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-063-08

フーリエ級数展開を用いた軽量伝達関数のオンライン適応による音源定位・分離の向上

Online Adaptation of Fourier series based Lightweight Transfer Function to Improve Sound Source Localization and Separation

周藤 唯[1*]　瀧ケ平 将行[1]　中臺 一博[2]　中島 弘史[3]
Yui Sudo[1]　Masayuki Takigahira[1]　Kazuhiro Nakadai[2]　Hirofumi Nakajima[3]

[1] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[1] Honda Research Institute Japan Co., Ltd.
[2] 東京工業大学　[3] 工学院大学
[2] Tokyo Institute of Technology　[3] Kogakuin University

Abstract: 本論文では，マイクロホンアレイ信号処理に基づくロボット聴覚システムのための，フーリエ級数に基づく音響伝達関数モデルのオンライン適応手法について述べる．伝達関数は音源からマイクロホンへの信号伝搬特性を表すものであり，音源定位や分離など，実環境の分析には不可欠である．伝達関数に基づくアレイ信号処理を実環境に応用するには，2 つの特徴が必要である．1) 音響環境の変化に適応できること，2) メモリや計算資源が限られたロボットなどの組み込みシステムで使用するため，伝達関数モデルが軽量であることである．本論文では，上記 2 つの特徴を併せ持ったフーリエ級数展開を用いた軽量な伝達関数モデルのオンライン適応手法を提案する．実験の結果，提案手法を用いてオンラインで適応した伝達関数を用いることで，既存のオンライン伝達関数適応手法よりも音源定位・分離性能が向上することを示した．

1　はじめに

ロボット聴覚 [1, 2] は，ロボットが周囲の音響環境を理解し，人間とロボットのコミュニケーションを実現することを目的とした研究分野である．ロボットは騒音環境や複数の音源が同時に存在する環境でも音を聞き分ける必要があるため，音源定位や音源分離が重要な技術として盛んに研究されている．一般的なロボット聴覚のフレームワークでは，これらの技術を音声認識や音声翻訳，話者識別など様々な音声タスク [3, 4, 5, 6] の前処理として使用することで，実環境における音声対話を実現することができる [7, 8]．音源定位や音源分離の手法は，主に伝達関数ベースの手法と非伝達関数ベースの手法に分けられる．

伝達関数に基づいた手法は，固定ビームフォーミングと適応ビームフォーミングに分類される．典型的な固定ビームフォーマである Delay-and-Sum や Weighted Delay-and-Sum は，与えられた伝達関数セットだけを用いて分離行列を推定する．Maximum Likelihood [9, 10], Minimum Variance Distortionless Response [11] は，半固定ビームフォーミングに分類され，一旦，室内音響を考慮した分離行列を推定するが，推定後は固定ビームフォーマーとして振る舞うため，音響環境の変化に分離行列を適応することができない．適応型ビームフォーマーとしては，Linear Constrained Minimum Variance [12] や Griffith-jim [13] などが提案されている．固定ビームフォーマーとは異なり，適応的に分離行列を推定するため，固定ビームフォーマーよりも優れた環境適応を行うことができる．

非伝達関数ベースの手法には，ブラインド音源分離や深層学習を用いた手法がある．代表的なブラインド音源分離の手法である独立成分分析 (Independent Component Analysis) [14] や独立ベクトル分析 (Independent Vector Analysis) [15, 16] は，伝達関数を用いずに音源分離を行うことができるが，パーミュテーション問題の扱いが困難である．深層学習を用いた手法も活発に研究されている [17, 18, 19, 20, 21, 22]．これらの手法は，伝達関数を測定する代わりに大量のデータを用いて音響環境を学習し，ニューラルネットワークを用いて音源定位，音源分離を実現する．また，音源定位，音源分離，識別モジュールのカスケード接続による誤差蓄積を防ぐため，ニューラルネットワークを用いて複数モジュールを統合する試みもなされている [23]．こ

*連絡先：(株) ホンダ・リサーチ・インスティチュート・ジャパン
〒 351-0188 埼玉県和光市本町 8-1
E-mail: yui.sudo@jp.honda-ri.com

れらの手法は，十分な学習データを用いることで伝達関数ベースの手法と比べて高い性能を示すものの，大量の学習データと高い計算能力が必要であり，現時点ではロボットに適用することは現実的ではない．

したがって，音源定位や音源分離手法をロボットへ適用することを考慮すると，伝達関数ベースの手法が望ましいが，伝達関数ベースの手法には 2 つの問題がある．1) 一つ目は，伝達関数と音響環境とのミスマッチである．通常，伝達関数は時不変な関数として定義され，自由音場を想定した幾何学計算や無響室での音響測定によって得られることが多い [24, 25]．しかし，このようにして得られた伝達関数は，実際の環境での直接測定された伝達関数と一致しないため，音源定位や音源分離の性能が低下する．また，実環境で伝達関数を直接測定したとしても，音響環境が変わるたびに伝達関数を測定し直す必要がある．2）二つ目は，伝達関数のメモリサイズが大きいことである．伝達関数を音源定位や音源分離に利用するためには，各音源から各マイクロホンへの伝搬特性を表す伝達関数が大量に必要となる．すなわち，マイクロホンと考慮する音源方向の数が増えるにつれて，より多くのメモリを必要とする．特に，3 次元空間の音源方向を考慮する場合，伝達関数のサイズは爆発的に増大する [26]．

本論文では，上記の 2 つの問題を解決するために，フーリエ級数展開を用いた軽量伝達関数モデルのオンライン適応手法を提案する．さらに，提案手法を音源定位と音源分離に適用し，その有効性を検証する．なお，本稿は [27] の提案手法をもとに，評価実験を追加した．

2 関連研究

本節では，前節で述べた 2 つの問題 1) 伝達関数と音響環境のミスマッチ，2) 伝達関数のメモリサイズに関連する研究について述べる．

2.1 伝達関数のオンライン適用

伝達関数の適応に関する研究は，マイクロホンアレイのキャリブレーション問題として暗黙に研究されてきた．例えば，Kaung らは，手拍子音を利用して複数のマイクロホン間の時間オフセットを非同期に推定する方法を提案した [28]．Miura らは，手拍子音を用いて Simultaneous Localization And Mapping [29] によりにより，マイクロホン位置，音源位置，オフセット時間を同時に推定するキャリブレーション手法を開発した [30]．この方法は，伝達関数補間と統合し，マイクロホンアレイの伝達関数を直接キャリブレーションすることができる [31, 32]．Dan らは，ベイズモデルと Expectation-Maximization アルゴリズムを用いて，マイクロホンの位置やオフセットなどのパラメータをキャリブレーションする統合的なフレームワークを提案した [33]．

しかし，これらの手法はオフライン処理をベースとしており，手拍子音や Time Stretched Pulse [34] などの特殊な音が必要なため，音源定位や音源分離を行いながらリアルタイムにキャリブレーションを行うことは困難である．さらに，ほとんどの手法は，マイクロホンアレイと音源の位置のキャリブレーションに着目しており，音源定位と音源分離に必要な伝達関数を直接推定するわけではない．そのため，得られたマイクロホン位置と音源位置から幾何学的に伝達関数を推定しなければならず，前節で述べた音響環境とのミスマッチが生じてしまう．

Nakadai らは，これらの問題を解決するために，伝達関数のオンライン適応を提案した [35]．この方法は，上記の方法とは異なり，伝達関数を直接推定することができ，音響環境とのミスマッチを解消することができる．しかし，このオンライン適応手法は，音源方向ごとに離散的な伝達関数を必要とする（以下，離散伝達関数モデルと呼ぶ）．そのため，高い角度分解能を実現するためには，より多くの伝達関数を用意する必要があり，メモリサイズが増大してしまう．

2.2 補間を用いた伝達関数サイズの削減

伝達関数のメモリサイズを小さくするために，補間を用いた手法がいくつか提案されている [26, 36, 37, 38]．Nishino らは，補間にスプライン法を用い，単純な線形補間と比較してその有効性を示した [36]．この方法は，補間により伝達関数の測定回数を減らすことができ，伝達関数のサイズと計算コストを削減することが可能であるが，位相の補間ができないため，音源定位や音源分離の性能が低下する．Duraiswami らは，球面調和関数モデルに基づく頭部関連伝達関数（HRTF）の補間および外挿方法を提案した [37]．このモデルでは，HRTF を高精度に補間することができるが計算コストが高い．

Asahara らは，フーリエ級数展開に基づく軽量な伝達関数モデル（以下，フーリエ伝達関数モデルと呼ぶ）を提案した [26]．この方法は，伝達関数をあらかじめ決められた角度分解能で離散的に伝達関数を持つ離散伝達関数モデルとは異なり，フーリエ級数展開を用いて任意の方向の伝達関数を連続的に補間することで，伝達関数のメモリサイズを削減することができる．しかし，いずれの手法も環境変化に対応することはできないため，伝達関数と音響環境のミスマッチが生じてしまう．

3 提案手法

本節では，離散伝達関数モデル，フーリエ伝達関数モデル [26]，およびフーリエ伝達関数モデルのオンライン適応手法について説明する．

3.1 離散伝達関数モデル

伝達関数は通常，音源方向ごとに離散的に測定され，伝達関数セットとして保持される．伝達関数は $H_m(\omega, \theta_k)$ と表すことができ，ω，$m = 1, 2, ..., M$，θ_k はそれぞれ周波数，m 番目のマイク，k 番目の音源 $(k = 1, 2, ..., K)$ 到来方向を表す．高速フーリエ変換（FFT）により，ω は $\omega = \omega_0 f$ に離散化される．ここで，f は周波数インデックス $(f = 0, 1, 2, ..., F-1)$，F は FFT サイズを表す．本論文では簡単のため，ω は省略する．マイクロホンの数と位置はマイクロホンアレイの配置によって制約されると仮定し，伝達関数セットの角度分解能はあらかじめ決められた音源方向の数 K によって決定される（$360/K$ 度）．全て音源方向の伝達関数 $H_m(\theta_k)$ を保持するために必要なメモリは βKM となる．ここで，β は 1 つの伝達関数に必要なメモリサイズであり，FFT サイズの半分 $(F/2)$ と 1 つの複素数に必要なメモリサイズの積として計算される．例えば，$F = 512$, $K = 72$ (5 度ステップ), $M = 8$, (double) = 8B のとき，必要なメモリサイズは $\beta KM = 1.91$ MiB となる．すなわち，角度分解能を上げるためには K を大きくしなければならず，伝達関数サイズが大きくなる．

3.2 フーリエ伝達関数モデル

前節で述べたように，すべての音源方向ごとに伝達関数を保持する代わりに，伝達関数 $H(\theta_k)$ はフーリエ級数展開を用いて次のように展開することができる．

$$H(\theta_k) = \Sigma_{n=-N}^{N} C_n \exp(in\theta_k), \tag{1}$$

ここで C_n と N はそれぞれ n 番目の複素係数とフーリエ級数展開の次数である．N が $K/2$ より小さい場合，伝達関数モデルには有限次のフーリエ級数展開を用いた近似による誤差が含まれる．離散伝達関数モデルと同じ音源到来方向（例えば，5 度ステップ）を用いて伝達関数を測定する場合（$\theta_k = 2\pi k/K$ の場合），上式は次のように記述される．

$$H(\theta_k) = \Sigma_{n=-N}^{N} C_n \exp(\frac{i2\pi kn}{K}). \tag{2}$$

フーリエ係数 C_n は離散フーリエ変換を使って次のように計算できる．

$$C_n = \Sigma_{k=0}^{K-1} H(\theta_k) \exp(\frac{-i2\pi kn}{K}). \tag{3}$$

また，離散伝達関数モデルと異なる音源到来方向を使用して測定される場合（$\theta_k \neq 2\pi k/K$ の場合），以下のように最小二乗推定法を用いてフーリエ係数を求めることができる．

$$\boldsymbol{H} = \boldsymbol{SC}, \tag{4}$$

ここで，$\boldsymbol{S}$，$\boldsymbol{H}$，$\boldsymbol{C}$ はそれぞれ複素指数関数行列，伝達関数のベクトル，フーリエ係数を表し，以下のように表せる．

$$\boldsymbol{S} = [\boldsymbol{s}(\theta_1), \boldsymbol{s}(\theta_2), ..., \boldsymbol{s}(\theta_K)]^T, \tag{5}$$

$$\boldsymbol{s}(\theta) = [e^{-iN\theta}, e^{-i(N-1)\theta}, ..., e^{i(N-1)\theta}, e^{iN\theta}]^T, \tag{6}$$

$$\boldsymbol{H} = [H(\theta_1), H(\theta_2), ..., H(\theta_K)]^T, \tag{7}$$

$$\boldsymbol{C} = [C_{-N}, C_{-N+1}, ..., C_{N-1}, C_N]^T, \tag{8}$$

また，フーリエ係数は以下のように表される．

$$\boldsymbol{C} = \boldsymbol{S}^{+}\boldsymbol{H}, \tag{9}$$

ここで，$\boldsymbol{S}^{+}$ は $\boldsymbol{S}$ の擬似逆行列である．

フーリエ級数ベースの伝達関数モデルに必要なメモリサイズは $\beta(2N+1)M$ で表されるが，離散伝達関数モデルに必要なメモリサイズは βKM である．例えば，3.1 節で述べたように，$K = 72$ (5 度ステップ) のとき，離散伝達関数モデルのメモリサイズが 1.91MiB であるのに対し，フーリエ伝達関数モデル（$N = 15$ のとき）のメモリサイズは 0.82MiB に削減することができる．また，離散伝達関数モデルは，あらかじめ決められた角度分解能 $(360/K)$ を持つのに対し，フーリエ伝達関数モデルは，任意の角度 $\boldsymbol{s}(\theta)$ が利用可能であるため，メモリサイズを増加させることなく任意の角度分解能 θ で利用することができる．

3.3 フーリエ伝達関数モデルのオンライン適応

フーリエ伝達関数モデルにおけるオンライン適応手法のブロック図を図 1 に示す．

1) 観測信号 $\boldsymbol{X} = [X_1, X_2, ..., X_M]$（$M$ はマイクの数を表す）および，幾何計算や事前測定により求めた伝達関数を用いて音源定位を実行する．観測信号には，手拍子のような特殊な信号は使用しないことに注意されたい．音源定位には，Delay-and-sum や MUSIC (MUltiple SIgnal Classification) [39, 40] などのアルゴリズムを用いることができる．音源定位処理は，伝達関数と入力信号 $\boldsymbol{X}$ が与えられたとき，空間スペクトル A_{sp} が最大になる音源到来方向，θ を求める問題として，以下の式で一般化できる．

$$\theta' = \underset{\theta}{\operatorname{argmax}}(A_{sp}(\boldsymbol{C}, \boldsymbol{X}, \theta)). \tag{10}$$

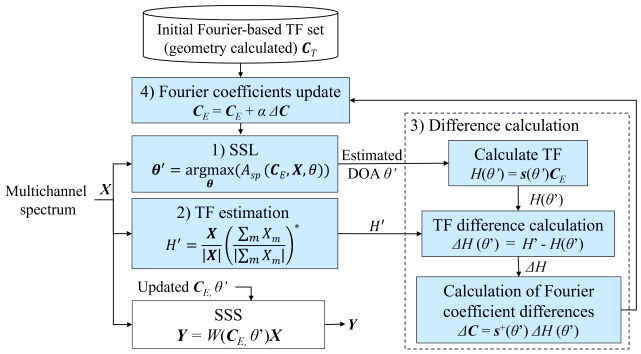


図 1: フーリエ伝達関数モデルにおけるオンライン適応

適応処理の開始直後は，初期伝達関数セットまたはそれに近い値を用いて θ' を推定するため，推定誤差が大きくなる可能性がある．提案する適応手法では，推定された音源到来方向に基づいて $\boldsymbol{C}$ が常に更新されるため，できるだけ正確に音源到来方向を推定することが重要である．音源到来方向の推定誤差を低減するために，過去の L サンプルを用いた線形平滑化を適用し現在の推定値の差が 15 度以上の場合は外れ値除去を行う．このアプローチは，音源がごく短い時間 $L = 13(0.1$ 秒) では連続的に移動すると仮定している．

2) 音源定位によって音源が検出されたら，入力 $\boldsymbol{X}$ を用いて正規化伝達関数を以下の式を用いて推定する．

$$H' = \frac{\boldsymbol{X}}{|\boldsymbol{X}|}\left(\frac{\Sigma_m X_m}{|\Sigma_m X_m|}\right)^*, \tag{11}$$

ここで m と $*$ はマイクインデックスと共役演算子を表す．

3) 次に，現在のフーリエ係数 $\boldsymbol{C}$ を用いて，推定された音源到来方向 θ' の伝達関数 $H(\theta')$ を以下のように計算する：

$$H(\theta') = \boldsymbol{s}(\theta')\boldsymbol{C}. \tag{12}$$

現在の伝達関数と上式を用いて推定された伝達関数の差 $\Delta H(\theta')$ は以下のように計算される．

$$\Delta H(\theta') = H' - H(\theta'). \tag{13}$$

次に，フーリエ係数の差分を計算し，現在の伝達関数と推定伝達関数の差分を補正する．

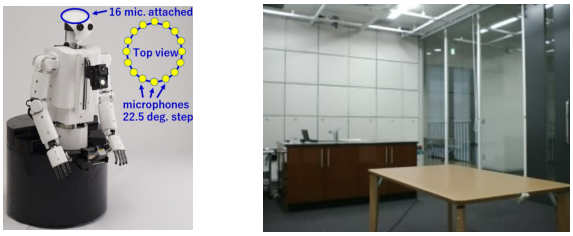
$$\Delta \boldsymbol{C} = \boldsymbol{s}^+(\theta')\Delta H(\theta'), \tag{14}$$

ここで，$\boldsymbol{s}^+(\theta)$ は $\boldsymbol{s}(\theta)$ の擬似逆行列である．

4) 最後に，以下の式を用いてフーリエ係数 $\boldsymbol{C}$ を更新する．

$$\boldsymbol{C} = \boldsymbol{C} + \alpha\Delta\boldsymbol{C}, \tag{15}$$

ここで，α は適応率 (0 - 1) を表す．更新されたフーリエ係数は次の音源定位処理で用いられる．これらの更新処理は，音源が定位される度に随時繰り返される．



(a) HEARBO (15ch)　　(b) 実験室

図 2: 実験環境

4 実験

提案手法の有効性を評価するために 3 つの実験を行った．実験 1 では，提案するフーリエ伝達関数の適応手法が，実環境において既存の離散伝達関数の適応手法と同等に伝達関数を更新できるかどうかを検証するために，更新された伝達関数の振幅スペクトルを比較した．実験 2 と実験 3 では，音源定位と音源分離における性能向上を評価する．音源定位と音源分離には遅延和ビームフォーマを用いた．

4.1 実験 1: 適用後の伝達関数比較

実験 1 では，図 2a に示す HEARBO ロボットの頭部に取り付けた 16 チャンネルの円形マイクアレイを使用し，このうち 15 チャンネルの信号を使用した．実験は，図 2b に示す残響時間 RT60=0.3[s] の 4.0 × 7.0 × 3.0m の部屋で行った．椅子やテーブルなどの障害物はすべて取り除いた状態で HEARBO ロボットを部屋の中央に配置し，HEARBO ロボットの半径 1.5m を 2 周移動しながら，サンプリングレート 48kHz で白色雑音を収録した．収録された音源を用いて伝達関数の適応処理を行った後，伝達関数の振幅スペクトルを比較した．

4.2 実験 2: 音源定位評価

実験 2 では，8 チャンネルの円形マイクアレイを用い，マイクアレイの半径 1.5m を 2 周移動しながら，サンプリングレート 16kHz で白色雑音を収録した．実データを用いた移動音源の音源定位では，音源の基準方向の測定に誤差が生じる可能性があるため，実験 1 の条件を再現したシミュレーション環境を使用した．また，鏡像法 [41] を用いて 3 次反射まで考慮したシミュレーションを行った．音源定位性能の評価には，以下に示すように，音源定位誤差の標準偏差を用いた．

$$\sigma = \sqrt{\frac{1}{I-1}\Sigma_{i=1}^{I}(\theta_{\mathrm{i}} - \theta'_{\mathrm{i}})^2}, \tag{16}$$

ここで，θ_{i}，θ'_{i}，I はそれぞれ参照音源到来方向, 推定された音源到来方向, 総サンプル数を表す．さらに，伝達関数のメモリサイズを測定した．

4.3 実験 3: 音源分離評価

実験 3 では，実験 2 と同様に 8 チャンネルの円形マイクアレイおよびシミュレーション環境を用いて，男性と女性の音声信号を 16kHz のサンプリングレートで収録した．男性と女性の音声信号の到来方向はそれぞれ 103 度，12 度方向とした．音源定位誤差の影響を受けないよう，各音源到来方向は固定した．各音声信号は，CSJ コーパス [42] からランダムに選択した．また，以下に示す SDR（signal-to-distortion ratio）[43] を用いて音源分離性能の評価を行った

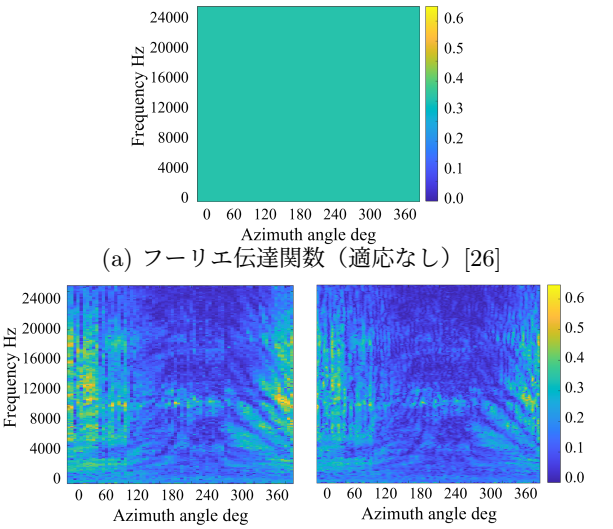
$$SDR(y) = 10\log10(||y_{\mathrm{t}}||^2/||e_{\mathrm{r}}||^2), \qquad (17)$$

ここで，y_{t} は，y に含まれているクリーン音声，e_{r} は含まれている雑音を表す．

5 実験結果

5.1 実験 1: 適応後の伝達関数比較

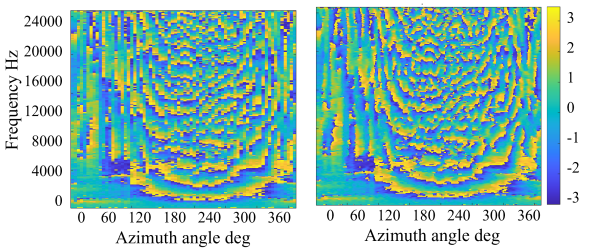
図 3 は，(a) 幾何計算によって算出したフーリエ伝達関数（適応なし）[26]，(b) オンライン適応手法を用いた離散伝達関数 [35]，(c) 提案したオンライン適応手法を用いたフーリエ伝達関数の振幅スペクトルを示す．自由音場を想定した幾何計算により算出された伝達関数は，音源到来方向や周波数成分によらず一定の振幅特性を持つ（図 3a）のに対し，オンライン適応手法を用いることで，実験環境の音響特性を反映するように伝達関数が更新された（図 3b, 3c）．提案手法と従来手法はほぼ同等の振幅特性を示していることから，提案手法が従来手法と同様に実際の環境に適応することができたことがわかる．また，適応後の離散伝達関数モデルは，あらかじめ決められた角度分解能を持つ離散伝達関数であるため，各角度方向間の伝達関数は不連続であるのに対し，提案手法はフーリエ級数展開に基づいて各角度方向の伝達関数を補間することができるため，得られた伝達関数は滑らかである (図 4a, 4b)．



(a) フーリエ伝達関数（適応なし）[26]

(b) 離散伝達関数 [35]　(c) フーリエ伝達関数 (ours)

図 3: 伝達関数の振幅スペクトル



(a) 離散伝達関数 [35]　(b) フーリエ伝達関数 (ours)

図 4: 伝達関数の位相スペクトル

5.2 実験 2: 音源定位評価

5.2.1 音源定位性能

シミュレーション環境における式 (16) の音源定位誤差 σ を表 1 に示す．離散伝達関数モデルとフーリエ伝達関数モデルによらず，オンライン適応により一貫して音源定位誤差を削減することができた (A1-2 vs. B1, C2-3)．提案したフーリエ伝達関数モデルは，補間により任意の角度分解能で音源定位を実行することができる．従来の離散伝達関数モデルでは，角度分解能を高めるためには各角度方向に対して伝達関数を必要とするため，伝達関数のメモリサイズが増大してしまうのに対し，フーリエ伝達関数モデルは，メモリサイズを増やすことなく高い角度分解能で音源定位を実行することができるため，音源定位誤差を削減することができる（B1 vs. C1-2）．また，フーリエ伝達関数モデルのもう一つの利点は，フーリエ級数の次数を減らすことで伝達関数サイズを小さくできることである（C2 vs. C3）．これは近似による音源定位誤差をわずかに増加させるが，それでも音源定位誤差は離散伝達関数モデルと同等であった（B1 vs. C2）．

5.2.2 適応率の影響

次に，図 5 に，式 (15) における適応率 α の効果を示す．適応率が小さい場合，幾何学的情報に基づいて計算された伝達関数と実験環境のミスマッチが十分に改善されないため，音源定位誤差は十分に改善されなかった．反対に，適応率 α を 0.3 より大きくすると，音源定位誤差は発散した．適応率 α=0.03 の時，音源定位誤差が最も小さくなった．

5.2.3 フーリエ次数の影響

図 6 に，フーリエ次数 N の影響を示す．本節では，前節で述べたような伝達関数の更新不十分や発散を防ぐ

表 1: 音源定位の標準誤差

ID	伝達関数モデル	適応	フーリエ次数 N	分解能 (deg)	伝達関数サイズ [MiB]	音源定位誤差 (deg)
A1	離散伝達関数	なし	N/A	5	1.91	8.70
A2	フーリエ伝達関数	なし [26]	35	1	1.91	8.56
B1	離散伝達関数	あり [35]	N/A	5	1.91	**7.41**
C1	フーリエ伝達関数	あり (ours)	15	5	**0.82**	**7.50**
C2	フーリエ伝達関数	あり (ours)	15	**1**	**0.82**	**7.40**
C3	フーリエ伝達関数	あり (ours)	25	**1**	**1.36**	**7.30**

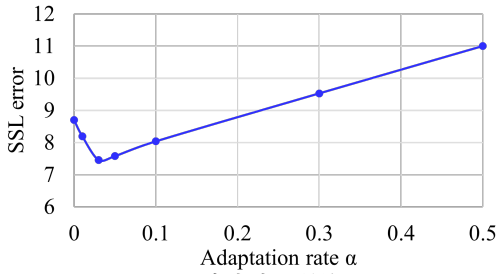


図 5: 適応率の影響

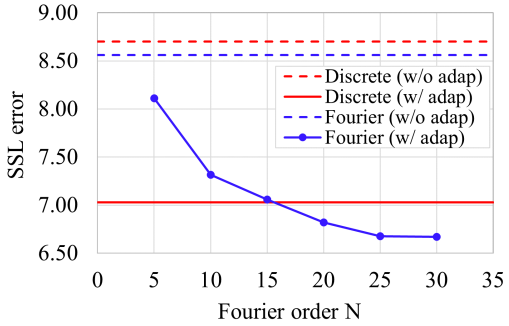


図 6: フーリエ次数の影響

ため，適応率 α=0.01 とし，4.2 節で述べた適応ステップを 10 回繰り返した．また，角度分解能は離散伝達関数モデルと同様に 5 度とした．適応なしの場合（破線）と比較して，フーリエ次数によらず適応を行うことで音源定位誤差が小さくなった．さらに，フーリエ次数 N が大きいほど，伝達関数サイズは大きくなるが，音源定位誤差は小さくなった，本実験条件では，N=35 未満で離散伝達関数モデルよりもメモリサイズを小さくすることができるため，$15 < N < 35$ の範囲において，提案手法は従来手法よりも小さいメモリサイズで小さい音源定位誤差を達成できることがわかった．

5.3 実験 3: 音源分離評価

表 2 に音源分離による SDR の改善効果を示す．提案した適応方法を用いたフーリエ伝達関数モデルは，従来の離散伝達関数モデルよりも大きい SDR 改善効果が得られた．これは，フーリエ伝達関数モデルでは，補間を用いて任意の高い角度分解能を利用できるためと考えられる．さらに，音源定位タスクと同様に，フーリエ伝達関数もでるではフーリエ級数の次数を減らすことで伝達関数サイズを小さくすることができる．フーリエ級数の次数を減らすと近似誤差が増えるため，SDR 改善向上がわずかに減少するが，離散伝達関数法よりも大きい SDR 改善が見られた．

表 2: 音源分離タスクにおける SDR 改善 (dB)

伝達関数モデル	適応	フーリエ次数 N	伝達関数サイズ [MiB]	SDR 改善
離散伝達関数	なし	N/A	1.91	1.38
フーリエ伝達関数	なし [26]	35	1.91	2.23
離散伝達関数	あり [35]	N/A	1.91	3.92
フーリエ伝達関数	あり (ours)	15	**0.82**	**6.02**
フーリエ伝達関数	あり (ours)	25	**1.36**	**6.11**

6 議論

音源定位タスクと音源分離タスクにおける提案手法の性能を比較した結果，提案手法は音源分離タスクにおいて SDR が 3 倍向上したのに対し，音源定位タスクでは約 15%の誤差低減に留まった．この差の理由としては，音源分離タスクでは，音源方向が固定されていたため，音源方向の伝達関数が完全に適応されたのに対し，音源定位タスクでは，音源が円周方向に動き続けていたため，伝達関数の更新が十分でなかったことに起因すると考えられる．適応率を上げることで伝達関数の適応を高速化する可能性がある一方で，発散につながる可能性もあるため，発散することなくより高速に環境に適応できる伝達関数更新手法のさらなる検討が必要であると考える．

7 結論

本論文では，フーリエ級数展開に基づく軽量伝達関数モデルのオンライン適応手法を提案した．提案手法は，伝達関数と音響環境とのミスマッチを防ぐことにより，音源定位と音源分離性能を改善した．また，フーリエ級数に基づく伝達関数適応手法は，補間により任意の高い角度分解能を使用することができるため，従来の離散伝達関数モデルのオンライン適応手法よりも高い性能を示した．さらに，フーリエ級数展開の次数を小さくすることで，性能劣化を小さく抑えながら伝達関数サイズを小さくすることができた．今後は，より高速で高精度な適応手法の研究を行う予定である．

参考文献

[1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2000, pp. 832–839.

[2] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. ICASSP*. IEEE, 2015, pp. 5610–5614.

[3] Y. Peng, J. Tian, B. Yan, D. Berrebbi, X. Chang, X. Li, J. Shi, S. Arora, W. Chen, R. Sharma *et al.*, "Reproducing whisper-style training using an open-source toolkit and publicly available data," *arXiv preprint arXiv:2309.13876*, 2023.

[4] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[5] Y. Peng, Y. Sudo, S. Muhammad, and S. Watanabe, "Dphubert: Joint distillation and pruning of self-supervised speech models," in *Proc. Interspeech*, 2023, pp. 62–66.

[6] Y. Sudo, M. Shakeel, B. Yan, J. Shi, and S. Watanabe, "4D ASR: Joint modeling of CTC, attention, transducer, and mask-predict decoders," in *Proc. Interspeech*, 2023, pp. 3312–3316.

[7] K. Nakadai, G. Ince, K. Nakamura, and H. Nakajima, "Robot audition for dynamic environments," in *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*. IEEE, 2012, pp. 125–130.

[8] K. Nakadai and H. G. Okuno, "Robot audition and computational auditory scene analysis," *Advanced Intelligent Systems*, vol. 2, no. 9, 2020.

[9] V. Barroso and J. Moura, "Maximum likelihood beamforming in the presence of outliers," in *Proc. ICASSP*, 1991, pp. 1409–1412 vol.2.

[10] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.

[11] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[12] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," in *Proceedings of the IEEE*, vol. 60, no. 8, 1972, pp. 926–935.

[13] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[14] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[15] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Independent Component Analysis and Blind Signal Separation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 601–608.

[16] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Process.*, vol. 87, no. 8, pp. 1859–1871, 2007.

[17] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Sound event aware environmental sound segmentation with Mask U-Net," *Advanced Robotics*, vol. 34, pp. 1280–1290, 2020.

[18] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[19] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Improvement of DOA estimation by using quaternion output in sound event localization and detection," in *Proc. DCASE*, 2019, pp. 244–247.

[20] Z. Zhang, T. Yoshioka, N. Kanda, Z. Chen, X. Wang, D. Wang, and S. E. Eskimez, "All-neural beamformer for continuous speech separation," in *Proc. ICASSP*. IEEE, 2022, pp. 6032–6036.

[21] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Multi-channel environmental sound segmentation utilizing sound source localization and separation U-Net," in *2021 IEEE/SICE International Symposium on System Integration (SII)*, 2021, pp. 382–387.

[22] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "A sequence matching network for polyphonic sound event localization and detection," in *Proc. ICASSP*. IEEE, 2020, pp. 71–75.

[23] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Multichannel environmental sound segmentation with separately trained spectral and spatial features," *Applied Intelligence*, vol. 51, pp. 8245–8259, 2021.

[24] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1119–1123, 1995.

[25] G.-B. V. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of The Audio Engineering Society*, vol. 50, pp. 249–262, 2002.

[26] Y. Asahara, K. Matsuda, H. Nakajima, and K. Nakadai, "A Fourier series based data compression model for acoustic transfer function," in *2020 IEEE/SICE International Symposium on System Integration (SII)*, 2020, pp. 664–668.

[27] Y. Sudo, M. Takigahira, H. Tsuru, K. Nakadai, and H. Nakajima, "Online adaptation of fourier series based acoustic transfer function model to improve sound source localization and separation," in *Proc. RO-MAN*, 2023.

[28] Kuang, Yubin and Åström, Karl, "Stratified Sensor Network Self-Calibration From TDOA Measurements," in *Proc. EUSIPCO*, 2013.

[29] T. Bailey, J. Nieto, J. Guivant, M. Stevens, and E. Nebot, "Consistency of the EKF-SLAM algorithm," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 3562–3568.

[30] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based online calibration for asynchronous microphone array," *Advanced Robotics*, vol. 26, no. 17, pp. 1941–1965, 2012.

[31] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 694–699.

[32] K. Nakamura, S. Ambrose, and K. Nakadai, "Slam-based online calibration of asynchronous microphone array for robot audition," in *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, 2011, pp. 524–529.

[33] K. Dan, K. Itoyama, K. Nishida, and K. Nakadai, "Calibration of a microphone array based on a probabilistic model of microphone positions," in *Trends in Artificial Intelligence Theory and Applications. Artificial Intelligence Practices: 33rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2020, Kitakyushu, Japan, September 22-25, 2020, Proceedings*. Springer, 2020, pp. 614–625.

[34] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1484–1488, 1981.

[35] K. Nakadai, M. Takigahira, Y. Kawai, and H. Nakajima, "Fully-online always-adaptation of transfer functions and its application to sound source localization and separation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2100–2105.

[36] T. Nishino, S. Kajita, K. Takeda, and F. Itakura, "Interpolating head related transfer functions in the median plane," in *Proc. WASPAA*. IEEE, 1999, pp. 167–170.

[37] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs [head related transfer functions]," in *Proc. ICASSP*, vol. 4. IEEE, 2004, pp. 45–48.

[38] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[39] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[40] F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[41] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[42] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[43] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.