ISSN 2186-5647

9-5

強化学習を用いた集束超音波の焦点制御に関する基礎検討

日大生産工(院)　○佐藤 龍成　　日大生産工　沖田 浩平

1. 緒言

近年，アレイトランスデューサをデバイスとして超音波を体内に集束させて治療する集束超音波(HIFU：High intensity focused ultrasound)が注目され，メス等で身体に傷をつけることなく最小限かつ高効率な治療が可能となってきている．しかしながら，人体の部位による超音波の屈折が発生し焦点を定めることが困難であることや，焦点に発生したキャビテーション気泡によって音場が変化し，集束超音波焦点がターゲットである病巣から外れる問題がある．これらの問題は，複数の超音波振動素子からなるフェーズドアレイトランスデューサの位相遅延量の制御によって解決可能だが，制御法は確立されていない．

人工知能の発展により，その技術は自動運転をはじめとする様々なロボットに搭載されるようになっている．人工知能の代表例として，強化学習が挙げられる．強化学習を用いた事例として，様々な地震波に対する建物の挙動をシミュレーションで再現し，その制御方法を深層強化学習で用いて学習させることで優れた制振効果が得られている[1]．このように，強化学習は現実を模したシミュレーションとの相性がよく，与えられた環境で効率的に試行錯誤することで，最適な行動を選択することが可能である．

本研究では，フェーズドアレイトランスデューサの最適な制御方法への強化学習の適用に向けて，音響不均質媒体中における強化学習による焦点制御の基礎検討として，簡略化したモデルでその有効性について調べた．

2. 強化学習

2.1 強化学習の概要

強化学習とは，与えられた環境（状態）において学習者と呼ばれるエージェントが観測，行動し，目的とした価値（報酬）を最大化するまで試行錯誤し続ける学習方法である．Fig.1に強化学習の概念図を示す．時点tにおいて，エージェントが環境から状態S_tを観測し，与えられた方策をもとに行動A_tを出力する．次に，エージェントの行動によって変動した環境は状態S_{t+1}に更新され，それに応じた報酬R_tをエージェントが得る．エージェントは報酬R_tに基づいて方策の改善を繰り返す．その結果，累計して得られる累積報酬を最大化する方策は次式で表される．

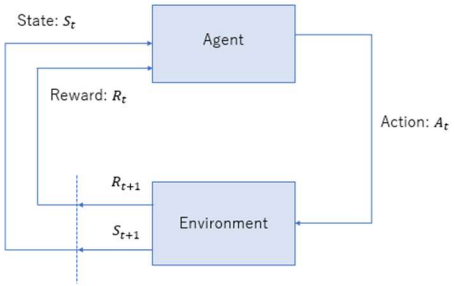

Fig.1　強化学習の概念図

$$\pi^* = argmax_{\pi \in \Pi} \mathrm{J} \qquad (1)$$

ここで，π^*は最適方策，Jは期待値，Πは方策集合である．環境の状態，行動の組み合わせにおける期待値Jの合計が最大となる行動をとる方策を方策集合Πから見つけることで最適方策π^*が表される[2]．

2.2 強化学習のアルゴリズム

強化学習の目的である最適方策を得るためのアルゴリズムには，エージェントの行動によって離散行動空間で行うものと連続行動空間で行うものの２つに分けられる．離散行動空間では，エージェントはあらかじめ定められた行動の中から選択するため，アルゴリズム設計が簡単になるメリットがあるが，選択できる行動が限られてしまう．一方，連続行動空間では，エージェントは任意の数値を行動として選択することができるため，行動の選択肢が豊富で複雑な環境に対応できるメリットがある．代表的なアルゴリズムとして，離散行動空間に対してはQ-Learning[3]とSARSA[3]が，連続行動空間にはDDPG[3]とPPO[4]が挙げられる．

3. 簡略化モデルに対する強化学習

Fig.2にフェーズドアレイトランスデューサによる集束超音波の簡略化モデルを示す．トランスデューサ素子に対応する音源A，Bより

Focus Control of Focused Ultrasound by Reinforcement Learning

Ryusei SATO, Kohei OKITA

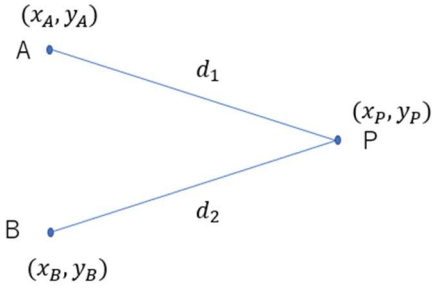

Fig.2　集束超音波の簡略化モデル

発した超音波が一様な媒質中を伝播して焦点Pに到達する状況を考える．ここで，音源Aから発せられる音は，周波数ωを用いて，

$$\sin(\omega(T + \Delta T_A)) \qquad (2)$$

と表される．ここで，ΔT_Aは位相遅延量である．音源から焦点Pまでの距離d_A進むのにかかる時間を$T_A{}'$とすると，点Pにおける音圧は音源Bから発せられ音の影響も考慮して，

$$\sin\left(\omega(T + \Delta T_A - T_A')\right) + \sin(\omega(T + \Delta T_B - T_B')) \qquad (3)$$

と表される．簡略化モデルでは，この焦点Pにおける音圧値を最大化する位相遅延量ΔT_AおよびΔT_Bを強化学習によって求めた．

焦点音圧を強化学習における状態として，焦点音圧を最大化するための位相遅延量を行動とし，音圧値の増減によって正負の報酬を与えた．本研究では，位相遅延量を$[-\pi \sim \pi]$の範囲で連続値として制御するため，連続行動空間に対応した強化学習アルゴリズムであるPPO[4](Proximal Policy Optimization)を用いた．これにより，方策を更新する際の大幅な変更を回避し，安定した学習を実現した．

4.　学習結果と考察

Fig.3に学習過程における報酬の履歴を示す．報酬が正負に変動しており，報酬が大きく増加するエピソードもあるものの，収束には至っていないことがわかる．次に，Fig.4に学習回数50回ごとの累積報酬を示す．こちらについても，累積報酬の値がほとんど場合で負となっており，累積報酬を最大化する最適方策が見つけられていないと考えられる．原因として，連続行動空間の学習に多くの学習回数が必要であること，PPOアルゴリズムの初期方策への依存性が高いことから初期方策が誤っていたことなどが挙げられる．

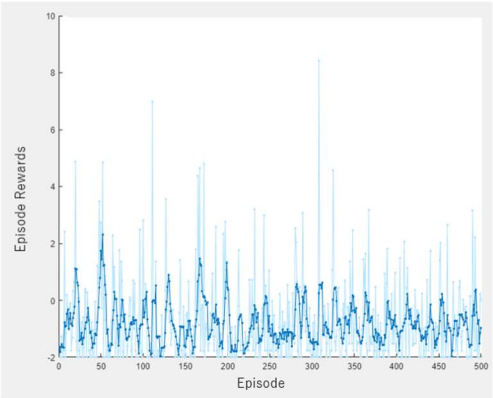

Fig. 3　学習過程における報酬の履歴

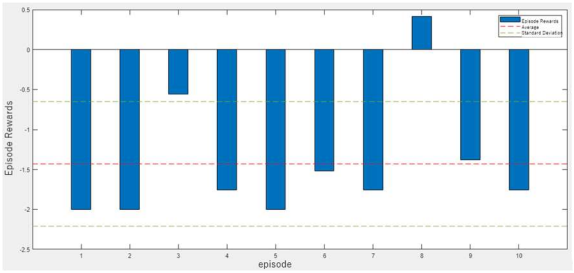

Fig.4　累積報酬の履歴

5.　結言

集束超音波治療におけるフェーズドアレイトランスデューサの最適な制御方法への強化学習の適用に向けて，簡略化モデルでその有効性について調べた．その結果，学習が十分に収束せず，累積報酬を最大化する方策が得られていない．今後は，連続行動として扱っている位相遅延量を離散行動として扱うことで，学習を高速化し，効率的に最適方策が得られる方法などについて検討する．

参考文献

1)　https://www.msi.co.jp/technology/machinelearing.html(参照 2023-09-25)
2)　R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction 2nd ed (2018).
3)　斎藤康毅，ゼロから作るDeep Learning4-強化学習偏，株式会社オライリー・ジャパン，(2022)
4)　John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv: 1707.06347 (2017)