位結果が分布していることがわかる．

4.2 実験 2: ホワイトノイズ音源での音源定位評価

現実には，4.1 節で行ったようなインパルス信号で音源定位をする状況は非常に限られている．そこで，実用的な長さの音響信号で音源定位をするため，音源を 100s のホワイトノイズとしたときの音源定位評価を行った．具体的には，100s のホワイトノイズをインパルス応答と畳み込み，STFT 後に時間フレーム T ごとに区切って音源定位を行った．各時間フレームでの音源定位方向の中央値を全体での音源定位方向とし，DoA-A 誤差，DoA-B 誤差を評価した．

検証データに対する NAF+の推定インパルス応答を用いて，音源をホワイトノイズとしたときの音源定位評価を行った．まず，最適なハイパーパラメータを決めるために，時間フレーム T を 16，32，64，128，256 の 5 パターン，さらに，STFT のパラメータ（窓幅，hop size）を（512，128），(512，256），(1024，256），(1024，512），(2048，512），(2048，1024）の 6 パターン使用して，最適なパラメータの検討を行った．なお，STFT の窓関数はハニング窓を使用した．各ハイパーパラメータでの DoA-A 誤差の結果を図 6a，図 6b，図 6c にまと

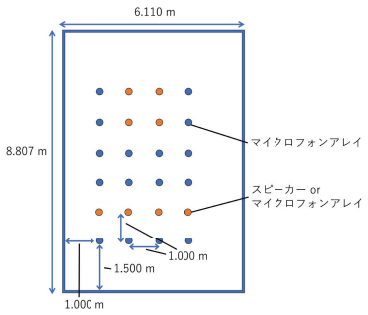


図 7: 音源位置 (図 3 より，音源位置（橙点）を変更)

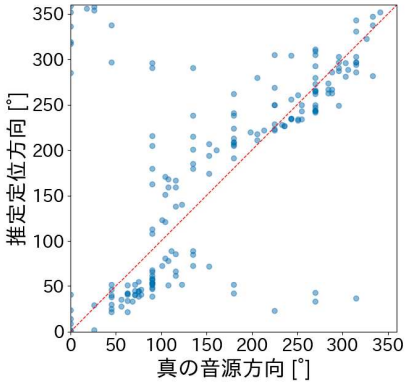


図 8: インパルス音源

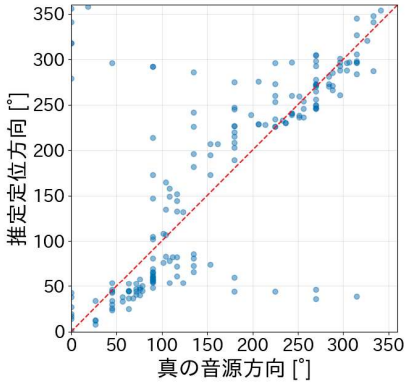


図 9: ホワイトノイズ音源

表 4: 音源による誤差の違い

(a) DoA-A 誤差

音源	実データ	シミュレーション 1	シミュレーション 2
インパルス	20.25 ° ± 14.54 °	27.25 ° ± 20.67 °	33.50 ° ± 28.71 °
ホワイトノイズ	18.50 ° ± 16.20 °	27.72 ° ± 20.06 °	30.34 ° ± 26.73 °

(b) DoA-B 誤差

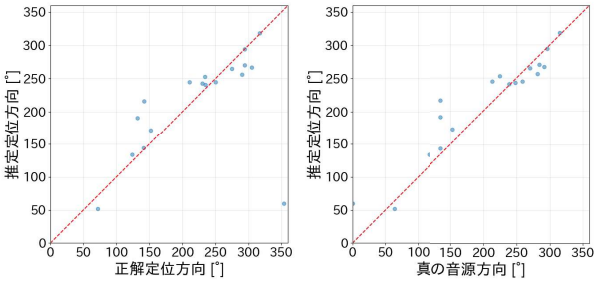
音源	実データ	シミュレーション 1	シミュレーション 2
インパルス	18.65 ° ± 12.98 °	27.40 ° ± 20.37 °	30.81 ° ± 26.37 °
ホワイトノイズ	16.71 ° ± 14.16 °	27.77 ° ± 19.65 °	28.60 ° ± 26.04 °

めた．STFT パラメータについては，いずれのデータセットでも窓幅 = 1024，hop size= 512 において平均的に誤差が小さくなっていることがわかる．一方，時間フレーム T の変化に対しては，あまり誤差が変化していないことがわかる．

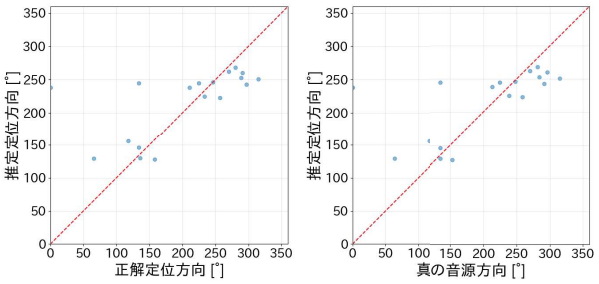
次に，時間フレーム $T = 64$，STFT パラメータ (窓幅=1024，hop size=512) としてホワイトノイズ音源での音源定位評価を行った．その結果を図 10a，図 10b，図 10c に示す．左側の図が誤差 DoA-A，右側の図が誤差 DoA-B を表す．いずれの図も，斜めに引かれた赤の破線上に点が近いほど誤差が小さいことを表す．いずれのグラフでも，数点の外れ値を除いて，破線上に近い位置に音源定位結果が分布していることがわかる．また，音源がインパルス信号の場合の音源定位誤差と比較した結果を表 4a，表 4b に示す．なお，表の値は誤差最大値の 90 パーセンタイルを基準として外れ値除去を行った結果となっている．インパルス音源のときと比較してホワイトノイズ音源では誤差がやや小さくなっていることがわかる．

4.3 実験 3: 未知の音源位置に対する音源定位評価

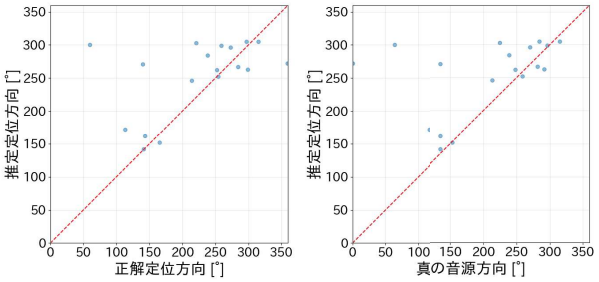
4.1 節，4.2 節において DoA-A 及び DoA-B 誤差が最小だった，実データセットで学習した NAF+について、同一環境の未知音源位置に対する音源定位評価を行った．学習データに含まれる音源位置と重ならないように，図 7 における橙色の 8 箇所をスピーカーの配置候補箇所とし，スピーカー位置を除く格子上の 23 点をマイクロフォンアレイ配置候補場所とした．なお，格子点の配置は学習データの配置図 3 と同じものを使用した．合計で $8 \times 23 = 184$ 個の音源，マイク位置ペアを作成し，学習済みの NAF+でインパルス応答の推定を行った．本実験では，正解の信号がなく DoA-A 誤差は計算できないため，インパルス音源及びホワイトノイズ音源での DoA-B 誤差の結果を，それぞれ図 8，図 9 に示す．90 パーセンタイル基準で外れ値除去すると，インパルス音源での DoA-B が 26.23° ± 17.33°，ホワイトノイズ音源では 24.00° ± 16.68° であった．このことから，未知の音源位置に対しても，検証データと同等の音源定位誤差となることが確認できた．



(a) 実データ



(b) シミュレーション 1



(c) シミュレーション 2

図 10: ホワイトノイズ信号での音源定位分布

4.4 おわりに

本研究では，多チャンネル音場推定手法の評価指標として，マイクロフォンアレイによる音源定位誤差に着目した検討を行った．実環境で計測したインパルス応答データセットに加え，AcoustiX および Pyroomacoustics により構築したシミュレーションデータセットを用いて，AVR，NAF 及びそれらを多チャンネルデータに拡張した AVR+，AVR++，NAF+の比較を行った．その結果，提案したチャンネル埋め込みを導入した NAF+が最も小さい音源定位誤差を示すことが確認された．また，ホワイトノイズを音源とする実用的な長さの音響信号を用いて音源定位評価を行ったところ，インパルス音源の場合と比較して誤差がやや小さくなることを確認した．さらに，実データセットで学習した NAF+を用いて未知の音源位置に対する音源定位評価を行い，検証データと同程度の誤差で音源方向推定が可能であることを確認した．

今後の課題として，より多いデータ数や多様な音響環境で提案手法を学習させることで，推定信号からの音源定位誤差をさらに小さくすることなどが考えられる．

Acknowledgements

本研究は CREST JPMJCR22D3 および RIKEN TRIP initiative (AGIS) の助成を受けた．

参考文献

[Heath 24] Heath, B. E., Suzuki, R., Le Penru, N. P., Skinner, J., Orme, C. D. L., Ewers, R. M., Sethi, S. S., and Picinali, L.: Spatial Ecosystem Monitoring with a Multichannel Acoustic Autonomous Recording Unit (MAARU), *Methods in Ecology and Evolution*, Vol. 15, No. 9, pp. 1568–1579 (2024)

[Hoydis 22] Hoydis, J., Cammerer, S., Ait Aoudia, F., Nimier-David, M., Maggi, L., Marcus, G., Vem, A., and Keller, A.: Sionna (2022), https://nvlabs.github.io/sionna/

[Lan 24] Lan, Z., Zheng, C., Zheng, Z., and Zhao, M.: Acoustic Volume Rendering for Neural Impulse Response Fields, *arXiv preprint arXiv:2411.06307* (2024)

[Luo 22] Luo, A., Du, Y., Tarr, M., Tenenbaum, J., Torralba, A., and Gan, C.: Learning neural acoustic fields, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 3165–3177 (2022)

[Mildenhall 20] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, in *ECCV* (2020)

[Mošner 24] Mošner, L., Serizel, R., Burget, L., Plchot, O., Vincent, E., Peng, J., and Černocký, J.: Multi-Channel Extension of Pre-trained Models for Speaker Verification, in *Interspeech 2024* (2024)

[Nakadai 20] Nakadai, K. and Okuno, H. G.: Robot Audition and Computational Auditory Scene Analysis, *Advanced Intelligent Systems*, Vol. 2, No. 9, p. 2000050 (2020)

[Salvati 14] Salvati, D., Drioli, C., and Foresti, G. L.: Incoherent Frequency Fusion for Broadband Steered Response Power Algorithms in Noisy Environments, in *IEEE Signal Process. Lett.*, Vol. 21, pp. 581–585 (2014)

[Scheibler 18] Scheibler, R., Bezzam, E., and Dokmanić, I.: Pyroomacoustics: A Python package for audio room simulations and array processing algorithms, in *IEEE ICASSP* (2018)

[加藤 24] 加藤 雅大, 小島 諒介：多チャンネル音響生成を評価するためのマイクロフォンアレイ室内インパルス応答データセットの構築, 人工知能学会 第 128 回人工知能基本問題研究会（SIG-FPAI） (2024)

[平塚 24] 平塚 謙良, 小島 諒介：世界モデルベースの深層強化学習による音源追跡の検討, 第 66 回人工知能学会 AI チャレンジ研究会 (2024)

[平塚 25] 平塚 謙良, 小島 諒介：音環境情報に基づくピックアンドプレースの模倣学習, RSJ2025 (2025)