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-058-07

伝達関数の常時オンライン適応による音源定位・分離の向上

Improvement of Sound Source Localization and Separation with Fully-Online Always-Adaptation of Transfer Functions

中臺 一博[1,2*]　瀧ケ平 雅行[1]　河合 熊輔[3]　中島 弘史[3]
Kazuhiro NAKADAI[1,2]　Masayuki TAKIGAHIRA[1]　Yusuke KAWAI[3]　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，ロボット聴覚システムのような実環境で用いることを前提としたマイクロホンアレイ処理を対象に，マイクロホンアレイ処理で用いる伝達関数を常時オンライン適応する手法について述べる．マイクロホンアレイ処理における伝達関数とは，マイクロホンと音源の間の信号伝播特性を表し，音源定位や音源分離処理に欠かせない情報である．一般には，伝達関数は，時不変・静的な関数として定義することが多いが，実用性を考えると，室内音響やその環境変化を伝達関数として考慮する必要があり，動的な関数として定義することが好ましい．本研究では，音源定位や音源分離の実行中でも，観測信号から伝達関数を動的かつ連続的に推定できる常時オンライン適応手法を提案する．提案手法を，ロボット聴覚オープンソースソフトウェア HARK 上に，オンライン動作可能なモジュールとして実装し，実装したモジュールを用いて構築した音源定位・分離システムを，オフィス環境で実収録したデータを用いて評価した．結果として，対象とする実験環境であらかじめ収録したデータから作成した静的な伝達関数を用いた場合と同等の音源定位・分離性能を得ることができ，提案手法の有効性を示すことができた．

1　はじめに

ロボット聴覚 [1] は，人・ロボットコミュニケーションや，ロボットによる音環境理解を実現することを目的に提案された研究分野である．ロボットは，騒音下や複数音源が同時に存在する場合でも音を聞き分ける必要があるため，音源定位と音源分離は，その主要技術として，活発に研究が行われてきた．音源定位・音源分離手法は主に，固定ビームフォーミング，適応ビームフォーミング，ブラインド分離，深層学習ベースの手法の 4 つのグループに分類することができる．

固定ビームフォーミングは，“fully-fixed” 型と “semi-fixed” 型の 2 種類に分けることができる．Delay-and-Sum (DS)，NULL (NULL)，Weighted Delay-and-Sum (WDS) は fully-fixed 型固定ビームフォーミングであり，その分離行列はあらかじめ与えられた伝達関数 (Transfer Function, TF) のみを用いて推定される．Maximum Likelihood (ML) [2, 3]，Minimum Variance Distortionless Response (MVDR) [4] は semi-fixed 型ビームフォーミングであり，一旦，室内音響を考慮して分離行列を推定するものの，推定後は，fully-fixed ビームフォーミングとして振る舞うため，環境変化に適応できない．Linear Constrained Minimum Variance (LCMV) [5] や Griffith-Jim (GJ) [6] などは，適応型ビームフォーミングに属するが，分離行列の推定には，伝達関数を用いる．固定ビームフォーミングとの違いは，分離行列が適応的または逐次的に推定されることで，固定ビームフォーミングよりも優れた環境適応性能を発揮することである．Independent Component Analysis (ICA) [7] や Independent Vector Analysis (IVA) [8, 9] はブラインド音源分離の代表的な手法である．これらの手法は，伝達関数を用いずに分離処理が可能であるが，パーミュテーション問題が発生するなど処理の制御が難しい．この問題を解決するため，Geometric Source Separation (GSS) [10], Geometric Independent Component Analysis (GICA) [11]，Geometric High-order Decorrelation based Source Separation (GHDSS) [12] など，伝達関数から得られる空間情報と統合する手法が提案されている．なお，これらのオンライン処理版も，適応型ビームフォーミングの一種といえる．深層学習を利

*（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: nakadai@jp.honda-ri.com

用した手法は近年，活発に研究されている [13, 14, 15]．これらの手法は，伝達関数を用いる代わりに，大量のデータを用いて音響環境を学習し，高い性能を得ているが，大量のデータと学習用に高い計算能力を持った計算機を必要とするため，現時点ではロボットにとって実用的とはいえない．

以上から，音源定位・分離の実用性を考えると，伝達関数を用いる手法の方が好ましいと考えられる．一方で，伝達関数は，通常，静的な関数として定義され，自由音場を仮定した幾何計算や，無響室での多方向からの音響測定によって得ることが多い [16, 17]．しかし，このようにして得られた伝達関数は，直接対象環境で測定した結果から得られる伝達関数と比較すると，ミスマッチが生じるため，音源定位・分離性能が低下してしまう．また，対象環境で測定した結果から得られる伝達関数を使う場合には，対象環境中の物体の配置が変わったり，対象環境自体が変わったりする度に伝達関数の再測定が必要となってしまうといった問題もある．

この問題は，これまでに報告されてきた手法では，伝達関数はマイクロホンと音源間の幾何学的関係を記述するためだけに用いる静的関数であると定義していることに起因する．実用性を考えれば，室内音響やその変化を含めた動的関数として定義するべきである．本稿では，伝達関数を動的関数として再定義し，伝達関数のオンライン適応法を提案する．さらに，提案手法を音源の定位と分離に適用し，オンラインデモを通じて，その有効性を検証する．

2 関連研究

伝達関数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のうちに研究されてきた [18, 19, 20, 21, 22]．Thrun ら [18] は，マイクロホンで観測された音響信号を用いて，未知のマイクロホン位置を推定するオンラインキャリブレーション技術を提案した．彼らは，数値シミュレーションと実環境評価実験を通じて，この手法の有効性を示している．Kaung ら [19] は，音源・マイクロホン間の距離情報を事前に与える必要があるものの，手拍子音を用いて非同期で複数マイクロホン間の時間オフセットを推定する方法を報告している．Ono ら [20] は，位置が不明な非同期マイクロホンのキャリブレーションをブラインドアライメント問題と定義し，この問題を解くために必要なマイクロホン数や観測数といった条件を明らかにしている．また，この問題をオフラインで高精度に解く補助関数ベースの手法を提案した．Miura ら [21] は，手拍子音を用いて Simultaneous Localization And Mapping (SLAM) により，マイクロホンの位置，音源位置，オフセット時間を同時に推定するオンラインキャリブレーション手法を提案した．この手法を，伝達関数補間 [23] 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よう拡張した手法も提案されている [24]．さらに，話者ダイアリゼーションに適用するため，オフラインクラスタリング処理に拡張した手法も報告されている [25]．Dan ら [22] は，マイクロホンアレイの位置やオフセットなどのパラメータをベイズモデルを用いて Expectation-Maximization (EM) 法でキャリブレーションを行う統一的なフレームワークを報告した．この手法は，キャリブレーション時に，混合音源を用いることができるという特長を有している．

しかし，これらの方法は，いずれもオフライン処理がベースとなっている [18, 19, 20, 25]．また，キャリブレーション処理をオンラインで行うことができる手法も，キャリブレーション処理自体は，事前に行っておく必要がある [21, 24]．また，手拍子音や Time Stretched Pulse (TSP) などのキャリブレーション用の特殊な音が必要であること [26]，計算コストも高いことから，音源定位・分離を行いながら，実時間でキャリブレーション処理を行うことは難しい．さらに，ほとんどの手法はマイクロホン位置や音源位置のキャリブレーションに重点を置いており，音源定位・分離で必要な伝達関数を直接推定できるわけではない．このため，得られたマイクロホンや音源位置から，幾何計算で伝達関数を推定しなければならず，前節で述べたようなミスマッチ問題が生じてしまう．

ロボット用の音源定位・分離を対象としたマイクロホンアレイ処理のためには，以下の 4 つの要件を満たす必要がある．つまり，伝達関数の「キャリブレーション」ではなく「オンライン適応」が必要といえる．

1. 手拍子音や TSP などの特殊な音源を使わずに，音声を含む任意の音源を用いることができる．
2. マイクロホンや音源の位置をキャリブレーションする代わりに，伝達関数を直接得ることができる．
3. キャリブレーションのような前処理を必要とせず，常時オンライン適応ができる．
4. オンライン適応は，音源定位・分離などのマイクロホンアレイ処理と同時に行うことができる．

3 提案手法

図 1 は，システム全体の構成を示しており，伝達関数オンライン適応ブロックとロボット聴覚機能ブロックの 2 つのブロックから構成される．このシステムは，M チャネルのマイクロホンアレイから得られるマルチ

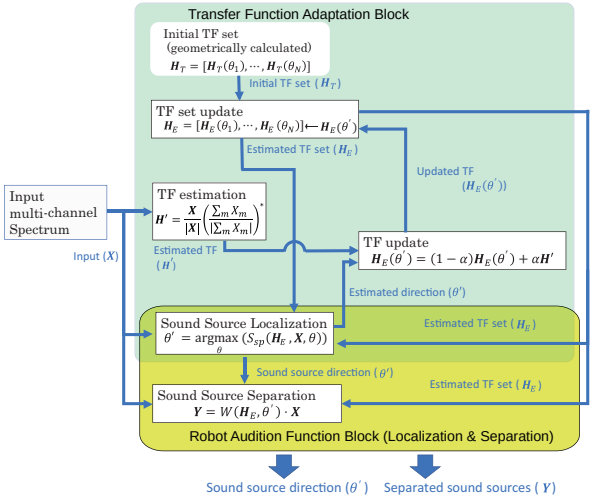


図 1: Proposed Framework for fully-online always-adaptation of TFs with sound source localization and separation

チャネルの音声信号 $\boldsymbol{X}(\omega) = [X_1(\omega), \cdots, X_M(\omega)]^T$ を入力とし，音源方向 θ' と分離音源 $\boldsymbol{Y}(\omega)$ を出力する典型的なロボット聴覚システムとして構成されている．以下，説明を簡単にするために，ω を省略する．

3.1 伝達関数オンライン適応ブロック

このブロックは，常時オンライン適応を行う．入力は観測信号 $\boldsymbol{X}$ である．つまり，入力信号は，手拍子音や TSP のような特別な信号は想定せず，観測信号をそのまま用いる．$\boldsymbol{X}$ は，まず，音源定位処理に入力される．音源定位では，処理開始時には，あらかじめ幾何計算や測定から得られた初期伝達関数セット $\boldsymbol{H}_T$ を用いて音源定位を行う．

$$\boldsymbol{H}_T = [\boldsymbol{H}_T(\theta_1), \boldsymbol{H}_T(\theta_2), \cdots, \boldsymbol{H}_T(\theta_N)], \tag{1}$$

ここで，N は水平角方向の分割数である．本稿では，簡単のために 1 次元の方位角 θ を想定しているが，$[\theta, \phi]$ に置き換えることで，理論的には容易に 2 次元に拡張することができる．

定位には，DS や MUltiple SIgnal Classification (MUSIC) [27] など，任意のアルゴリズムを適用できる．音源定位処理は，伝達関数 $\boldsymbol{H}_T$ と入力信号 $\boldsymbol{X}$ が与えられたときに，空間スペクトル S_{sp} が最大になる θ を求める問題として，以下のように一般化した形で定義できる．

$$\theta' = \underset{\theta}{\mathrm{argmax}}(S_{sp}(\boldsymbol{H}_E, \boldsymbol{X}, \theta)), \tag{2}$$

処理開始直後，しばらくは，$\boldsymbol{H}_T$，またはそれに近い値を用いて θ' を推定するため，大きな推定誤差を生じる可能性がある．しかし，$\boldsymbol{H}_T$ は $\boldsymbol{H}_E$ として常時更新されるため，十分に適応が進んだ後は，正確な θ' を出力することが期待できる．この仮説は，評価実験の節で検証を行う．

伝達関数推定モジュールでは，以下の式により，入力 $\boldsymbol{X}$ から正規化伝達関数を推定する．

$$\boldsymbol{H}' = \frac{\boldsymbol{X}}{|\boldsymbol{X}|} \cdot \left(\frac{\sum_m X_m}{|\sum_m X_m|} \right)^*, \tag{3}$$

ここで，m はマイクロホンのインデックスを示し，* は共役演算子を示す．つまり，振幅は各周波数でのノルムが 1 になるように正規化され，位相は各周波数での平均が 0 になるように正規化される．

次に，伝達関数更新モジュールでは，$\boldsymbol{H}'$ と θ' から，$\boldsymbol{H}_E(\theta')$ を次のように更新する．

$$\boldsymbol{H}_E(\theta') = (1-\alpha)\boldsymbol{H}_E(\theta') + \alpha\boldsymbol{H}', \tag{4}$$

ここで，α は重みパラメータである．本稿では，実験的に 0.5 としている．

更新された $\boldsymbol{H}_E(\theta')$ は，伝達関数セット更新モジュールに送られる．このモジュールでは，伝達関数セット $\boldsymbol{H}_E$ の中の θ' に対する伝達関数を，単純に $\boldsymbol{H}_E(\theta')$ に置き換える．更新された伝達関数セットは，次のタイムフレームでの音源定位処理で使用される．これらの処理を，入力信号を受信するたびに繰り返す．

3.2 ロボット聴覚機能ブロック

このブロックでは，音源定位と音源分離処理を行う．この 2 つの処理は，いずれも伝達関数セットが必要であり，伝達関数セットには常に最新の伝達関数セット $\boldsymbol{H}_E$ を使用する．音源定位は，伝達関数オンライン適応ブロックと共通となっており，式 (2) にしたがい，音源定位を行う．

音源分離は，一般に，下式の線形分離過程として記述できる．

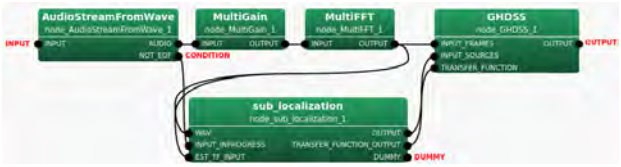
$$\boldsymbol{Y} = \boldsymbol{W}(\boldsymbol{H}_E, \theta') \cdot \boldsymbol{X}, \tag{5}$$

ここで，$\boldsymbol{Y}$ は分離信号であり，$\boldsymbol{W}$ は伝達関数セット $\boldsymbol{H}_E$ と音源方向 θ' から得られる分離行列である．音源定位から得られた音源方向 θ' を対象とした処理であるため，分離処理では完全なセットではなく，部分セットを用いる．このため，$\boldsymbol{W}(\boldsymbol{H}_E, \theta')$ は $\boldsymbol{W}(\boldsymbol{H}_E(\theta'))$ と書き換えることができる．

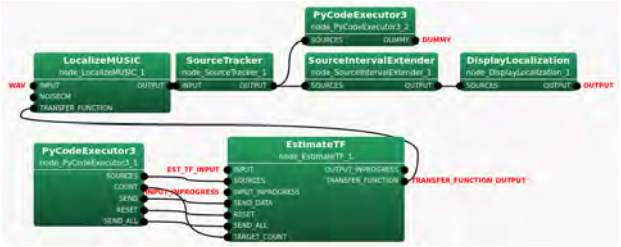
4 実装

提案したフレームワークを，HARK[1] [28] 上に実装した．HARK は 2008 年に OpenCV の音声版を目指

[1]Honda Research Institute Japan Audition for Robots with Kyoto University の略. https://www.hark.jp/



a) main network



b) sub_localization network

図 2: Implemented HARK network

してオープンソース化したロボット聴覚ソフトウェアである．HARK には，MUSIC ベースの音源定位アルゴリズム，12 種類の一般的な音源分離アルゴリズム，深層学習ベースの音声認識をはじめとしたロボット聴覚の主要機能に加え，ロボット聴覚システムの構築に必要な一通りの機能が含まれている．総ダウンロード数は，2021 年 10 月時点で約 22.5 万件となっている．ユーザーの使い勝手を考慮した GUI プログラミング環境の採用，ロボットでの使用を想定した実時間処理といった特長を有している．提案フレームワークの実装では，HARK の既存の資源を活用し，伝達関数適応用の新しいモジュールを 1 つと，音源定位・分離用に既存の 2 つのモジュールの修正のみで実現することができた．図 2は，新たに実装したモジュールと修正したモジュールから構成された HARK ネットワーク (プログラムに相当) を示している．図 2a) はメインのネットワークで，図 2b) は図 2a) 中の仮想モジュール sub_localization の詳細を示している．新たに実装した EstimateTF は，図 1の音源定位部分を除いた伝達関数オンライン適応ブロックに相当している．音源定位・分離のアルゴリズムには，MUSIC と GHDSS を選択した．この 2 つの手法に対応するモジュール LocalizeMUSIC, GHDSS は，もともと静的な伝達関数セットしか利用できなかったため，随時伝達関数の更新情報を受け取ることができるように修正を行っている．

実装の際には，伝達関数セットのサイズが 2〜5MB と大きいことを考慮し，LocalizeMUSIC と EstimateTF/GHDSS 間の通信は，伝達関数セット $\boldsymbol{H}_E$ 全体の更新だけでなく，部分的な伝達関数セット $\boldsymbol{H}_E(\theta)$ の更新にも対応させ，処理速度の向上を図っている．Tab. 1に，図 2で使用したモジュールとその機能の一覧を示す．

表 1: List of HARK modules used in Fig. 2

Module	Function
AudioStreamFormWav	Reading the multi-channel audio data, dividing it into frames, and send them
MultiGain	Gain normalization of the multi-channel audio data
MultiFFT	Frequency analysis for the multi-channel audio data
GHDSS	Sound source separation with GHDSS (modified)
sub_localization	Virtual node containing the network in Fig. 2b)
LocalizeMUSIC	Sound source localization with MUSIC (modified)
SourceTracker	Sound source tracking
SourceIntervalExtender	Adjustment of sound source tracking results
DisplayLocalization	Display of sound source localization/ tracking results
EstimateTF	The proposed TF adaptation (newly implemented)
PyCodeExecutor3	Module written with python3 to communicate localization/tracking results across frames

5 評価

以下の 4 種類の性能評価実験を通じて，提案手法の検証を行った．

1. 伝達関数推定
2. 音源の定位
3. 音源分離
4. 常時オンライン適応デモを通じた音源定位性能実証

5.1 利用したマイクロホンアレイ

実験では，図 3に示すように 2 種類のマイクロホンアレイを使用した．1 つは 8 ch の円形マイクロホンアレイ Tamago [2] で，もう 1 つは，ロボット Hearbo の頭部に設置した 16 ch の円形マイクロホンアレイである．なお，実際に利用した信号は，このうち，15 ch 分である．実験はすべて，広さ $4 \times 7 \times 3$ m，残響時間 $RT_{60} = 0.3$[s] の部屋で行った．Tamago を使用する際には，机と椅子を設置，机上にノートパソコンやいくつかのオブジェクトを置いた状態で実験を行った．Tamago も机の上に設置した (床からの高さは 0.9 m)．Hearbo を使用する際は，椅子，机，を含め，すべての物体を取り除き，ロボットを部屋の中央に配置して実験を行った．

[2] http://www.sifi.co.jp/

5.2 データ準備

まず，Tamago と Hearbo 用の幾何計算伝達関数として，TF_T^G, TF_H^G をマイクロホンと音源位置から計算して作成した．

次に，Tamago では，24bit，16kHz サンプリングで，以下のデータの収録/作成を行った．

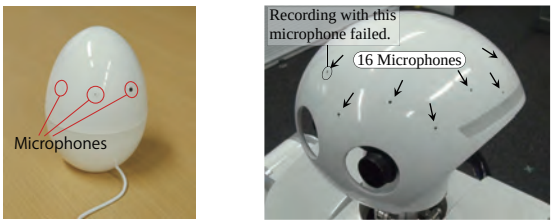
TF_T^L: 計測伝達関数セット（スピーカ高: 低）
TF_T^M: 計測伝達関数セット（スピーカ高: 中）
W_T: Tamago の周りを移動しながら録音した白色雑音
S_T: Tamago の周りを移動しながら録音した音声
M_T: 音源分離用の同時発話音声

また，Hearbo では，24bit，48kHz のサンプリングで，以下のデータの収録/作成を行った．

TF_H^H: 計測伝達関数セット（スピーカ高: 高）
W_H: ロボットの周りを移動しながら録音した白色雑音

上述のように，Tamgo 用の測定伝達関数は TF_T^L と TF_T^M の 2 種類を作成した．TF_T^L については，Tamago の周りを 0 度から 360 度まで 30 度間隔で収録した TSP 信号から作成した．この際，Tamago とスピーカの距離は 0.78 m，スピーカの高さは Tamago に対して 15.8 度下向きとした．TF_T^M については，スピーカの高さを 1.0 m とし，人が椅子に座ったときの口の高さを想定して，上方向に 7.3 度としたことを除き，TF_T^L と同じ条件で収録したデータから作成した．Hearbo 用の伝達関数 TF_H^H は，0 度から 360 度まで 5 度間隔で収録した TSP 信号から作成した．Hearbo までの距離は 1.5 m とし，スピーカの高さは人が立ったときの口の高さを想定して 1.5 m とした．

白色雑音 W_T は，スピーカを人が手に持って Tamago の周りを時計回りに 1 周した後，反時計回りにもう 1 周することを 6 回繰り返し，6.8 分間のデータとして収録した．この際，Tamago までの距離は，約 0.78 m，スピーカの高さは約 1.0 m になるように手動で調整した．音声データ S_T は，Tamago までの距離とスピーカの高さは W_T と同じだが，Tamago の周りを時計回りに 3 周し，発話長 20 分程度のデータとして収録した．音源には，日本語話し言葉コーパス (CSJ) [29] から選んだ男性の音声を用いた．さらに，音源分離の評価用の同時発話音声として，M_T を作成した．作成のため，まず，2 本のスピーカを Tamago から 0 度と 60 度の方向に，高さ 1.0 m，距離 0.78 m となるよう配置し，各スピーカから CSJ から選んだ 2 人の男性の音声データを同時に再生，100 秒の同時発話音声として録音した．さらに，0 度方向の音声信号に対し，SN 比が 20dB になるように白色雑音を重畳し，M_T とした．



a) Tamago (8 ch)　　b) Hearbo (used 15 ch)

図 3: Microphone arrays

Hearbo については，白色雑音 W_H のみを収録した．W_H は，人がロボットから 1.0m の距離を保ちながら，スピーカを手で持ち，ロボットの周りを 2 周し，15 秒の白色雑音として収録した．

5.3 実験と評価指標

伝達関数推定の性能は，白色雑音 W_T を用いて提案手法で推定した伝達関数セットと，3 つの伝達関数セット TF_T^L, TF_T^M, TF_T^G それぞれとの差を，平均二乗誤差 (MSE) で比較することで測定した．方向 θ に対する，2 つの伝達関数セット，TF_i と TF_j の MSE は次のように定義した．

$$MSE_{ij}(\theta) = \frac{1}{MF}\sum_m\sum_f |TF_i(m,f,\theta) - TF_j(m,f,\theta)|^2, \tag{6}$$

ここで，M と F はマイクと周波数ビンの数を示し，m と f はそれらのインデックスである．

音源定位性能の評価には，W_H を用いて推定した伝達関数，TF_H^H, TF_H^G の Hearbo 用の 3 種類の伝達関数セットを用いた．評価指標には定位誤差（L_E）を用い，次のように定義した．

$$L_E = \frac{N_E}{N_T}, \tag{7}$$

ここで N_E は削除誤りを含む定位を誤ったフレーム数を示す．定位結果は正解方向と同じであれば成功とみなし，これをフレームベースで評価した．N_T は，定位したフレームのうち，ある閾値以上のパワーを持つフレームの総数を示す．閾値には，-5 dB と -10 dB を選択した．定位アルゴリズムには，DS を用い，評価データとして W_H を使用した．この評価実験は，提案手法の基本性能を確認することが目的であるため，最も単純なアルゴリズムである DS を用いたクローズドテストとして実施した．

音源分離の性能評価には，GHDSS, DS, LCMV, NULL, MVDR の 5 種類のアルゴリズムを選択し，M_T に対して分離を行い評価した．DS と NULL は fully-fixed 型固定ビームフォーミング，MVDR は semi-fixed 型固定ビームフォーミング，LCMV と GHDSS は適応型ビームフォーミングである．M_T には，拡散性の白

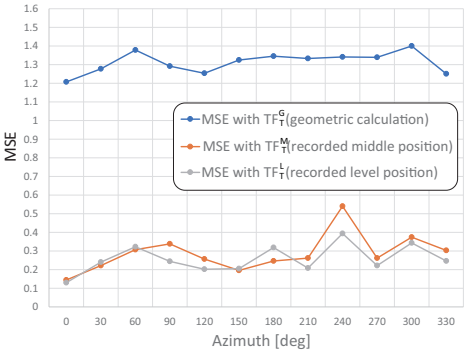


図 4: MSE in TF estimation

色雑音が含まれているため，音源分離の後処理として Histogram-based Recursive Level Estimation (HRLE) [30] による音声強調処理を行った．各分離アルゴリズムにおいて，提案手法で推定した伝達関数セット W_T および，TF_T^M, TF_T^G の 3 種類の Tamago 用伝達関数セットを用いた場合の結果を比較した．評価指標には，Signal-to-Distortion (SDR) と Signal-to-Interference (SIR) [31][3] を用いた．SDR と SIR の定義は以下の式で表される．

$$SDR(s) = 10\log_{10}\left(||s_{\text{target}}||^2/||e_{\text{residue}}||^2\right), \quad (8)$$

$$SIR(s) = 10\log_{10}\left(||s_{\text{target}}||^2/||e_{\text{interf}}||^2\right), \quad (9)$$

ここで，s_{target} は s に含まれるクリーン音声信号を，e_{residue} は分離信号 $\hat{s} = s_{\text{target}} + e_{\text{residue}}$ に含まれる雑音の残差項を，e_{interf} は e_{residue} に含まれる干渉雑音を示す．SDR と SIR の改善は，分離信号と観測信号の SDR と SIR の差として定義する．

また，実際に W_T を用いて明示的なキャリブレーションを行う従来型の手法，S_T を定位と伝達関数適応の両方に用いながら常時オンライン適応する提案手法の 2 種類の伝達関数適応手法を音源定位と組み合わせたオンラインデモを通じて結果を比較した．

5.4 結果

図 4は，推定した伝達関数セットと，TF_T^G, TF_T^M, TF_T^L との MSE を示している．TF_T^G は誤差が大きく，音源定位・分離の性能が低いことがわかる． TF_T^M と TF_T^L は，方位角によらず，MSE が小さく保たれていることから，提案手法による環境適応が良好に働いているといえる．スピーカの高さは，TF_T^L よりも TF_T^M の方が評価データ収録環境に近いが，両者の MSE 値はほぼ同等となっている．これは，評価データ収録時にスピーカの高さを手動で調整したため，高さ調節が正確でなかったためと考えられる．

[3] http://bass-db.gforge.inria.fr/bss_eval/

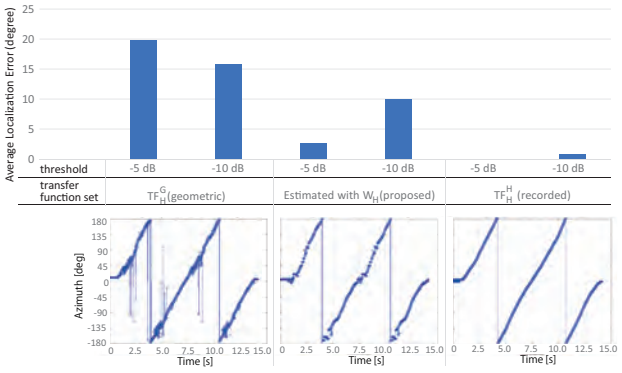


図 5: Sound source localization results: the upper panel shows the average localization errors and the lower panel shows localization results at the threshold of -10 dB.

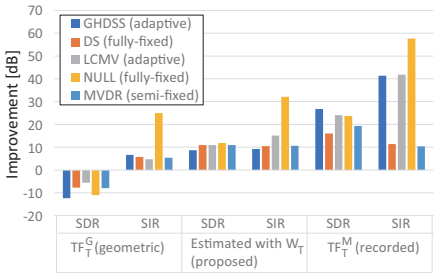


図 6: Sound source separation results

図 5 に定位結果を示す．計測ベースの伝達関数セット TF_H^H では誤差がほぼ 0 であったのに対し，幾何計算伝達関数セット TF_H^G では誤差が大きくなっている．これは，下段の定位結果から明らかなように，外れ値が多く検出されたためである．推定伝達関数セットは TF_H^H を用いた場合と近い性能を示し，外れ値も少ないことから，適切に提案方法が機能しているといえる．

図 6 に分離結果を示す．全体としては，TF_T^M が最も良い性能を示し，次いで提案手法による推定伝達関数セット， TF_T^G の順となっている．入力は 60 度方向からの音声雑音と拡散性の白色雑音のため，幾何計算伝達関数 TF_T^G による分離では SDR を改善できていない．分離アルゴリズムについては，fully-fixed 型固定，semi-fixed 型固定，適応型の 3 つのビームフォーミンググループすべてが，提案手法によって改善できていることがわかる．これにより，伝達関数を動的な関数として定義すること，またその適応方法が有効であることが示されたといえる．

図 7は，2 種類の適応方法を用いた場合の音源定位の様子を示したスナップショットである．図 7a）は，明示的な事前キャリブレーションを行う従来型手法を提案方法を用いて構成した場合の結果である．適応を行う前は，提案法と TF_T^G に差がないが，白色雑音 W_T を用いたキャリブレーションを行うことで，TF_T^M に近い結果

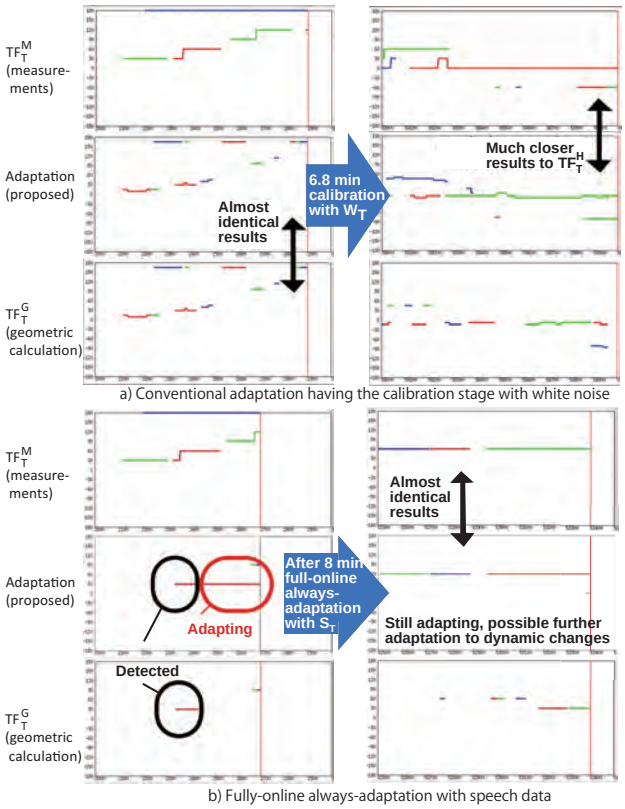


図 7: Two Online Demonstrations. The vertical and horizontal axes are the azimuth in degree and time frame in 10 ms of each subplot.

が得られるようになることがわかる．図 7b) は，提案手法による常時オンライン適応を行った場合の結果である．提案手法は，音源が定位される (黒丸) やいなや，適応処理が開始される (赤丸) ことがわる．このように明示的なキャリブレーションを行うことなく，TF_T^M に常時適応することができる．また，この後に，音響環境がさらに変化する場合も，提案手法は常時オンライン適応により，その変化に追従することができる．

6 おわりに

ロボット聴覚に代表されるように，実環境でのマイクロホンアレイ処理が求められる場面では，室内音響環境の変化に応じてマイクロホンと音源間の伝達関数を動的に適応する必要性があることから，本稿では，伝達関数の常時オンライン適応を報告した．提案手法をロボット聴覚オープンソースソフトウェア HARK 上に実装し，提案手法を用いたオンライン音源定位・分離システムを構築した．また，構築したシステムを用いた，実環境データ，およびオンラインデモによる評価を通じて，提案手法の有効性を示した．今後は，伝達関数更新アルゴリズムを拡張し，複数音源の扱いや，音声認識による評価を行う予定である．

謝辞

本研究は JSPS 科研費 JP19K12017, JP19KK0260 および JP20H00475 の助成を受けた

参考文献

[1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, “Active audition for humanoid,” in *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*. AAAI, 2000, pp. 832–839.

[2] V. Barroso and J. Moura, “Maximum likelihood beamforming in the presence of outliers,” in *IEEE ICASSP-91*, 1991, pp. 1409 – 1412.

[3] M. L. Seltzer, B. Raj, and R. Stern, “A bayesian framework for spectrographic mask estimation for missing feature speech recognition,” *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

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

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

[6] L. Griffiths and C. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Transactions on Antennas and Propagation*, vol. AP-30, no. 8, pp. 27–34, 1982.

[7] 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.

[8] A. Hiroe, “Solution of permutation problem in frequency domain ica, using multivariate probability density functions,” in *Independent Component Analysis and Blind Signal Separation*, J. Rosca, D. Erdogmus, J. C. Príncipe, and S. Haykin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 601–608.

[9] I. Lee, T. Kim, and T.-W. Lee, “Fast fixed-point independent vector analysis algorithms for convolutive blind source separation,” *Signal Processing*, vol. 87, no. 8, pp. 1859 – 1871, 2007, independent Component Analysis and Blind Source Separation.

[10] L. C. Parra and C. V. Alvino, “Geometric source separation: Merging convolutive source separation with geometric beamforming,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[11] M. Knaak and S. Araki, “Geometrically constrained independent component analysis,” *IEEE Transactions on Speech and Audio Processing*, vol. 15, no. 2, pp. 715–726, 2007.

[12] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, “Blind source separation with parameter-free adaptive step-size method for robot audition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1476–1485, 2010.

[13] K. Nakadai, S. Masaki, R. Kojima, O. Sugiyama, K. Itoyama, and K. Nishida, "Sound source localization based on von-mises-bernoulli deep neural network," in *2020 IEEE/SICE International Symposium on System Integration (SII)*, 2020, pp. 658–663.

[14] 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.

[15] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Sound event aware environmental sound segmentation with mask u-net," *Advanced Robotics*, vol. 34, no. 20, pp. 1280–1290, 2020.

[16] 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," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, 1995.

[17] J. J. E. G. B. Stan and D. Archambeau, "Comparison of different impulse response measurement technique," *Journal of the Audio Engineering Society*, vol. 50, pp. 249–262, 2002.

[18] S. Thrun, "Affine structure from sound," *Advances in Neural Information Processing Systems*, vol. 18, pp. 1353–1360, 2006.

[19] Y. Kuang and K. Astrom, "Stratified sensor network self-calibration from tdoa measurements," in *European Signal Processing Conference (EUSIPCO)*, 2013.

[20] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 161–164.

[21] 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.

[22] 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*, ser. Lecture Notes in Computer Science, H. Fujita, P. Fournier-Viger, M. Ali, and J. Sasaki, Eds., vol. 12144. Springer, 2020, pp. 614–625.

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

[24] K. Nakamura, S. Ambrose, and K. Nakadai, "On-the-spot calibration of microphone array transfer functions for robot audition," in *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*. IEEE, 2015, pp. 3354–3359.

[25] K. Nakamura and T. Mizumoto, "Blind spatial sound source clustering and activity detection using uncalibrated microphone array," in *25th European Signal Processing Conference, EUSIPCO 2017, Kos, Greece, August 28 - September 2, 2017*. IEEE, 2017, pp. 2438–2442.

[26] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. Am.*, vol. 69, no. 5, pp. 1484–1488, 1981.

[27] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[28] K. Nakadai, H. G. Okuno, and T. Mizumoto, "Development, deployment and applications of robot audition open source software HARK," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 16–25, 2017.

[29] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. of the Second International Conference on Language Resources and Evaluation (LREC'00)*. ELRA, 2000.

[30] H. Nakajima, G. Ince, K. Nakadai, and Y. Hasegawa, "An easily-configurable robot audition system using histogram-based recursive level estimation," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 958–963.

[31] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.