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-058-01

鳥類学におけるロボット技術や AI の関わり

Ornithology will meet Artificial Intelligence and Robot technology

森本　元

Gen Morimoto

山階鳥類研究所
Yamashina Institute for Ornithology

Abstract: 鳥類学とロボット技術や AI は，将来，現在以上に学際的研究が増加し，融合していくことは想像に難くない．近年，急速に発展を続けるロボット技術や AI は，私たちの生活に活用されつつある．他方，鳥類学をはじめとした野生動物研究では，AI については活用のための試行錯誤が始まったばかりであり両者の関わりは薄い．この状況を打破し両者が協同し発展するためには，鳥類学と工学の双方の相互理解が重要である．そこで本発表では，鳥類学の概要を整理しつつ，両者が融合しうる領域はどこなのか，どのような発展が考えられるのかについて紹介する．

1 はじめに ～AI やロボット技術と鳥類学～

目覚ましい発展を遂げている人工知能（Artificial Intelligence: 以下，AI）やロボット技術は，科学技術の専門的領域だけでなく，一般の人々の生活や企業活動といった域にまで飛躍的に普及してきており，その勢いと一般化は驚くべきものがある．近年，日々のニュース等でも耳にする自動車に関する自動運転や，人々が日々利用している翻訳サイトといったインターネット越しのいわゆるウェブサービス，他にも，教師データや大量のデータを活用した将棋ソフトや，医療における画像診断など，実例には事欠かない．こうした機械学習や深層学習に代表される AI 新技術は，さらに発展と普及の更なる加速が予想される．

それは，人々の生活だけでなく，学術領域においても同様であり，野生生物を対象とした研究分野においてもこうした技術の活用が普及し，分野の融合が進んでいくと予想される．他方，野生生物の研究分野はある意味，AI 技術の対局にあるような学術分野でもある（鳥類研究方法のいずれかの教科書[e.g. 1]を参照いただければ，詳細を知ることができるだろう）．なぜならば，観察者（測定者）の有する熟練の野鳥観察技術に依存した観測と測定が行われており，同時に，そうした特殊な技術を有する人材のマンパワーに観測結果が左右される側面をもつからである．これは見方によっては，職人技というアナログから学んだデジタル化と自動化という，AI やロボット技術とのコラボレーションの余地がまだまだ大きい学術分野ともいえよう．

こうした融合的な発展には，AI 分野の研究者と鳥学研究者の双方の相互理解が必要であろう．そこで本稿では，鳥類学とはどのような分野なのか，AI やロボット技術と融合しうる領域はどこであるかといった点を，非生物学分野の方々へ紹介することを目的とする．野生鳥類を対象とする学術分野の存在を御理解いただき，AI 分野とともに相互発展する一助になれば幸いである．

連絡先: 山階鳥類研究所
〒270-1145 千葉県我孫子市高野山 115
E-mail: morimoto@g.nifty.jp

2. 鳥類学とは

野生生物を対象とする研究はさまざまあるが，野生動物を対象とする研究者のアプローチ方法には学問分野別の視点からの研究アプローチと，対象生物視点からの研究アプローチがある．たとえば AI の研究者は，その AI 研究が，自動車や医療といった様々な他分野に活用できることを考えるだろう．他方，自動車の技術者は AI やエンジン技術，駆動系の研究といった分野が大きく異なる様々な領域を扱うことになる．これは鳥類学においても似た構造がある．

例えば「性選択」という進化の仕組みの一つとして異性間選択がある．メスが，より派手なオスを選

り好むことでオスの持つ形態的な装飾形質（派手であったり特徴的な形態である羽毛，体色など）が，遺伝的な背景によって次世代へ受け継がれる．この進化プロセスを通じて，オスはメスよりも派手な姿へと進化したというものである[鳥の色彩での事例: e.g. 2]．こうした研究は，鳥類に限らず，節足動物，魚類，両生類，爬虫類，哺乳類といったあらゆる動物が研究対象となる．鳥類学分野に限っても，特定の 1 種だけでなく，さまざまな鳥種が対象とできる．研究者は，自身の科学的興味に基づき，その研究テーマに適した最適な研究対象生物を探し，研究を遂行する．これはテーマ視点の研究アプローチである．

この例にあげた異性間選択という研究テーマでは，オスがメスと異なる派手な外見を獲得する進化の仕組み，その解明に焦点をあてている．検証し解明しようとする対象はそのメカニズム，つまり，進化理論である．専門的な表現で前述の内容を繰り返すなら，「メスがより派手なオスを選り好み繁殖することで，オスの色や飾り羽といった装飾形質がより明瞭になる方向へと淘汰圧が生じ，形態形質の発現を司る遺伝子を基盤として，年月をかけてその種のオスがより派手な外見へと進化する」ということだ．これが異性間選択という進化理論なのだが，この研究を実施するには研究対象を鳥類に絞る必要はなく，昆虫や哺乳類などでも構わない（そして実際，この進化的なメカニズムはさまざまな生物で多数の研究が行われている）．行動学を扱う学会や生態学を扱う学会は，こうした研究テーマをベースとした学協会といえよう．

他方，研究対象を基盤とする学問も存在する．鳥類学はまさにこの鳥類を研究材料（対象）としている学問である．その中身は，生態学，行動学，形態学，生理学，古生物学，保全生物学などの理系の学問分野だけに限らず，さらには鳥類を扱った文化人類学など多岐に渡る．極端な表現をすれば，鳥類を材料としていれば，どのような学術分野の研究でも可能といえる．それゆえ，自動的に分野横断的・学際的な学会となる．魚類の学会や昆虫といった節足動物の学会，哺乳類の学会などと同様に，鳥類の学協会は，多様な学問領域を含む研究材料系の学協会となる．このため，一言に「鳥類学」といっても，個々の研究者が扱っている研究テーマや領域はさまざまであり，それらは大きく異なっている．

3 鳥類学と工学（ロボット技術等）

鳥類学が生物学の一分野であるのと同様に，広い視点において AI やロボット技術が工学の一分野であることに誰も異論はないだろう．生物学と工学という学術的な枠組みが大きく異なるこれらがどのように関わるのかを考えると，その方向性を意識せざるを得ない．一つは鳥類学分野の知見を工学分野へ活用する方向性である．もう一つは，その逆であり，工学分野の知見や技術を鳥類学分野で活用するという方向性である．両者の関係性がこのように真逆の二つある点について，コラボレーションにおいては，その認識が重要となるだろう．

生物分野の知見を工学分野にて活用する事例の代表の一つは，バイオミメティクスといえよう．生物模倣とも呼ばれるこの学術分野は，生物のもつ特徴的な機能を参照するだけでなく，その機能を産みだすメカニズムを研究し活用することで新たな技術を生み出すものである．身近なものの例では，ある植物の種子のもつトゲの接着機能を真似た面ファスナー（いわゆるマジックテープ）が古くから有名である．近年では，ハスの葉の微細な表面構造による撥水機能を模倣して作成されたシートを用いて，内ブタにヨーグルトがくっつかない容器などが市販化されるなど，生物の機能を応用した工学技術はさまざま存在する（バイオミメティクスの事例に関しては複数の文献[e.g. 3,4]が出ているので参照されたい）．なお，どのように参照すればバイオミメティクスとするか，バイオインスピレーションとの違いといった，用語の定義の狭義・広義の議論はあるが，ここでは生物からヒントを得るものを含めて取り扱う．

バイオミメティクスと鳥類学との関わりは古く，もっとも著名なものは航空機だろう．かつて空を飛ぶ術を持たなかった人類は，飛行技術の開発にあたり，鳥類の飛翔を参考にした多くの研究を行ってきた．他の近年の事例として，日本国内における 500 系新幹線の形状もよく知られている．空中から勢いよく水中へ飛び込んで魚類などを捕食する鳥種であるカワセミの頭部形状は，低い抵抗で水中へ進入できる形状へと進化しており，この形状が，新幹線の空気抵抗を軽減する形状を産み出す際に参考にされた．他にも，鳥の羽毛を参考にして開発されたパターン形状によって，500 系新幹線のパンタグラフの騒音（風切り音）対策技術が開発されているし，鳥類の脚部を真似たロボットアーム[e.g. 5,6]や，鳥の歩行を参照した 2 足歩行ロボット[7]などの研究が行われている．鳥のように羽ばたいて飛行するロボットやドローンも開発，実用化されはじめている[e.g. 8]．静かに音を立てずに空中を飛び回ることができる点は，従来のプロペラやジェットエンジンによる飛行と異なる特徴である．他には羽毛の構造色発色メカニズムを参考にした発色材料の研究例[9]などもあり，鳥類を参照したバイオミメティクス研究は現在も発展を続けている．

こうしたコラボレーションは，工学が鳥類の特徴を拾い上げる（つまり鳥から工学）という方向性の側面が強いといえよう．もちろん，こうした研究においては相互にフィードバックがある．機能は知られていてもそのメカニズムが未知であった鳥類の特徴が，それを工学的に応用する研究を行う中で得られた成果が鳥類学へフィードバックされ，生物学的もその駆動メカニズムが判明することで，相互に発展する．異なる学術領域の学際的協同では，互いが対等な融合の側面もあるが，同時に，この例では研究の起点（シード）が鳥類の機能にあるという点もまた事実である．

なお近年では，循環型社会の構築のために，生態学の知見（生物の相互作用やネットワーク）を活用するといったエコミメティクス研究も進んできており，メカニクスといったハード面だけでなく，ソフト面でも，研究が進んでいる[10].

4 鳥類学と AI

他方，AI と鳥類学の関係はその逆の方向性が強いのではなかろうか．機械学習や深層学習に代表される AI 技術は，鳥類学が抱える様々な学術領域に対し，新たな視点と解決法を提供し，次の次元に鳥類学を飛躍的に発展させてくれるものと期待される．同時に，鳥類学の特性をキャッチアップすることにより，AI 研究へのフィードバックとともに相互発展するだろう．この相互理解のためには，鳥類学が内包する様々な学術領域の種類や特性への理解と，AI への理解が同時に必要になる．

そのために，両者の立ち位置をまず確認しておきたい．現状，鳥学者を含む野生生物学分野の研究者は，AI への理解が十分とはいえないだろう．また，工学者は野生生物学への理解が十分とはいえないと思われる．だが両者は全くの異分野ではなく，相互理解しやすい素養をもつ側面もある．野生生物のデータは，物理法則に従うものでないため，値が暴れ変動が大きい．言い換えれば，ノイズが非常に多いデータである．また，結果へ影響する要素が一つではないことや，生物と環境要因の相互関係からなる複雑な生態メカニズムを解き明かすことを目指してきた分野でもある．それゆえ，生物学者は統計学を重視してきた．検定にはじまり，統計的推定，複雑な統計モデル，さらに，近年ではベイズ推定も普及してきた．このバックグラウンドは，野生生物学者が AI を理解しやすい状況を自然と作り出していると思う．今後，着目するパラメーターや，教師データと解析対象のデータの関係などを把握しつつ，何をめざしているのかを認識しながら，両分野の研究者が協業することで，鳥類学をはじめとした野生生物研究は，AI 研究と大きく協業できると考えられる．

他方，工学者サイドには，鳥学者の扱う学術領域をさらに御理解いただけると幸いである．鳥類学は前述したように，多様な学術領域が含まれる．代表的なものをいくつか記す．

行動学や行動生態学は，前項でも触れた性選択研究や採食戦略，個体間闘争，渡りルートの研究等，さまざまな行動や形態要因に着目し，進化的な意義を研究する分野である．ここでは，なわばりの移動データや，個体間の争いの闘争行動のエソグラムデータ，さえずり行動での音声データや定位データなどが蓄積される．こうした研究では，観察者が野外（または飼育下）で対象の鳥類を肉眼や双眼鏡で観察し，調査用紙へ監察結果を記入していくことでデータを得ている．また，多様な機器も使用されている．数十年前はビデオカメラによる動画撮影やカセットテープによる音声録音などであったが，現在ではそれらは長時間の録画や録音が行われるようになっている．集められた映像データの解析には，かつては人力でひたすら画面を目視し続けるしかなく，撮影時間以上の多大な労力のかかる解析作業であり，現在でも苦労の多いたぐいのデータである．以前より，動きを検知するソフトウェアの活用[e.g.11]など，現在ではプログラマブルに解析しているケースも増えてきているとはいえ，研究者が個々にこうした工夫をして対応しているのが実情であろう．

音声についても，ただ耳で聞き直していた作業方法から発展し，現在では PC の高性能化に伴い，PC へ音声データを取り込み可視化しソナグラムを描く方法が一般化した．このような作業を行い，必要箇所を聞き出すといった方法を用いている研究者が多いと思われる（近年の野生動物の音声データの調査・解析については文献[12]に詳しい）．使用されるソフトウェアは録音方法については，野生生物の音声解析に特化したソフトウェアが存在し（実質的にほぼ鳥類用の要素が強い），Avisoft SASLab Pro[13], Raven Pro[14]が長らく多くの研究者に用いられている．これらのデータの解析には，パターンマッチングや機械学習などさまざまなアプローチが現在も試みられており（文献[15]に詳しい），人力と機械化の途上にあるといえよう．なお鳥類の音声録音については，古くはパラボラマイクにテープレコーダーや MD, DAT 等を接続し，対象個体を狙って録音することが一般的であった（文献[16]に詳しい）．近年では，IC レコーダーを設置し，長時間録音を行う研究が普及してきている[e.g. 17]．そうした録音データからの鳥種の自動識別や，他の音を含む音源からの対象種の抽出[e.g.18]の研究も進んでおり，大量に蓄積された

録音データをどう処理するかは，この分野の大きな課題となっている．

近年では音のデータの質が劇的に向上したことは鳥類の音声を用いた研究の大きなブレイクスルーとなる可能性を秘めているだろう．マルチマイクアレイによる立体的なデータ取得とそれを活用した研究[e.g. 19,20]は, 行動追跡研究の新たな時代の到来を予感させる．

移動データでは，大型の人工衛星発信機に始まった追跡研究が，今では小型の GPS ロガーやジオロケータなどによる小型の追跡機器の登場により，以前よりも様々な鳥種を対象と出来るようになると同時に，多量のデータが蓄積されるようになってきている（文献[21]に詳しい）．この分野の，急速に移動ログの多量のデータが増加し，急速に発展している分野といえよう．

生態学もまた鳥類学の中で大きな分野の一つである．生態学はたいへん幅広い学問ゆえ，説明することがなかなか難しい．ここでは一例として鳥類の個体数に関する話題を取り上げたい．全国各地の調査地点において, ボランタリーの鳥類調査員によって, 毎年，または数年ごとに調査が行われ，出現鳥種や個体数を記録するタイプのモニタリング調査が複数[e.g. 22,23]行われている．こうした調査では，長年の時系列，かつ，多地点の種構成や個体数データが蓄積されている．他にも，バードウォッチャーの観察結果を蓄積するタイプの調査が国内外で実施されており，一般の人々が，観察した鳥種を web サイト上で登録し続けている[e.g.24,25]．こうしたデータは，気候変動に伴う鳥の個体数変化や渡りの移動のタイミングの変化などを観測する研究などに活用されるなど，オープンデータ（またはそれに近い形）で運用されていることが多い. 気温といった環境要因や, 複数の生物種の捕食-被食関係，炭素循環などの物質循環など，さまざまな生物と要因がからみあった複雑な生態系に関する研究など，さまざまな生態学的研究が存在する．

形態学もまた, 鳥類学の代表的な研究分野である. 嘴や翼，骨格，羽毛の構造といった形態的特徴に着目し，その機能や特徴を明らかにすることなどを目的とする．こうした研究では，古典的にはノギス等で計測した測定データだったが，今では，形態的な三次元座標データや画像データなどが得られるようになっている. 例えば, X 線 CT による立体データ，羽毛の内部構造を調べた電子顕微鏡画像（SEM 等）の大量の画像データ[26]が蓄積されてきている．

分類学や系統学は鳥類の各種のグループ分けや進化的系統を明らかにする研究分野だが，こうした研究では，形態的な研究に加えて，この数十年の間に遺伝子データの活用が一般化した．アクセス可能なデータベースが様々あるが，代表例は世界中の（ほぼ）全鳥種だけでなく様々な生物の特定の領域の DNA 情報を収録した DNA バーコーディング[27]のプロジェクトなどが著名である．シーケンサーの発達により，その読み取り（解読）の速度は劇的に速くなっており，今では次世代シーケンサーを活用した網羅的なゲノム解析が普及してきている．バイオインフォマティクスという遺伝学と情報学の学際領域が生まれて久しいように，塩基配列の大量データの解析には，情報学的なアプローチが必要不可欠である．

このほかにも，絶滅の危機にある野生動物の保護などを扱う保全生物学や，文学作品の中でどのように鳥類が扱われているかなどの人類学的領域や人文学的領域での文理融合的な研究分野など，紹介しきれないがさまざまな学術領域が存在する．たとえば後者では，大量の文章データの自然言語処理などが関わりうるだろう．

5 むすび

近年，技術の発達に伴い，前述した様々なデータの蓄積が加速している．画像データ，音声データなどは，その膨大さに反して，それを人力で解析することには限界がある．解析よりも，データ蓄積の方がずっと先行する状況になってきているといえよう．さらに，マルチマイクロフォンアレイといった新しい試みが加わることで，従来は観察者が 1 名で移動しながら時間差で短時間だけ実施していた鳥個体の追跡観察（識別と定位の同時判断）を，多地点で同時に長期的に行えるようにもなっていくに違いない. こうしたデータの量と質の飛躍的な向上は，既存の鳥類学の知見を超えて，劇的に発展させる可能性を秘めるものである．そしてその分析には，AI は欠かせないもっとも重要な要素であろう. 鳥類学と AI 研究の学際的発展が益々加速することに期待したい.

謝辞

本発表の機会を与えていただいた, 鈴木麗璽博士, 中臺一博博士に御礼申し上げます．本発表の一部は JSPS 科研費 20H00475 からの助成をいただいた.

参考文献

[1] 山岸　哲（編著）『鳥類生態学入門』(1997, 築地書館 193pp)

[2] 森本　元. 鳥類の羽色と機能 ~羽毛の発色と生物学的

背景~. 色材協会誌 Vol. 89 No.6 p184-190, 2016.
[3] 下村政嗣. バイオミメティクスと表面技術. 表面技術 Vol. 64, No.1, p2-8, 2013.
[4] 篠原現人・野村周平（編著）『生物の形や能力を利用する学問バイオミメティクス』（2016, 東海大学出版部 151pp）
[5] Doyle, C.E., et al. Avian-inspired passive perching mechanism for robotic rotorcraft. In: 2011 IEEE/RSJ international conference on intelligent robots and systems. IEEE, p4975-4980, 2011.
[6] Thomas, J., Polin, J., Sreenath, K., & Kumar, V. Avian-inspired grasping for quadrotor micro UAVs. In: International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, p. V06AT07A014, 2013.
[7] Gong, Y., Hartley, R., Da, X., Hereid, A., Harib, O., Huang, J. K., & Grizzle, J. Feedback control of a cassie bipedal robot: Walking, standing, and riding a segway. In: 2019 American Control Conference (ACC). IEEE, p. 4559-4566, 2019.
[8] Festo (参照 2021-11-16) BionicSwift. https://www.festo.com/group/en/cms/13787.htm.
[9] Kawamura, A., Kohri, M., Morimoto, G., Nannichi, Y., Taniguchi, T., & Kishikawa, K. Full-color biomimetic photonic materials with iridescent and non-iridescent structural colors. Scientific reports, Vol. 6 No.1, p1-10, 2016.
[10] 下村政嗣. 持続可能な循環型社会を目指す生態系サービスとしてのエコミメティクス. In: バイオミメティクス・エコミメティクス—持続可能な循環型社会へ導く技術革新のヒント—. CMC 出版, p339-350, 2021.
[11] 植田睦之・田中啓太. 鳥の巣のビデオ録画の動体監視ソフトウェアによる自動解析. Bird Research, 2, T1-T7, 2006.
[12] 阿部聖哉. 音声データによる野生生物調査の研究動向. 環境アセスメント学会誌, 18(2), 3-9, 2020.
[13] Avisoft Bioacoustics. (2021-11-16 閲覧) http://www.avisoft.com.
[14] Cornell Lab of Ornithology. (2021-11-16 閲覧) https://ravensoundsoftware.com/software/raven-pro/.
[15] 阿部聖哉. 音声データによる野生生物調査の研究動向. 環境アセスメント学会誌, Vol.18, No.2. p3-9, 2020.
[16] 大庭照代. 鳥類音声録音の意義と方法. Strix, Vol.7, p35-82, 1988.
[17] 白井聰一. 針葉樹林ギャップ地を落葉広葉樹林に再生する過程における鳥相の変化: 録音によるデータの収集. 日本鳥学会誌, Vol.67 No.2, p227-235, 2018.
[18] 宇根健一郎, 藏屋英介, 野口健太郎, 神里志穂子, 金城道男, 長嶺隆, & 嘉手苅修. 環境音を含む音データからのヤンバルクイナの鳴き声検出の検討. 情報処理学会第 74 回全国大会講演論文集, p589-590, 2012.
[19] Sumitani, S., Suzuki, R., Matsubayashi, S., Arita, T., Nakadai, K., & Okuno, H. G.. Extracting the relationship between the spatial distribution and types of bird vocalizations using robot audition system HARK. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. p2485-2490, 2018.
[20] Suzuki, R., Sumitani, S., Naren, N., Matsubayashi, S., Arita, T., Nakadai, K., & Okuno, H. G. Field observations of ecoacoustic dynamics of a Japanese bush warbler using an open-source software for robot audition HARK. JEA, Vol.2, No.2 p1-11, 2018.
[21] 樋口広芳 (編)『鳥の渡り生態学』(2021, 東京大学出版会 330pp）
[22] 環境省 (2021-11-16 閲覧) モニタリングサイト 1000. http://www.biodic.go.jp/moni1000/.
[23] 鳥類繁殖分布調査 (2021-11-16 閲覧) https://bird-atlas.jp.
[24] Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. eBird: A citizen-based bird observation network in the biological sciences. Biological conservation, 2009, 142.10: 2282-2292, 2009.
[25] 環境省 (2021-11-16 閲覧) いきものログ. https://ikilog.biodic.go.jp.
[26] 山階鳥類研究所 (2021-11-16 閲覧) 標本データベース https://decochan.net.
[27] BOLD SYSTEMS (2021-11-16 閲覧) BARCODE OF LIFE DATA SYSTEM. http://www.boldsystems.org.

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-058-02

鳴禽類のメスのさえずりの役割の理解に向けた音源定位手法の活用に関する一検討

A study on the use of sound source localization techniques to understand roles of female songs in songbirds

古山諒[1*]　鈴木麗璽[1]　炭谷晋司[1]　有田隆也[1]
Ryo Furuyama[1]　Reiji Suzuki[1]　Shinji Sumitani[1]　Takaya Arita

[1] 名古屋大学，Nagoya University

Abstract: 鳥類の特定の分類にみられる，なわばり宣言やメスの誘引の役割がある鳴き声であるさえずりは，従来そのほとんどの種においてオスにのみ存在すると考えられてきたが，最近の研究ではメスにも多くの種で確認されており，祖先ではさえずりは雌雄ともに持っていた可能性も指摘されている．本研究は鳴禽類のメスのさえずりの役割の理解の音源定位手法の活用に基づく一検討として，オオルリのオスとメスのさえずりのプレイバックに対する野生のオス個体の応答をマイクロホンアレイを用いて観測・分析することを目的とする．具体的には，森林内のオオルリのオス１個体に対してオスとメスのさえずりを再生した．最初の試みとして，対象個体の反応として得られたさえずりの音響構造とレパートリー，移動パターンについて，HARKBird を用いて分析した．予備的実験・分析の結果からは，オス・メスのさえずりを再生した場合，両場合において対象個体がさえずりで応答し，前者の方が後者の場合より強い反応が生じうる可能性や，前者においてより頻繁に生じるさえずりは対象個体がスピーカ周辺を移動後に優先的に発せられる傾向がありうることなどが示唆された．より統制の取れた実験の元で本手法を活用することで，詳細な行動傾向の比較が可能であることが示された．

1　はじめに

さえずりと呼ばれる鳴き声を持つ種である鳴禽類は，さえずりを学習し生成する固有の脳核を持っており，ほかの鳥類に比べ鳴き声に豊富なレパートリーを持つ [3]．従来，さえずりはオスによるなわばり宣言やメスの誘引のための鳴き声のことを指し，「繁殖期のオスが発する長く複雑な鳴き声」と一般的に定義されてきた [2]．しかし，近年の研究 [5] では鳴禽類の７割を超える種にメスのさえずりが存在し，系統的に広く存在していることが指摘されている．また，系統樹の再構築に基づく分析によれば共通祖先はオス，メスのどちらもさえずりを持つことが一般的であり，現在メスがさえずりを持たない種は，進化の過程でさえずりが失われた可能性がある．

北半球の温帯地域では，メスのさえずりを持つ種は南半球の温帯地域や熱帯地域と比べ比較的少ない．この地域でメスのさえずりが失われた理由は，北半球の温帯地域に生息する鳥類によくみられる渡りという行動が関係していると考えられている [3]．渡り鳥では，オスがメスよりも先に繁殖地に到着し，短期間でなわばりを確立することが一般的なため，メスが年間を通して縄張り意識を持つ種に比べ，なわばり防衛に関する行動を行う必要性が低く，縄張り行動としてのメスのさえずりが失われていった可能性がある．また，かつてはさえずりが社会的結合の維持の役割を果たしており，留鳥と比べて渡り鳥はその必要性が低いためという指摘もある．

一方で，北半球の温帯に生息する種でもメスがさえずりを持つものもいる．その一つがオオルリ (*Cyanoptila cyanomelana*) である．日本国内ではメスがさえずりを持つ種の報告はまだ少なく，メスの鳴き声を収集するウェブサイト[1]では，このオオルリとサンコウチョウのみが登録されている，オオルリは，日本には夏鳥として南西諸島を除く九州から北海道までの全国各地に飛来し繁殖する渡り鳥である [11]．オオルリは，オスがメスより先に繁殖地に到着しなわばりを作るとされており，メスがさえずりを失いやすい条件に該当していながらも，現在までメスがさえずりを失っていない種である [10]．現在の観測では，メスのさえずりには巣の周囲での敵に対する警戒 [9] やヒナに注意を促す信号

*連絡先：名古屋大学大学院情報学研究科
〒 464-8601 愛知県名古屋市千種区不老町
E-mail:

[1] http://femalebirdsong.org/

[11] の役割があることが推測されており，メスがさえずる状況は限定的であるといえる．一方，メスの鳴き声が現在も消失せずに残る稀な形質であると考えると，実生態に関わらずメスのさえずりに対するオス個体の反応を調べることは従来とは異なるメスのさえずりの役割を探るきっかけになる可能性があると考えられる．

Riebel らはメスのさえずりの役割を理解するための課題として，1）メスのさえずりに関する記述の蓄積，2）雌雄の歌を比べる指標の開発，3）発生のメカニズムの理解，4）機能・役割の理解，5）雌雄両方のさえずりを考慮した系統比較解析を挙げている [6]．4）において，縄張り防衛，パートナーの誘引・識別・刺激，個体の質の宣伝，社会関係の仲介等の役割を検討するための実験手法を提示しており，特に社会関係の仲介の理解において，インタラクティブなプレイバック実験によるさえずりの特徴の分析や，マイクロホンアレイを用いた音源定位に基づく社会ネットワーク分析を提案している．これらの多様な役割の可能性を検討するにあたり，マイクロホンアレイに基づく詳細な行動観測が雌雄ともにさえずる状況での複雑なコミュニケーションの理解のための手立てとなることが期待されているといえる．

以上を踏まえ，本研究は鳴禽類のメスのさえずりの役割の理解の音源定位手法の活用に基づく一検討として，オオルリのオスとメスのさえずりのプレイバックにおける個体の応答をマイクロホンアレイを用いて観測・分析することを目的とする．オオルリは日本三鳴鳥の一種とされ，そのさえずりは美しく種固有の基本的特徴に加えて個体固有のレパートリーがあり，その頻度等により多様なパターンが存在する．メスのさえずりもオスと類似した特徴を持ち，オス程の多様さはないがレパートリーがあり美しい [11]．これらの点でも，マイクロホンアレイに基づく観測の利点を検討するのに適した種であると考えられる．

具体的には，Suzuki らのウグイスに対するプレイバック実験におけるマイクロホンアレイ活用の方法 [8] を踏襲し，森林内のオオルリのオス 1 個体に対してオスとメスのさえずりを繰り返し再生した．これは，オスのさえずりであればなわばりへの侵入個体の再現と考えられ，対象個体はなわばり防衛のため近隣でさえずることが期待される．一方メスのさえずりの場合は同様の状況でさえずりのみメスに差し替えた人工的な状況と考えられる．対象個体の反応の違いをマイクアレイの音源定位情報を活用して詳細に調べることで，メスのさえずりがどのように認識されうるかを詳細に検討可能と考えた．今回はその最初の試みとして，対象個体の反応として得られたさえずりの音響構造とレパートリー，音源到来方向に基づく移動パターンについて，ロボット聴覚オープンソースソフトウェア HARK[4] を利用した鳥類の鳴き声の観測のためのスクリプト集である HARKBird[7] を用いて分析した．

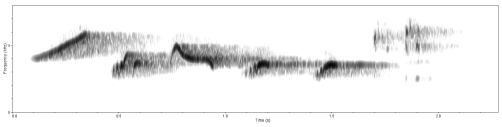


図 1: プレイバック実験で使用したオオルリのオスのさえずり

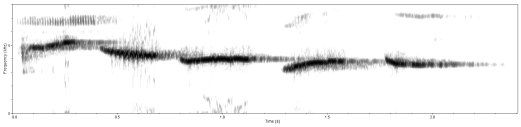


図 2: プレイバック実験で使用したオオルリのメスのさえずり

2 方法

2.1 実験

2021 年の 6 月 24 日に名古屋大学大学院生命農学研究科附属フィールド科学教育研究センター稲武フィールドにて現地のオオルリのオス 1 羽に対してプレイバック実験を行った．フィールドは主にスギ・ヒノキ・アカマツからなる針葉樹人工林であり，コナラ・シデ・カエデ等の小さな広葉樹パッチが点在する．実験当時はオオルリに加え，ウグイス，センダイムシクイ，コルリ等がさえずり，さらに，エゾハルゼミが鳴く状況であった．実験場所は開けた舗装の無い駐車場であり，周囲を針葉樹・広葉樹に囲まれている．駐車場の中心付近に 8 チャネルの USB マイクロホンアレイ（TAMAGO-03, System in frontier 社製）を複数台設置し，1 台のノート PC，または，Raspberry Pi を用いた録音ノード [13] に接続した．各マイクで同時に録音するのに加え，PC に Bluetooth で接続したポータブルスピーカを設置し，一定時間ごとにあらかじめ収録した鳴き声の録音を再生するプログラムを実行した．

実験では，現地を縄張りとするオスの個体に対して，前年に同地で録音されたオス（図 1），CD 鳴き声ガイド日本の野鳥 [12] に収録されたメスのさえずり (図 2) を 30 秒間隔で 15 分にわたって再生し，それに対するオスの個体のさえずりをマイクで収録した．ただし，再生機材の都合で一部音が鳴らない場合があり，再生回数には若干の違いがあった．実験ごとに 10 分以上の間隔を挟み，オス（M 条件），メス（F 条件）のさえずりのプレイバック実験をそれぞれ 3 回ずつ計 6 セッショ

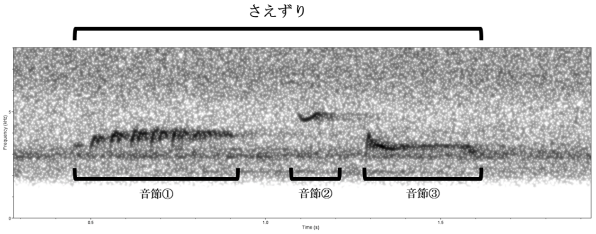


図 3: オオルリ（オス）のさえずりの例．さえずりは複数の音節で構成されており，この例では 3 つの音節で構成されている．

ン行った．なお，プレイバックの影響がない実験開始前の約 20 分の予備的な観測では，プレイバック時より対象個体のさえずりや移動がわずかであったことから，録音中の行動はプレイバックの影響を受けたものであると解釈して解析を行った．

2.2 解析

次の分析を行った．まず，2 つのプレイバックの条件において，対象個体の鳴き声の音響構造に違いがあるかを調べた．鳴き声の解析にはフィールド録音を対象とした生物音響解析のためのソフトウェアである Luscinia[2] で抽出，作成した各さえずりのスペクトログラムを用いた．図 3 はさえずりの一例であり，この例では 3 つの音節で構成されている．今回は，Luscinia を用いて取り出した各音節について，全時刻において最も強度が大きい周波数値（最大強度周波数）と，初期時刻と最終時刻における最も強度が大きい周波数値の差分（周波数帯域）を求めた．いずれもすべての音節の平均値をそのさえずり全体の指標とした．今回はオス，メスのさえずりのプレイバック実験で得られたさえずりのうち，20 個ずつを取り出し解析した．実験の各群の平均値の比較には F 検定を行い，等分散性が認められたため Student の t 検定を用いて比較を行った．有意水準は 5%未満とした．

また，各セッションにおける対象個体のさえずりパターンを，さえずりの種類と個体の移動に注目して解析した．各セッションにおいて，音声分析用ソフトウェアである Praat[1] を用いてオオルリのさえずり区間とその種類を手作業でアノテーションした．その結果，さえずりは 19 種類あることが判明した．次に，オオルリの鳴き声を含み，かつ，エゾハルゼミの鳴き声を含まないように周波数帯を制約（3700-7500Hz）して

[2]http://rflachlan.github.io/Luscinia/

表 1: 各セッションのさえずり頻度と移動頻度

セッション	M1	M2	M3	F1	F2	F3
さえずり頻度	124	165	173	124	75	70
平均		154.0			89.7	
移動頻度	9	31	17	24	3	3
平均		19.0			10.0	

HARKBird を用いて到来方向に関する音源定位を行った．HARKBird のアノテーションツール上に Praat での結果を読み込み，MUSIC スペクトルを参照しながら各鳴き声の到来方向を抽出した．今回手作業を中心にしたのは，セミ等の他種他個体の鳴き声が混在し，オオルリの鳴き声のみの情報を自動で取り出すことが容易でなかったこと，セッション数が限られているため，すべてを確認した確実なデータを取得することに注力したためである．

以上の結果を用いて，各セッションでのさえずり頻度とその到来方向が 5 度以上変化した回数，2 つの条件でのさえずりの種類の頻度分布，各セッションでのさえずり到来方向ごとの種類分布を比較した．

3 結果と考察

はじめに，2 種のプレイバック条件における対象個体への基本的な影響として，各セッション中のさえずり頻度と移動頻度を平均値とともに表 1 に示す．いずれの場合もプレイバック音に応じて対象個体の定常的なさえずりが観測されたが，オスのさえずり再生時（M 条件）の方がより頻繁にさえずり，かつ，頻繁に移動していることがわかる．これは，調査時間内において M 条件の方がプレイバックに誘引され周囲でさえずる時間がより長かったことによると考えられる．一方，メスのさえずりの再生時（F 条件）においては，開けた空間でメスの鳴き声が繰り返し聞こえる突飛な状況であるが，従来推測される警戒の役割の影響を受けた反応の結果であることや，オオルリのメスの歌はオスの歌とよく似ているために影響は弱いものの類似の反応を引き起こした可能性があると考えている．なお，プレイバックの影響がない実験開始前の約 20 分の予備的な観測では，74 回のさえずりをほぼ 1 か所でさえずっており，上記の結果がプレイバックへの応答であることを部分的に支持するといえる．

次に，音響構造の分析について述べる．図 4，5 はそれぞれさえずりの平均最高強度周波数，平均周波数帯域を示している．前者はいずれの場合も 3700Hz 程度でほぼ同様であり，オスとメスのプレイバック実験の間で有意な差はなかった（Student t(38), t= 0.62, p=0.54）．後者についても有意差は確認されなかった

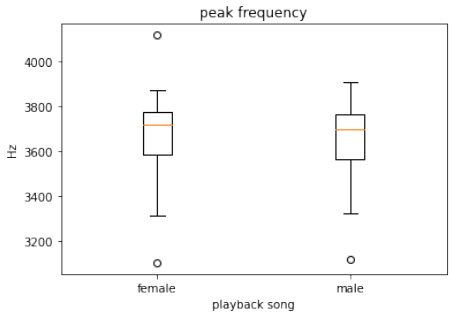


図 4: ２種のプレイバック条件におけるさえずりの最高強度周波数

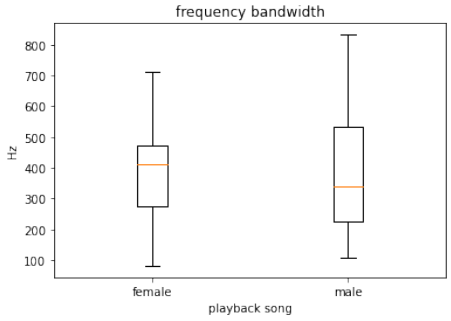


図 5: ２種のプレイバック条件におけるさえずりの周波数帯域

が（t= -0.41, p= 0.68），オスの場合は 320Hz，メスの場合は 410Hz 程度であり，メスのさえずりを再生した場合の方がさえずりの始めと終わりの間に若干幅が大きい結果となった．

より詳細なさえずりのパターンを分析するため，図 6 に２種のプレイバック条件におけるさえずりのレパートリー（種類）ごとの頻度を示す．条件に関わらずおおむね種類ごとに一定の割合で発せられているが，種類ごとに大きな差があることがわかる．より詳細には，1，6，7 番目のさえずりはそれぞれ頻繁であったが，M 条件においてより頻繁な傾向があった．また，14，15，16，18，19 番目は M 条件のみにおいて観測された．オスのさえずり再生の方がより頻繁な反応を引き起こしたことが，より多くのレパートリーを生じさせる要因になったことが考えられる．なお，前述の実験前の予備観測では，主要なさえずり頻度の多寡の傾向はプレイバック時と類似していたが，一部のレパートリーは他の種類と比べて相対的にプレイバック条件より頻繁，もしくはより少ない傾向が確認された．短時間の録音のため比較は難しいが，プレイバックの有無，つまり他個体の存在の有無によって使うレパートリーに違いがあること示唆しているかもしれない．

最後に，各セッションでのさえずりの種類ごとの方向頻度分布を図 7 に示す．同図から，両条件において

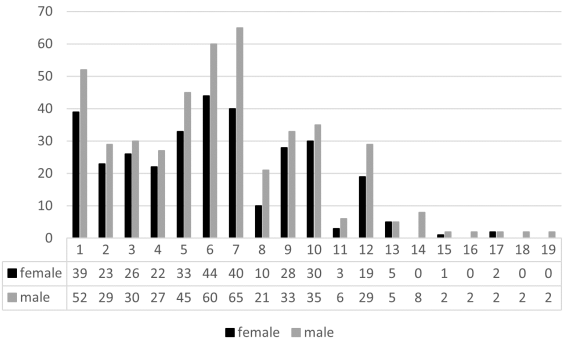


図 6: ２種のプレイバック条件とプレイバックの影響がない場合におけるさえずりのレパートリー別頻度

対象個体はいくつかの決まった方向のソングポストにおいてさえずっていたことがわかる．さえずりが頻繁であった方向では，主要な種類のさえずりが同程度の頻度で生じていることがわかる．一方，さえずり頻度が少なかった方向では，M 条件において頻度が高かった 1，6，7 番目のさえずりがよく観測された．これは，これらのさえずりが移動後に優先的に発せられるものであることを示唆していると考えられ，雌雄のさえずりの影響に関連して何らかの役割を持つことも考えられる．

4　おわりに

本稿では，鳴禽類のメスのさえずりが持つ役割を明らかにするための音源定位手法の活用の一検討として，オオルリのオスとメスの鳴き声を用いたプレイバック実験を行った．分析の結果からは，オス・メスのさえずりを再生した場合，両場合において対象個体がさえずりで応答し，前者の方が後者の場合より強い反応が生じうることが示唆された．メスさえずりへの弱い反応は，従来推測される警戒の役割の影響を受けた可能性や，オオルリのメスの歌はオスの歌とよく似ており類似の弱い反応を引き起こした可能性等が考えられる．また，オスのさえずり再生時により頻繁に観測されたさえずりの種類が，対象個体がスピーカ周辺を移動後に優先的に発せられる傾向があることも示され，何らかの機能を持つ可能性もありうることも示唆された．わずかな試行回数に基づく予備的な知見であるのに加え，個体の生態的状況の調査や通常時の個体の行動傾向との詳細な比較が必須であり，上記の結果の一般性やその意味するところはさらなる検討が必要であるのは明らかであるが，マイクロホンアレイを用いた観測により，役割が未知な要素の多いメスのさえずりに関わりうる詳細な個体の反応を抽出可能であることが示された．

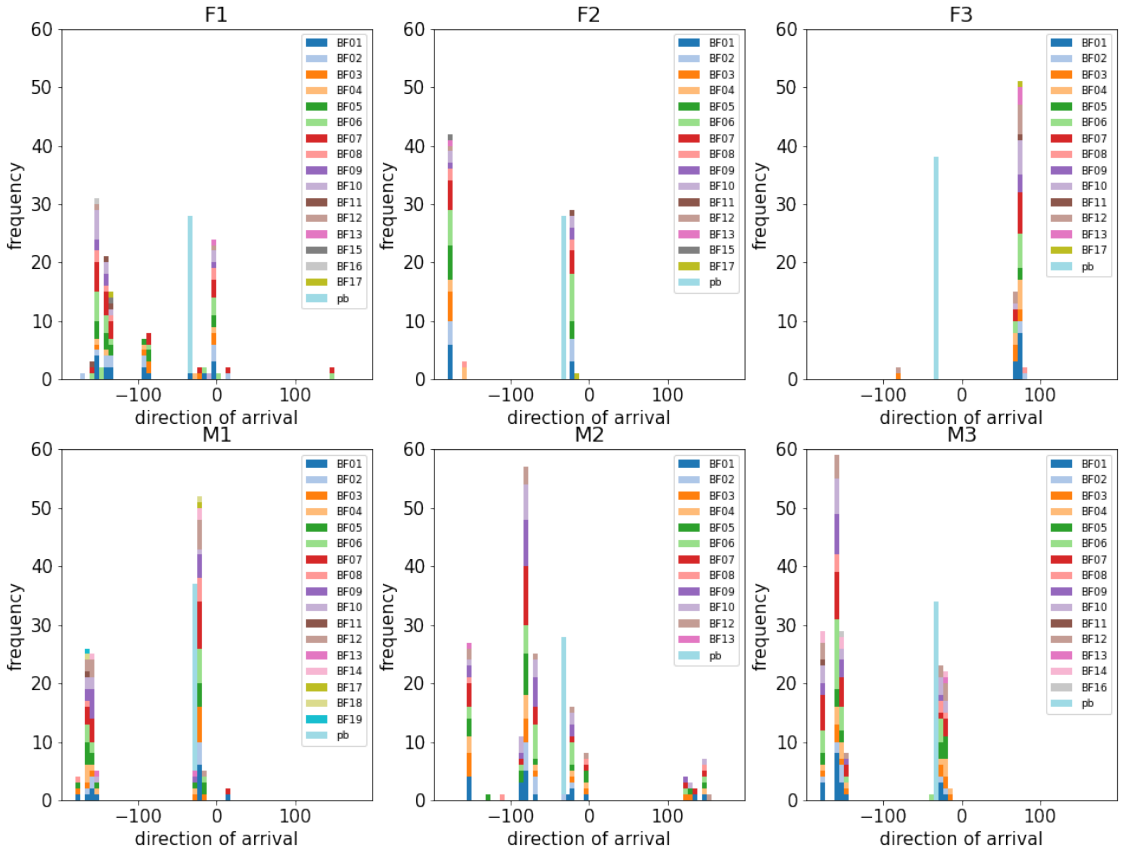


図 7: 各セッションでのさえずりの種類ごとの方向頻度分布（M*：オスさえずり再生，F*：メスさえずり再生．BF*：対象個体のさえずり種類，pb：プレイバック音）

今回は，オスとメスのさえずりを対象個体に再生しその反応をみて，メスのさえずりの機能を考察するという，これまでにあまり例のない探索的な実験であった．オスとメスのプレイバックにおいて反応の差がみられたが，実際の生態に近い状態で実験を行うことも重要である．例えば，実際の観察例のように実験個体の巣の付近でオスとメスのさえずりを再生し，比較するような実験である．鳴き声を出すタイミングや方向を抽出可能である HARK は，今回の実験のように複数の個体による相互作用を可視化し，移動頻度や移動方向ごとの鳴き声の種類の頻度を分析することで，従来の手法では困難だったオスとメスの社会的ネットワークの理解において助けになるはずである．今回の実験対象は一夫一妻の繁殖システムを持つオオルリだったが，一夫多妻などの一対以上の個体による繁殖をするものや，社会性を持つ種の実験において HARK を利用した実験はより有効なものとなる可能性がある．

謝辞

高部直紀氏（名古屋大学）の実験実施への協力，森本元氏（山科鳥類研究所）の助言に感謝する．本研究の一部は JSPS 科研費 JP21K12058，JP20H00475，JP19KK0260，JP17H06383（#4903）の助成を受けた．

参考文献

[1] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, pp. 341–345, 2001.

[2] C. K. Catchpole and P. J. B. Slater, *Bird Song: Biological Themes and Variations*. Cambridge University Press, 2008.

[3] N. E. Langmore, “Female birdsong,” *Current Biology*, vol. 30, pp. 789–790, 2014.

[4] 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. 27, pp. 16–25, 2017.

[5] K. J. Odom, M. L. Hall, K. Riebel, K. E. Omland, and N. E. Langmore, "Female song is widespread and ancestral in songbirds," *Nature Communications*, vol. 5, 2014, Art. no. 3379.

[6] K. Riebel, K. J. Odom, N. E. Langmore, and M. L. Hall, "New insights from female bird song: towards an integrated approach to studying male and female communication roles," *Biology Letters*, vol. 15, 2019, Art. no. 20190059.

[7] R. Suzuki, S. Matsubayashi, K. Nakadai, and H. G. Okuno, "HARKBird: Exploring acoustic interactions in bird communities using a microphone array," *Journal of Robotics and Mechatronics*, vol. 27, pp. 213–223, 2017.

[8] R. Suzuki, S. Sumitani, Naren, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno, "Field observations of ecoacoustic dynamics of a japanese bush warbler using an open-source software for robot audition hark," *Journal of Ecoacoustics*, vol. 2, Art. no. EYAJ46.

[9] 蒲谷鶴彦 and 松田道生, **日本野鳥大鑑**. 小学館, 2001.

[10] 徐敬善, "オオルリの繁殖生態と美しい構造色の羽," **野外鳥類学を楽しむ**, pp. 361–375, 2016.

[11] 徐敬善, "生態図鑑 オオルリ," *Bird Research News*, vol. 15, pp. 1–2, 2018.

[12] 松田道生, *CD* **鳴き声ガイド日本の野鳥**. 日本野鳥の会, 2016.

[13] 炭谷晋司, 大和祐介, 鈴木麗璽, 小島諒介, 有田隆也, 中臺一博, and 奥乃博, "野外での鳥類鳴き声観測のための web ベース録音ユニットと可視化ツールの試作," **第** *39* **回日本ロボット学会学術講演会予稿集**, 2021, Art. no. 2D4-03.

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-058-03

複数マイクアレイを用いたキンカチョウの時空間的発声パターンに基づく個体間相互作用の調査

A playback experiment on songbirds using simulated vocalizations based on a generative model

炭谷晋司[1*]　鈴木麗璽[1]　有田隆也[1]　和多和宏[2]　松林志保[3]　中臺一博[4,5]　奥乃博[6]
Shinji Sumitani[1]　Reiji Suzuki[1]　Takaya Arita[1]　Kazuhiro Wada[2]
Shiho Matsubayashi[3]　Kazuhiro Nakadai[3,4]　Hiroshi G. Okuno[5]

[1] 名古屋大学，Nagoya University
[2] 北海道大学, Hokkaido University
[3] 大阪大学, Osaka University
[4] 東京工業大学，Tokyo Institute of Technology
[5] （株）ホンダ・リサーチ・インスティチュート・ジャパン， Honda Research Institute Japan
[6] 京都大学，Kyoto University

Abstract: 本稿では，鳥類の音声コミュニケーションにおける個体間相互作用の観測・理解を目的として行ったマイクアレイを用いた音源定位実験について報告する．具体的には，空間的に限定された屋外環境としてテントを用い，その中にキンカチョウ（*Taeniopygia guttata*）を放ち，複数のマイクアレイを用いて録音を行い，ロボット聴覚オープンソースソフトウェア HARK で音源の定位・分離を行うことでキンカチョウが鳴き合う様子の観測を試みた．生成モデルの 1 つである VAE (Variational Autoencoder) と SVM（Support Vector Machine）を組み合わせた鳴き声の個体・種類の分類に関する取り組みと，分類結果と鳴き声の定位結果に基づく個体間相互作用に関する調査結果から確認された個体数の変化によって生じる歌・地鳴きの発声パターンの違いについて報告する．

1　はじめに

鳥類にとって，鳴き声は重要な意思伝達手段である．様々な発声を用いて他の個体と多様なコミュニケーションをとることが知られているが，その際に用いられる鳴き声の種類は大きく分けて「歌」と「地鳴き」の 2 つに分類される．歌は比較的長く複雑な発声で，縄張りの防衛や雌への求愛などに用いられ [1]，地鳴きは比較的短く単純な鳴き声で，捕食者への警告や社会的結合を形成するための信号など，より具体的な情報を交換するために用いられる [2]．鳥類の発声行動には，鳴くタイミングに関する個体間の時間的相互作用をはじめとして，個体ごとのなわばりの関係などの空間的相互作用，種間・種内それぞれに異なる歌の種類・周波数などの性質を持ち，互いに作用しあう音響的相互作用といった様々な次元での相互作用を持つ．

鳥類の鳴き声における相互作用の理解は，様々な鳥類で調査されているが，キンカチョウ（*Taeniopygia guttata*）も研究が盛んな種の 1 つである．キンカチョウは，歌の音声学習におけるモデル生物 [3] として調査されたり，地鳴きは集団内における社会的結合の形成や維持に関する調査対象 [4] として扱われるなど，鳴き声に関する研究事例は多くある．ここでキンカチョウの個体間相互作用に関する研究事例についていくつか例を挙げる．Gill らは，キンカチョウを対象として，ペアの形成から孵化までの繁殖期間における個体間相互作用について，実験個体に背負わせた小型のマイクを用いて録音行い調査した [5]．録音の分析結果からは，繁殖段階に応じてペア間で用いる地鳴きの種類が変化することが示唆された．また，Ikebuchi らは，ペアの形成に関してオス，メスを 2 羽ずつケージ内に入れ，録音とビデオ撮影による観測によって個体間の接近，発声の分析を行った [6]．その結果，オス個体は，ケージに入れられた直後に他のオス個体へ攻撃的な姿勢を取り，これによって集団内でのオスのヒエラルキーが形成されることが確認された．また，メスはオスの歌の質よりもこの社会的優位性によってオスを選好する傾向があることが示唆されている．

*連絡先： 名古屋大学大学院情報学研究科
〒 464-8601 愛知県名古屋市千種区不老町
E-mail:sumitani.shinji@h.mbox.nagoya-u.ac.jp

上記のように，キンカチョウの相互作用に関する知見は様々な手法によって多く報告されているが，多くの場合はケージ内といった実験室的環境で行われたものである．このような環境では，相互作用において重要となる空間的相互作用の抽出にも限界があり，本来は自然環境での空間的な制約を排除した観測が理想であるが，先行研究で用いられるようなビデオ撮影では木々などが障壁となって個体の追跡が困難である．また，マイクを個体に背負わせて行う録音手法では，空間的な相互作用の抽出ができず，個体への侵襲性も懸念される．

そこで，我々は従来の観測手法における課題を解決した観測手法の検討を目的として，マイクアレイを用いた手法によってキンカチョウ集団が鳴き合う様子を観測する取り組みを行ってきた．この手法は，市販のマイクアレイとロボット聴覚オープンソースソフトウェア HARK（Honda Research Institute Japan Audition for Robots with Kyoto University）[7] で構成される，複数の野鳥の発声行動のタイミングと方向，および，各鳴き声の音源が抽出可能な簡易システムである HARKBird[8, 9][1] を用いた観測手法である．本観測手法は，今回の取り組み以外にも鳥類集団における歌の時間的重複回避の観測 [8, 10] や，スピーカから同種の歌を再生するプレイバックが対象個体に与える影響の定量的観測 [11] などに適用されている．

これまでのキンカチョウの観測における取り組みとしては，屋外テント内に観測環境を構築し，複数のキンカチョウ（*Taeniopygia guttata*）を放つことで，集団で鳴き合う様子の録音調査を行っている [12]．この予備的調査では，複数マイクアレイを用いた高い精度での音源定位手法や，音源の類似度を利用した 2 次元定位手法の提案，歌・地鳴きの分類を試行しており，詳細な音源位置の定位・雑音の定位を低減できることや，低コストな手法で歌・地鳴きの区別ができること示し，マイクアレイによる非侵襲的な観測手法によってキンカチョウ集団が鳴き合う様子を詳細に観測できることが確認している．一方で，個体識別という個体間相互作用の観測には欠かせない課題が残されていた．

本稿では，マイクアレイを用いた上記研究の発展としてキンカチョウの個体・鳴き声分類に関する取り組みと，分類結果に基づく個体間相互作用に関する予備調査の結果について報告する．前者では，VAE (Variational Autoencoder)[13] と SVM (Support Vector Machine[14] を用いて実験個体の識別および歌・地鳴きの分類を検討した結果，比較的高い精度での識別を実現した．後者では，分類結果と音源定位結果に基づいて，テント内の個体数の違いに基づく歌と地鳴きの発声パターンの変化について調査した．キンカチョウは野生では集団で生活する鳥類であり，集団内でも個体間の関係はペア関係をはじめとして前述のオス個体の上下関係など，様々な社会的関係が個体間で存在することが考えられる．そのため，本研究では状況に応じた社会的関係の違いが観測できることが期待される．本稿では，オス 1 羽放鳥時，オス 2 羽放鳥時，オス 2 羽・メス 1 羽放鳥時の録音に着目して分析し，結果から条件によって発声の空間的分布や頻度に違いがあり，特にオス間の個体間関係に変化が生じることが観測された．

[1] https://sites.google.com/view/alcore-suzuki/home/harkbird

2 手法

2.1 実験環境

録音実験は，2020 年の 8 月から 9 月前半にかけて，北海道大学札幌キャンパス構内にある約 7m 四方のテント内で行った（図 1）．テント内には，止まり木や巣，エサ場を配置し，鳴き声の録音に使用するマイクアレイは，テント中央に 1 台，その周囲に 4 台配置した（図 2）．設置した止まり木や巣は，地上から約 1.5m ほどの高さに設置しており，マイクアレイも同じ高さに設置した．

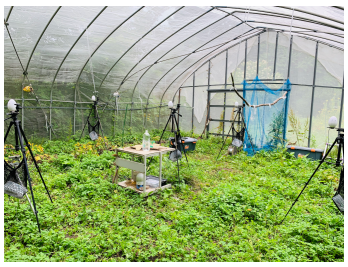


図 1: テント内環境．

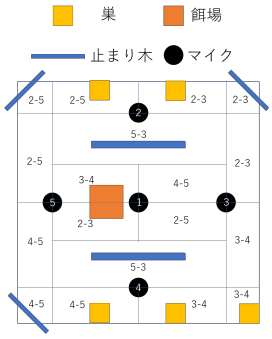


図 2: 配置図．

2.2 録音実験

録音実験は，1 個体のみ放鳥した条件，個体が放鳥されている環境に新たな個体を放鳥する条件の下で行った．各実験では，前日の 18 時以降に翌日の実験個体となるキンカチョウを実験環境に放鳥し，その翌日早朝 5 時から 18 時まで録音を行った．複数個体の実験では，9 時から 9 時 30 分の間に追加個体を放鳥し，13 時 30 分から 14 時の間に追加個体を再び捕獲した．実験は，異なるオスを 1 羽ずつ放った 2 条件（実験 1，実験 2）とオス 1 羽のいる環境に他のオス 1 羽を追加した条件（実験 3），オス 2 羽のいる環境にメス 1 羽を追加した条件（実験 4）で行った．実験に用いた個体は，実験室

のケージ内で飼育された個体であり，実験中以外はオスとメスを分けて別のケージで飼育した．オスとメスはそれぞれの発声は聞こえる状態ではあるが，視覚的に遮断された状態で飼育した．9 時 30 分から 13 時 30 分の 4 時間の録音に関して分析を行い，比較した結果について報告する．なお，本動物実験は，北海道大学動物実験委員会のガイドラインと承認（18-0053）を得て実施している．

2.3 2 次元定位

2 次元定位は，三角測量の要領で，2 台のそれぞれのマイクアレイから音源の定位方向へ直線を伸ばして交点ができた場合に定位位置とする方法で行うが，本実験では，先行研究結果 [12] に基づいて場所に応じて適したマイクアレイのペアを選定した上で定位行った．これは，音源の位置がペアのマイクアレイを結ぶ直線上に近いほど 2 次元定位は困難になることから，そのような位置での 2 次元定位を可能な限り避けることを実現している．図 2 の分割された場所ごと示されている数字のペアがその場所での 2 次元定位に利用したマイクアレイのペアである．2 次元定位を行う前に，HARKBird を用いてキンカチョウ個体の鳴き声をうまく定位するように定位のパラメータを適宜調整し，音源の定位・分離を行った．短時間フーリエ変換によって得られた各チャネルのスペクトログラムから MUSIC 法 [15] を用いて音源定位を行い，その定位結果に基づいて GHDSS 法 [16] を用いて対応する音源を抽出した．また，2 次元定位は以下の方法で行った．

1. 定位時間の重複するすべての各音源ペアに関して，HARKBird で抽出した音源定位情報を用いて三角測量に基づく 2 次元定位を行う．
2. 2 次元定位がなされた（交点が作られた）場合，交点ができた時間分だけ定位音源に対応する分離音を切り出す．
3. 切り出した音をグレースケール画像（128×128）に変換し，UMAP[17] を用いて 2 次元にデータの次元を削減，UMAP の特徴空間上で切り出した音源のペア間の距離が $d < 3$ である場合，その 2 次元定位情報を採用する．
4. 上記の処理を全てのマイクアレイのペアで行い，それぞれのペアが担当する定位範囲の 2 次元定位結果のみを抽出し，それらを結合する．これを最終的な 2 次元定位結果とする．

今回は，個体識別を別で行うため特徴空間上での音源ペア間の距離は大きく確保し，個体鳴き声の定位漏れが生じないようにした．また，採用する定位結果は中央のマイクアレイの位置を座標平面の原点として，縦軸，横軸でそれぞれ $[-4, 4]$ で限定した．

2.4 個体識別と歌・地鳴きの判別

複数個体を放鳥した場合の鳴き声および個体の識別は，VAE (Variational Autoencoder)[13] と SVM (Support Vector Machine[14] を用いる．判別の流れを図 3 に示す．まず，1 個体放鳥時の録音の定位結果を用いて，手作業で歌と地鳴きとそれ以外に分類する．今回，地鳴きは distance call と呼ばれる個体で比較的特徴のある鳴き声のみを分類した．distance call は，飛行時や警戒時，求愛など様々な場面で発することが報告されており，ペアとなったオスとメスが結合を維持するためのものだと考えられている [18]．また，ペア形成の前や直後で distance call を用いた相互作用が多いことも報告されており [5]，個体数条件の違いで発声に違いがみられることが期待される地鳴きである．分類したオス 3 羽，メス 3 羽の鳴き声の音源データをデータセットとして，VAE で学習する．各個体の鳴き声の分離音を時間幅 1.0 秒，シフト幅が地鳴きは 0.1 秒，歌は 0.2 秒で分割して複数の音声データを生成し，この音声データを Python ライブラリの matplotlib を用いて周波数 [1500,8000] の領域を 128×128 の画像データに変換する．VAE は，潜在変数の次元が 32 次元で，エンコーダ側が 8 層の畳み込み層と 3 層の全結合層からなり，デコーダ側がエンコードの逆関数となるモデルを機械学習ライブラリである PyTorch を用いて構築した．データセットの前処理として transform によるグレースケール化を施し，学習は，実験に用いたオス個体 3 羽の歌・地鳴き，メス個体 3 羽の地鳴きそれぞれ 3000 個の分割された画像データを用い，5000 回の学習を行った．

画像データの分類は，学習後の VAE に未知データを入力して得られた 32 次元の潜在変数を SVM (Support Vector Machine) に入力することで行う．SVM の学習・適用は Python ライブラリの scikit-learn を用いて行い，パラメータはデフォルト値を使用した．まず，分類を行う実験の録音から得た定位・分離音源に対して学習データと同様の音声分割と画像データへの変換処理を行い，生成した未知の画像データと録音実験で用いた個体の鳴き声の学習データとを共に VAE に適用し，それぞれの潜在変数を得る．次に，ラベルが既知である学習データのみを教師データとして SVM で学習し，学習した SVM モデルに残りの未知データを適用して分類結果を得る．各音源は画像データに変換する際に複数個のデータに分割されているため，それぞれの分類結果をもとに多数決をとり，もっとも多い分類結果を

音源の分類とした．また，未知データにはキンカチョウの鳴き声以外の音が多く含まれており，分類に大きく影響する．そのため，これらノイズの大部分を大まかに排除する方法として以下を行った．各未知データに対してその潜在変数からコサイン距離で直近の 50 個の学習データを取り出し，この距離の平均を未知データのノイズ度と定義する．ノイズ度が全未知データの平均値より大きい未知データをノイズとみなし，分類時の多数決には含めないようにした．また，分類結果は手直しによって誤ったラベルを正しいラベルになおし，これを最終的な結果とした．

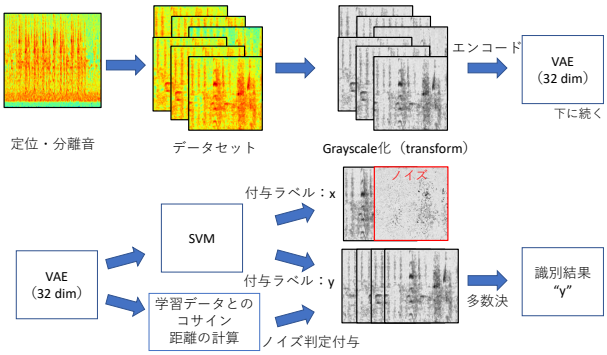
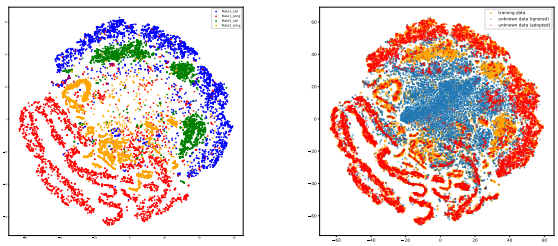


図 3: 鳴き声判別の流れ．

図 4 は，2 羽のオス個体の鳴き声を VAE に適用後,t-SNE[19] によって 2 次元に次元を削減して可視化した結果を示す．図 4 (a) からは各個体の歌・地鳴きがうまく分かれて分布していることがわかる．また，図 4 (b) では，中央付近に学習データでは確認できない分布が存在している．これらのほとんどが歌・地鳴き（distance call）以外の音であるが，提案手法により適切に除去されていることがわかる．



(a) 学習データ　　(b) ノイズ除去結果．

図 4: t-SNE による 2 次元での VAE 潜在空間の可視化．左図の各色は 2 羽のオス個体の歌と地鳴きを示す．右図の各色はそれぞれ橙：学習データ，赤：未知データ（採用），青：未知データ（除外）を示す．

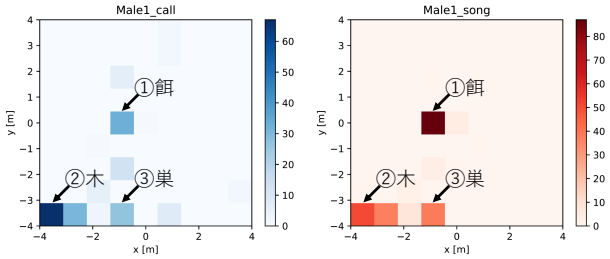
モデルの評価として，学習データには用いていない各ラベル 1000 個ずつのテストデータを SVM で分類を試みた結果，正しいラベルに分類されたものは最小で 861/1000 個（Female2 の地鳴き），最大で 963/1000 個（Male1 の地鳴き）であった．この分類結果に基づいて多数決を行うことを考慮すると，高い精度での個体識別，歌・地鳴きの判別が可能であるといえる．

3　分析結果

3.1　1 羽放鳥実験

図 5 はオス 1 羽（Male1）のみ放った場合の 2 次元定位結果をもとに各地点での歌・地鳴きの頻度分布を示したものである．この頻度分布は，手法で示した定位結果の採用範囲 $[-4, 4]$ を各軸で 10 分割している．歌と地鳴きの分布はある程度一致しているが，歌に関しては餌場（地点 1）で鳴く割合が高いことが確認できる．また，鳴き声の頻度の高い地点は餌場と図中左下の止まり木（地点 2）と巣（地点 3）の 3 か所に集中している．

図 6 は実験 1 とは異なるオス個体（Male2）を放った場合の 2 次元定位の頻度分布を示す．歌・地鳴きの分布は比較的類似しており，また歌の頻度が餌場（地点 1）で高い．これらの傾向は Male1 と一致する．また，鳴き声の頻度の高い場所は餌場と図中左下の止まり木（地点 2）の 2 か所に集中しており，これは Male1 と類似する．この結果から，オス個体のキンカチョウは 1 羽の場合，ある程度決まった場所で鳴き声を発する傾向があることが推測される．

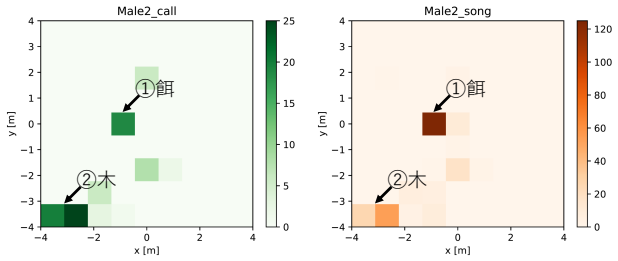


(a) 地鳴き（Male1）　　(b) 歌（Male1）

図 5: 鳴き声の頻度分布（実験 1）．

3.2　2 羽放鳥実験

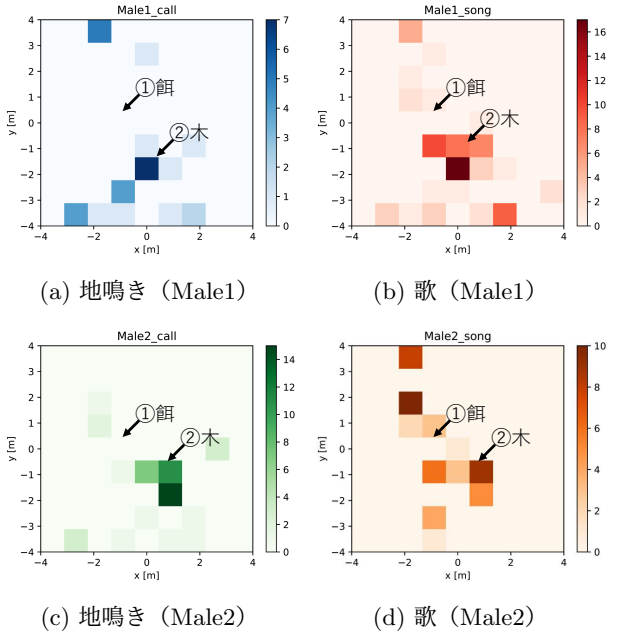
図 7 は，Male1 と Male2 の 2 羽をテント内に放鳥したときの 2 次元定位結果から得た鳴き声の頻度分布を示す．この結果からは，双方の個体で歌・地鳴きの分



(a) 地鳴き（Male2）　(b) 歌（Male2）

図 6: 鳴き声の頻度分布（実験 2）.

布が比較的一致していることが確認できる．これは，1 羽条件と同様の傾向にある．一方で，餌場（地点 1）で鳴く傾向は 2 羽ともに減少している．また，2 羽の分布は一方の個体のみが鳴いている地点もあるが，分布が重なる箇所もあり，特に図中中央下にある止まり木（地点 2）においてその傾向が確認できる．このことから，2 羽の個体は止まり木などの場所を共有しながら鳴いていることが示唆される．



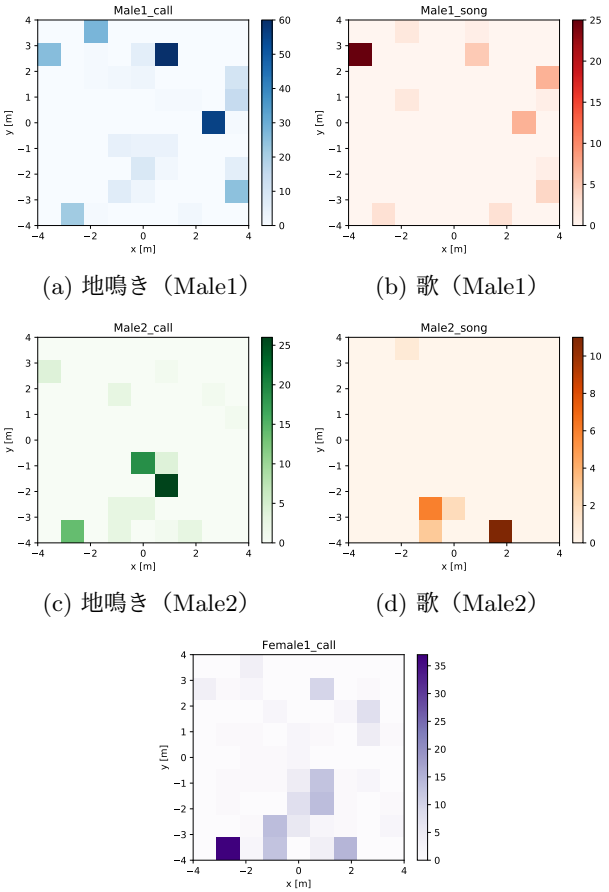
(a) 地鳴き（Male1）　(b) 歌（Male1）

(c) 地鳴き（Male2）　(d) 歌（Male2）

図 7: 鳴き声の頻度分布（実験 3）.

3.3　3 羽放鳥実験

図 8 は，オス 2 羽（Male1，Male2），メス 1 羽（Female1）をテント内に放鳥したときの 2 次元定位結果から得られた鳴き声の頻度分布を示す．オス 2 羽の条件



(a) 地鳴き（Male1）　(b) 歌（Male1）

(c) 地鳴き（Male2）　(d) 歌（Male2）

(e) 地鳴き（Female1）

図 8: 鳴き声の頻度分布（実験 4）.

と比較すると，Male1 の鳴き声の分布はテント全体に広がっており，Male2 の鳴き声はテントの図中下半分で多く確認できる．さらに，頻度の高い地点だけ注目してみると，Male1，Male2 でそれぞれ図中の上半分・下半分で明確に分かれており，メス個体の居ない 2 羽の条件とは傾向が大きく変化していることが確認できる．また，Female1 の地鳴きはテント全体に散らばっているが，概ね下半分に集中しており，これらの多くは Male2 の鳴き声の分布と重なっている．

3.4　発声頻度の比較

次に，条件の違いによるオス 2 羽の鳴き声の発声頻度の変化を調査した．表 1 は，各条件でのオス 2 羽の歌・地鳴きの定位回数をまとめたものである．それぞれの条件を比較してみると，地鳴きは個体で回数に差はあるものの，1 羽条件と 3 羽条件で多く，2 羽条件で少ない傾向があった．キンカチョウは，野生環境では群れで生活する動物であり，実験個体も複数個体と同

じケージで飼育されている．このことから，1 羽条件は普段と異なる状況であり，自身の存在の主張あるいは他個体への呼びかけとして多くの地鳴きを発声した状況であったことが推測される．また，2 羽条件の傾向に関しては 2 つの考察ができる．1 つは，2 羽条件は親しみのあるオス個体と一緒いる状況であり，1 羽の状況から脱して多くの呼びかけの必要が無くなったことにより地鳴きが減少したことである．もう 1 つは，distance call にはペア間での確認に用いられているという報告や，ペアの形成前や直後で多く用いられそれは繁殖段階が進むにつれ減少していく報告がある．オス同士においても同様の意味があると仮定すれば，結合の強いオス同士が時折地鳴きを用いて結合の確認を行っている状況だと考察できる．一方で，3 羽条件では Male1 で地鳴きが非常に多い．そこで，それぞれの個体の鳴き声に関して他個体の発声がどの程度重複しているか調査を行った．その結果，Male2 の歌に対する Male1 の地鳴きの重複が多くみられ，これは Male1 が Male2 の歌の発声に対して地鳴きによる阻害を行っているようにも見える．Distance call は様々な状況で用いられているとすれば，そのような利用方法も十分に考えられる．

一方で，歌は 1 羽条件で最も多く用いられ，2 羽条件，3 羽条件という順で回数は減る傾向があった．キンカチョウは，メスに向けた歌（directed song）とそれ以外の歌（undirected song）を区別して用いることが報告されている [20]．また自分の歌を維持するために，聴覚的なフィードバックが必要であることが報告されている [21]．以上のことから，1 羽条件で歌が多いのは，undirected song を用いた自身の歌の維持によるものであると考えられる．また，2 羽条件で 1 羽条件よりも歌の回数が減ったのは，聴覚的なフィードバックが必要という点で他個体が鳴いていないタイミングを計って発声するためであることが要因の一つだったと考えられる．最後に，3 羽条件での歌が最も頻度が少ない理由が 2 つ考えられる．1 つには，2 羽条件と同様の理由であり，メス個体の導入によって 2 羽条件よりも発声のタイミングが失われたということである．もう 1 つは，キンカチョウのペア形成においては，質の高い歌が好まれる傾向があることも報告されており [22]，発声回数を増やすよりも質の高い歌を歌うために自ずと回数が減ったということが考えられる．

3.5 発声と個体間距離の関係

2 次元定位の結果から，個体間で共通する発声分布があるなど個体の位置と発声に関係があることが示唆されたので，それらの関係についてより詳細な分析を行った．図 9 は，それぞれの個体の組み合わせについて，

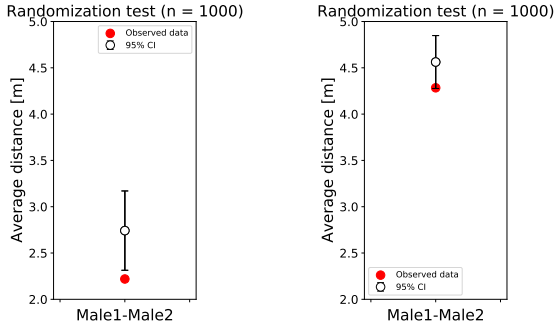
表 1: 各条件の Male1，Male2 の歌・地鳴きの定位回数.

		1 羽条件	2 羽条件	3 羽条件
Male1	地鳴き	203	29	302
	歌	240	79	61
Male2	地鳴き	90	48	86
	歌	269	52	23

一方が発声した場合にもう一方が前回の発声場所にいると仮定したとき，それぞれの定位位置間の距離を求めて平均を出したものである．赤点は観測値で，95 % 信頼区間は各個体に関して定位位置をランダムに入れ替えて平均距離を求めたランダマイゼーションテスト（$n = 1000$）の信頼区間である．Male1 と Male2 との距離間は，実験 3 で約 2.3m，実験 4 で約 4.3m と大きく異なっている．また，実験 3 では観測値が信頼区間より下回っている．この要因として，一方の鳴き声が他方を呼び寄せる役割があることが推測される．また，個体間の距離を結合の強さと仮定すれば，2 羽実験においてはオス個体同士は強い結合を持っていたが，メスの投入によってその結合が弱まったということが考えられる．実験 4 に関して，2 次元定位の頻度分布の結果からも確認できたように，Male1 は Male2，Female1 どちらとも距離を取って鳴いている傾向が確認できる．Male2 と Female1 は，観測値こそランダム化した場合の信頼区間内ではあるが，その他の組み合わせと比較すると Female2 との距離は短く，比較的近くで鳴き合っていることが示され，このペアは Male1 と Female2 のペアよりも結合の強い，親密な関係であることが示唆される．

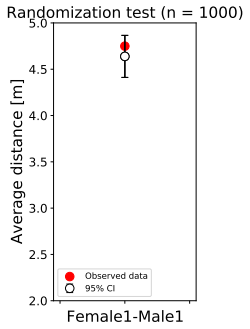
次に，オス個体の発声と個体間の距離に大きく差が見られたので，歌と地鳴きで違いがあるかを調査した．図 10 は，それぞれのオス個体に注目し歌と地鳴きそれぞれで鳴き声を発した場合の他方のオス個体との距離の関係を示したものである．歌，地鳴きともに実験 4 は実験 3 と比較して平均距離が大きくなっていることが確認できる．特に，歌はオス 2 羽の条件では近くでも多く歌われているが，メスを入れた実験では他方のオスの近くで歌うことが稀になっている．キンカチョウのオス個体は，ペアを形成する際に個体間のヒエラルキーを確立し，噛みつくなどの攻撃的な行動を行うことが報告されている [6]．これを考慮すると，観測された状況はメス個体をめぐって敵対し合う他のオスの近辺では容易に歌が歌えない状況であったことが 1 つの仮定として考えられる．

地鳴きに関しては平均距離は大きくなったものの，依然として近い距離でも用いていることが確認できる．特に，Male2 に関しては最も Male1 と近い距離で用いた地鳴きの頻度が最も高い．地鳴きが社会的結合の確認

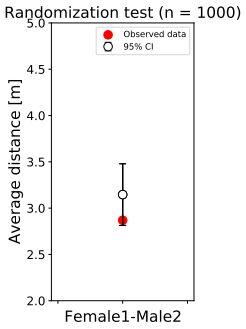


(a) Male1-Male2（実験 3）

(b) Male1-Male2（実験 4）



(c) Female1-Male1（実験 4）



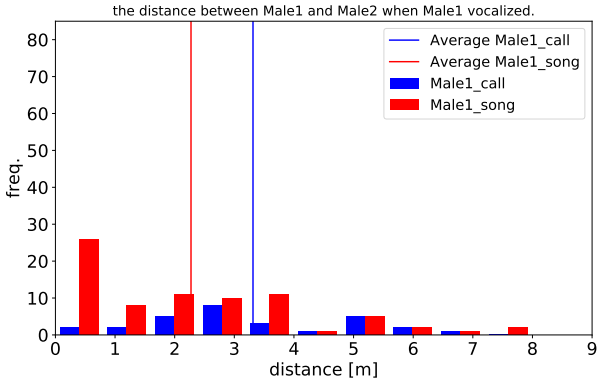
(d) Female1-Male2（実験 4）

図 9: 発声時の個体間の平均距離とランダマイゼーションテスト（$n = 1000$）による 95%信頼区間.

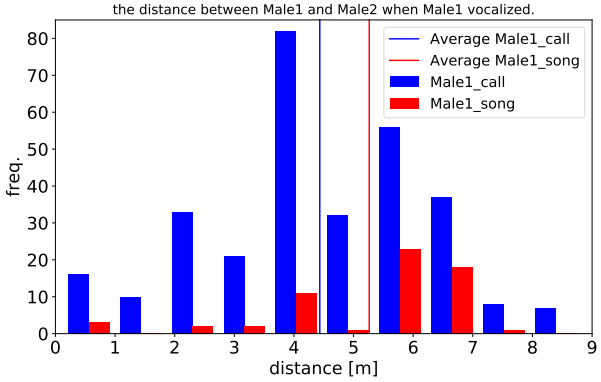
の際に用いられることを考慮すれば，敵対する状況においてもそれぞれの結合を確認し合うためにある程度の近距離においても地鳴きは発せられている状況ではないかと考えられる．あるいは，distance call は様々な用途があるということから，警戒の意味として用いたという可能性もある．一方で，この結果はオス間の関係を示したものであり，メス個体の状況が考慮されていない．そのため，これらの結果はメス個体の影響を受けた結果である可能性もある．

4　おわりに

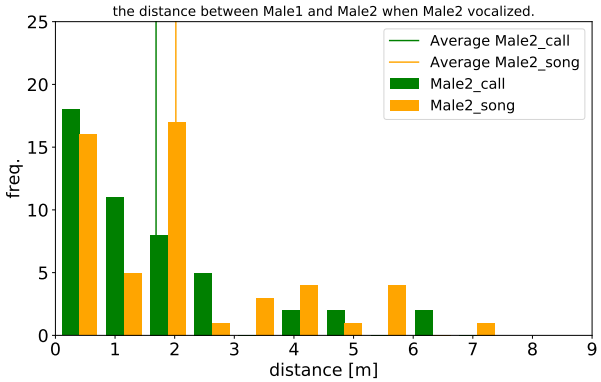
鳥類の鳴き声に基づく個体間相互作用の観測・理解のためにマイクアレイを用いてキンカチョウが鳴く様子を様々な条件で観測を試みた．提案した個体識別手法は，個体の鳴き声の特徴をうまく学習し，実際の分析にも適用ができた．また，分析結果は，状況に応じて挙動を変えている様子の観測に成功した．具体的には，キンカチョウ個体がオス 1 羽では比較的定位置で鳴く傾向があること，オス 2 羽では互いに止まり木などを共有しながら接近して鳴く傾向があること，メス 1 羽を入れるとオス個体が互いに距離を取りつつ鳴き合う傾向があり，歌と地鳴きで分布に差が生じること



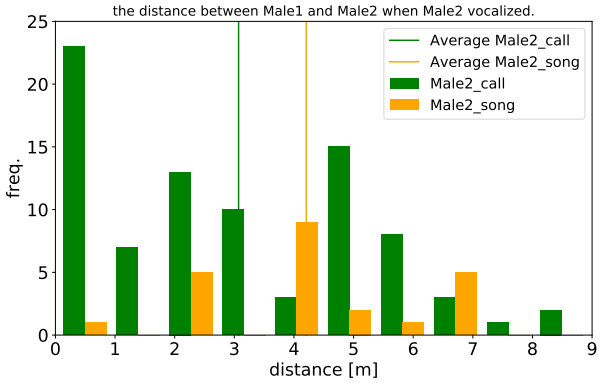
(a) Male1（実験 3）



(b) Male1（実験 4）



(c) Male2（実験 3）



(d) Male2（実験 4）

図 10: オス個体発声時の他方オス個体との距離の分布.

が確認できた．実際，実験を目視で確認した際に，2 羽条件では一方の移動後にはもう一方が続いて移動したり，3 羽条件では一方のオス個体が近づいた際にオス同士が威嚇をして鳴き合っている状況は多くあった．これは，ロボット聴覚技術によって実際の個体間で生じた相互作用が定量的に抽出できたことを意味する．

一方で，本実験は多くの課題が残る．例えば，地鳴きは distance call のみを抽出しており，その他の地鳴きに関する調査が不十分である．キンカチョウは，様々な地鳴きを介して社会的結合の確認を行うことが知られており [5, 23]，他の地鳴きのみで生じる相互作用の存在も考えられる．この意味で，今回の手法ではキンカチョウの個体間相互作用の理解には限界がある．歌に関しても directed song と undirected song の区別をしていない．条件の違いによる個体の挙動の変化をより詳細に調査するには，それぞれの条件でどちらの歌がどのように歌われたかを調査する必要がある．

今後の課題としては，上記の精査に加えて，個体数を増やしたより複雑な状況での観測，実際の自然環境における提案観測手法の適用可能性の検討が挙げられる．

謝辞

柴田ゆき野氏（北海道大学）の実験協力に謝意を表す．本研究の一部は JSPS 科研費 JP21K12058，JP20J13695，JP20H00475，JP19KK0260，JP17H06383（#4903）の助成を受けた．

参考文献

[1] C. K. Catchpole and P. J. B. Slater. *Bird Song: Biological Themes and Variations*. Cambridge University Press, 2008.

[2] P. Marler. Chapter 5 - bird calls: a cornucopia for communication. In Peter Marler and Hans Slabbekoorn, editors, *Nature's Music*, pp. 132 – 177. Academic Press, San Diego, 2004.

[3] L. A. Eales. Song learning in zebra finches: some effects of song model availability on what is learnt and when. *Animal Behaviour*, Vol. 33, No. 4, pp. 1293 – 1300, 1985.

[4] D. Stowell, L. Gill, and D. Clayton. Detailed temporal structure of communication networks in groups of songbirds. *Journal of The Royal Society Interface*, Vol. 13, No. 119, 2016. Art. no. 20160296.

[5] L. F. Gill, W. Goymann, A. Ter Maat, and M. Gahr. Patterns of call communication between group-housed zebra finches change during the breeding cycle. *eLife*, Vol. 4, , 2015. Art. no. e07770.

[6] Maki Ikebuchi and Kazuo Okanoya. Growth of pair bonding in zebra finches: physical and social factors. *Ornithological Science*, Vol. 5, pp. 65–75, 2006.

[7] 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. 27, pp. 16–25, 2017.

[8] R. Suzuki, S. Matsubayashi, K. Nakadai, and H. G. Okuno. HARKBird: Exploring acoustic interactions in bird communities using a microphone array. *Journal of Robotics and Mechatronics*, Vol. 27, pp. 213–223, 2017.

[9] S. Sumitani, R. Suzuki, S. Matsubayashi, K. Arita, T. Nakadai, and H. G. Okuno. An integrated framework for field recording, localization, classification and annotation of birdsongs using robot audition techniques - harkbird 2.0. In *Proceedings of ICASSP 2019*, pp. 8246–8250, 2019.

[10] R. Suzuki, S. Matsubayashi, F. Saito, T. Murate, T. Masuda, Y. Yamamoto, R. Kojima, K. Nakadai, and H. G. Okuno. A spatiotemporal analysis of acoustic interactions between great reed warblers (acrocephalus arundinaceus) using microphone arrays and robot audition software hark. *Ecology and Evolution*, Vol. 8, pp. 812–825, 2018.

[11] R. Suzuki, S. Sumitani, Naren , S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno. Field observations of ecoacoustic dynamics of a japanese bush warbler using an open-source software for robot audition hark. *Journal of Ecoacoustics*, Vol. 2, , 2018. Art. no. EYAJ46.

[12] Shinji Sumitani, Reiji Suzuki, Takaya Arita, Kazuhiro Nakadai, and Hiroshi Okuno. Non-invasive monitoring of the spatio-temporal dynamics of vocalizations among songbirds in a semi free-flight environment using robot audition techniques. *Birds*, Vol. 2, pp. 158–172, 2021.

[13] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint*, 2013. Art. no. arXiv:1312.6114.

[14] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, Vol. 20, pp. 273–297, 1995.

[15] R. Schmidt. Bayesian nonparametrics for microphone array processing. *IEEE Transactions on Antennas and Propagation (TAP)*, Vol. 34, No. 3, pp. 276–280, 1986.

[16] 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, pp. 1476–1485, 2010.

[17] L McInnes and J Healy. UMAP: Uniform manifold approximation and projection for dimension reduction. ArXiv e-prints 1802.03426, 2018.

[18] Richard Zann. Structural variation in the zebra finch distance call. *Zeitschrift für Tierpsychologie*, Vol. 66, No. 4, pp. 328–345, 1984.

[19] L. van der Maaten and G. Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2605, 2008.

[20] Richard A Zann. *The zebra finch: a synthesis of field and laboratory studies*, Vol. 5. Oxford University Press, 1996.

[21] Gerald E. Hough and Susan F. Volman. Short-term and long-term effects of vocal distortion on song maintenance in zebra finches. *Journal of Neuroscience*, Vol. 22, No. 3, pp. 1177–1186, 2002.

[22] Michelle L. Tomaszycki and Elizabeth Adkins-Regan. Experimental alteration of male song quality and output affects female mate choice and pair bond formation in zebra finches. *Animal Behaviour*, Vol. 70, No. 4, pp. 785–794, 2005.

[23] Pietro B.D'Amelio, Lisa Trost, Andries ter Maat. Vocal exchanges during pair formation and maintenance in the zebra finch (taeniopygia guttata). *Frontiers in Zoology*, Vol. 14, No. 13, 2017.

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-058-04

「間合い」としての共創コミュニケーション

Co-creative Communication as Subjective Synchrony

三宅 美博

Yoshihiro Miyake

東京工業大学情報理工学院情報工学系
Department of Computer Science,
School of Computing, Tokyo Institute of Technology
連絡先：226-8502 横浜市緑区 東京工業大学すずかけ台キャンパス
miyake@c.titech.ac.jp

Abstract: The co-creation system is a system that captures human communication from the inside. In this talk, I will focus on "Ma" as subjective time, and introduce the mechanism of its generation and interpersonal sharing. Specifically, based on the modeling of mutual entrainment of rhythmic movements using a cooperative tapping task, we show the effectiveness of human-artifacts interaction, especially walking rhythm, for rehabilitation support. This is the first step toward a new system theory that supports rhythmic movements and their synchronization from inside of humans.

概要

共創システムとはコミュニケーションを人間の内側から捉えるシステムである．本講演では，主観的時間としての「間(ま)」に注目し，その生成とインターパーソナルな共有の仕組みについて紹介する．協調タッピングを用いたリズム運動の相互引き込みのモデル化を踏まえ，人間と人工物のインタラクション，特に歩行リズムのリハビリテーション支援への有効性を示す．これはリズム運動とその同調を人間の内側から支援する新しいシステム論に向けての第一歩である．

われわれは、この間合いの共創を人間とロボットの間に再構成し，「間(ま)」の合う歩行リハビリ支援を実現してきた．人間の歩行リズムとロボットの歩行リズムの間合いを共創する人間・機械系である．「間」の合うトルクを人間に入力することで，人間が本来持つ歩行リズムを活かしたアシストが可能になる．これまで多くのパーキンソン病患者に適用されてきたが，歩行の賦活化だけでなく，すくみ足の改善など，間合いで初めて実現される臨床的な有効性が明らかにされている[1-6]．（https://walkmate.jp）

このような間合いの応用はリハビリ支援ロボットに留まらず，人間とサイバー空間の接続や人々を繋ぐメディアなど, 様々な領域への展開が考えられる．人間と人工物の共創コミュニケーションには大きい技術的な可能性が秘められている．

参考文献

[1] Miyake, Y., "Interpersonal synchronization of body motion and the Walk-Mate walking support robot," IEEE Transactions on Robotics 25, 638-644 (2009)

[2] Hove, M.J., Suzuki, K., Uchitomi, H., Orimo, S., Miyake, Y., "Interactive rhythmic auditory stimulation reinstates natural 1/f timing in gait of Parkinson's patients," PLoS ONE 7, e32600 (2012)

[3] Uchitomi, H., Ota L., Ogawa K., Orimo S., Miyake Y., "Interactive rhythmic cue facilitates gait relearning in patients with Parkinson's disease," PLoS ONE 8, e72176 (2013)

[4] Uchitomi, H., Ogawa, K., Suzuki, K., Nishi, T., Orimo, S., Wada, Y., Miyake, Y., "Effect of interpersonal interaction on festinating gait in rehabilitation for Parkinson's disease," PLoS ONE 11, e0155540 (2016)

[5] Yap, M.S.R., Ogawa, K., Nagashima, T., Hirobe, Y., Seki, M., Nakayama, M., Ichiryu, K., Miyake, Y., "Gait-assist wearable robot using interactive rhythmic stimulation to the upper limbs," Frontiers in Robotics and AI 6, 00025 (2019)

[6] Kishi, T., Ogata, T., Ora, H., Shigeyama, R., Nakayama, M., Seki, M., Orimo, S., Miyake, Y., "Synchronized tactile stimulation on upper limbs using a wearable robot for gait assistance in patients with Parkinson's disease," Frontiers in Robotics and AI 27, 010 (2020)

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-058-05

複数マイクロホンアレイを用いた NMFによる空間音源分離法の残響下での評価

Evaluation of spatial source separation using NMF with multiple microphone arrays under reverberation

鍵本泰宏[1*]　糸山克寿[1]　西田健次[1]　中臺一博[1,2]
Yasuhiro Kagimoto[1]　Katsutoshi Itoyama[1]　Kenji Nishida[1]　Kazuhiro Nakadai[1,2]

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

Abstract: 本稿では，複数台のマイクロホンアレイを用いた空間音源分離法について述べる．音源分離は，様々な音源やノイズが混在する中から所望の音源だけを抽出する技術である．音源分離の代表的な手法の一つであるビームフォーミングは，マイクロホンアレイと呼ばれる多チャンネルデバイスで収録した信号から，チャンネル間に生じる位相差に基づいて方向ごとに音源を分離することができる．しかし，ビームフォーミングは方向に基づく手法であり，同方向に複数の音源が存在する場合それらを分離することができないという課題があった．そこで，提案手法は複数台のマイクロホンアレイを用いた，音源の位置に基づく分離を行う．目的音源に対して複数マイクロホンアレイでビームフォーミングし，得られた分離音から非負値行列因子分解（Non-negative Matrix Factorization）によって目的音源だけを抽出することで，同方向に存在する別音源の影響を緩和する．提案手法は，シミュレーションにより遅延和法に比べSDR（Source to Distortion Ratio）がおよそ0.8〜2.3dB向上することがわかった．また，実環境での分離性能を評価し，残響の影響について検証した．

1　はじめに

近年，スマートホンやAIスピーカーの普及に伴い，様々な場面で音響処理技術が利用されるようになってきた．例えば，Apple社のSiriに話しかければ，音声による機器操作や検索，文字起こし等を行える．このようなアプリケーションを使う際に問題となるのが雑音である．実環境では他の音源やノイズが混在しているため，処理精度が低下してしまう．そのため，様々な音源の混合音から目的音源だけを分離する音源分離技術は，音響処理を行う上で重要性を増してきている．音源分離の代表的な手法として，ビームフォーミング（Beamforming，BF）が挙げられる．ビームフォーミングではマイクロホンアレイという多チャンネルデバイスで用い，収録信号のチャンネル間位相差に基づいて音源方向に指向領域を形成する．これにより各音源方向に対してビームフォーミングを適用することで各方向の分離音を得ることができる．しかし，ビームフォーミングは方向に基づく手法であり，同方向に複数の音源が存在する場合，それらを分離することができない．

このような課題に対処するため，本研究では複数のマイクロホンアレイを用いた位置に基づく音源分離手法を提案する．この手法は，音源を方向ではなく，位置に基づいて分離することにより同方向音源の影響を低減する．提案手法の処理は大きく2つのステップで構成されている．まず，図1のように，どのマイクロホンアレイから見ても目的音源S_0方向に別音源が存在するような状況を想定する．一つ目のステップでは，マイクロホンアレイを複数箇所に配置し，それぞれの位置から目的音源方向に対してBFを行う．この時，各マイクロホンアレイの指向領域が重なる部分はスポットと呼ばれ，スポット領域内に存在する音源を抽出することをスポットフォーミング（Spotforming，SF）と呼ぶ．各マイクロホンアレイでBFして得られた目的音源方向の分離音（以下，BF分離音と記す）には，目的音源と目的音源方向に存在する他音源が支配的になって分離される．目的音源は全てのマイクアレイで分離されるのに対し，その他の音源は一部のマイクアレイでしか分離されないと考えられる．そこで，2つ目のス

*連絡先：東京工業大学工学院システム制御系
〒 152-8552 東京都目黒区大岡山 2-12-1 W8W-W310
E-mail: kagimoto@ra.sc.e.titech.ac.jp

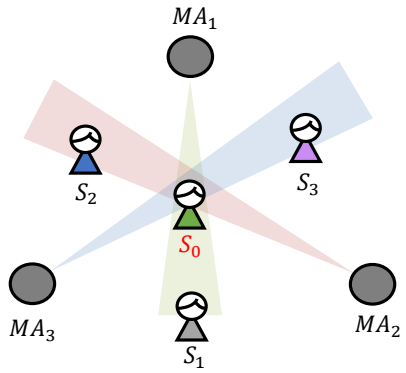


図 1: スポットフォーミングのイメージ．各マイクロホンアレイでビームフォーミングしたときに指向方向が重なる部分をスポット領域とし，スポット内の音源だけを分離する．

テップでは，ブラインド音源分離などで用いられる非負値行列因子分解（Non-negative Matrix Factorization，NMF）[1] を利用して，BF 分離音に共通する音響成分を取り出すことで目的音源を抽出する．ここで，NMF は入力スペクトログラムを構成する代表的な周波数スペクトルを学習し，低ランク近似する機械学習手法の一つである．各処理の詳細については第 3 節で説明する．

本稿では，提案手法について 4 人の音源が同時発話した場合を想定し，提案手法の性能を評価した．シミュレーション環境下で遅延和法に比べ SDR（Source to Distortion Ratio）がおよそ 0.8〜2.3dB 向上することが分かった．また，実環境下での残響の影響についての検証を第 4 節で行い，まとめと今後の課題について第 5 節で述べた．

2　関連研究

前述のとおり，従来のビームフォーミングは同方向に別の音源が存在する場合，それらを分離することが難しいという課題があった．この問題に対処するために，いくつかの手法が提案されている．

関口ら [2] は，複数マイクロホンアレイの配置最適化によるアプローチを提案している．マイクロホンアレイをロボットに搭載し，音源方向が重ならない位置に移動させることで，効率的に音源分離を行うことができる．ただし，このアプローチはロボットが移動するための空間的制約が大きく，使用できる状況が限定的である．

他のアプローチとして，スポットフォーミング法が提案されている．ビームフォーミングは特定方向に存在する音源を強調するのに対し，スポットフォーミングはスポットと呼ばれる領域を音源位置に形成し，スポット内の音源を抽出する手法である．鈴木ら [3] は，二本の超指向性マイクロホンの指向領域が重なる部分をスポットと捉え，収録信号を足し合わせることでスポット内の音源を強調する手法を提案している．指向性マイクロホンは動作が軽量で計算コストが低いという利点がある一方，指向性が一方向に固定されており，スポット位置を変えるためにマイクロホンを動かす必要があるという課題がある．Taseska ら [4] は，超指向性マイクロホンではなく，複数台のマイクロホンアレイを用いた手法を提案している．マイクロホンアレイは信号処理的に指向性を形成するため，任意位置の音を強調することができる．この手法は，分散させたマイクロホンアレイ全体を一つのマイクロホンアレイとみなし，スポット内の音源を強調する空間フィルタの設計を行う．しかし，マイクロホンアレイ間での厳密な同期や，マイクロホンアレイの配置を把握していることが前提となっており，実環境でこのような大規模なシステムを構築する場合配線コストが高くなるという課題がある．

これらを踏まえ，本研究では（1）固定化された複数台のマイクロホンアレイで運用可能である，(2）マイクロホンアレイ間で厳密に同期がされていなくても使える，という二点を重視した手法の構築を目指す．

我々は以前，複数台のマイクロホンアレイでビームフォーミングして得られた分離音のスペクトログラムを，k-means でクラスタリングすることでスポット内音源の分離を行えないか検討した [5]．この手法は，チャンネル間の位相差を用いる信号処理的な工程は各マイクロホンアレイ内で留め，マイクロホンアレイ間の処理は周波数成分のクラスタリングで行っている，これにより，マイクロホンアレイ間の厳密な同期なしに分離を行うことができる．しかし，この手法は空間的に音源を分離できる反面，クラスタリングを行う際に各フレームのスペクトルを一つのデータとして固定化してしまうため，複数音源が同時に発話する場合に分離ができないという欠点がある．そこで我々は，ブラインド音源分離などで用いられる NMF（Non-negative Matrix Factorization）を用いた手法を提案した [6]．NMF は入力される振幅スペクトルを複数音源の混合モデルとして仮定するため，時間的に音源が重なった場合でも分離が可能である．また，本手法ではマイクロホンアレイ間でフレーム単位の緩い同期が必要であるが，位相差などを使った処理は各マイクロホンアレイのビームフォーミング処理に留めているため，サンプル単位の同期を行わなくてもよいというメリットがある．提案手法の詳細については，次の第 3 節で述べる．

3 手法

本節では，提案手法の処理について説明する．全体の処理の流れは図 2 のようになっており，(1) 複数台のマイクロホンアレイによるビームフォーミング，(2) NMF による共通成分抽出，の 2 部で構成されている．以下では，それぞれの工程について説明する．

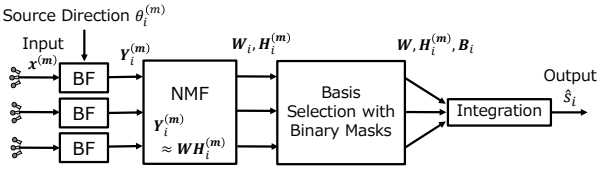


図 2: 提案手法の全体フロー図

3.1 複数台のマイクロホンアレイによる BF

まず，図 1 のような状況を考える．マイクロホンアレイは M 個存在し，各マイクロホンアレイでビームフォーミングを行う．m 番目のマイクロホンアレイで収録された信号を $\boldsymbol{x}^{(m)}(t) = [x_1^{(m)}(t), \ldots, x_N^{(m)}(t)]^\top \in \mathbb{R}^N, t = 1, \ldots, T$ とする．N, T はそれぞれチャンネル数，サンプル数を表している．収録信号に対して短時間フーリエ変換（short-time Fourier transform，STFT）を適用すると，$\boldsymbol{X}^{(m)}(f,\tau) \in \mathbb{C}^N$, $f = 1, \ldots, F$, $\tau = 1, \ldots, T_f$ に変換される．ここで，F および T_f はそれぞれ周波数ビン数とフレーム数である．また，マイクロホンアレイ m からみた目的音源 S_i の方向 $\theta_i^{(m)}$ に対してビームフォーミングすることで得られた分離音（以下 BF 分離音と記述）を $Y_i^{(m)}(f,\tau) \in \mathbb{C}$ とすると，$Y_i^{(m)}(f,\tau)$ は線形フィルタ $\boldsymbol{W}_i^{(m)}(f,\tau) \in \mathbb{C}^{1\times N}$ を用いて次のように得ることができる．

$$Y_i^{(m)}(f,\tau) = \boldsymbol{W}_i^{(m)}(f,\tau)\boldsymbol{X}^{(m)}(f,\tau) \tag{1}$$

ビームフォーミングではこの $\boldsymbol{W}_i^{(m)}(f,\tau)$ を推定することで目的音源方向だけを強調することができる．提案法では，ロボット聴覚用オープンソースソフトウェア HARK（Honda Research Institute Japan Audition for Robots with Kyoto University）[9] に含まれる GHDSS（Geometric High-order Dicorrelation-based Source Separation）[8] を使用する．GHDSS は音源信号間の高次無相関化を行うブラインド音源分離と空間的指向性を形成する BF とのハイブリッド手法である．

BF によって得られた分離音 $Y_i^{(m)}(f,\tau)$ には，目的音源 S_i と目的音源方向に存在するその他の音源が支配的になっていると考えられるため，すべての BF 分離音に共通する音響成分は目的音源の可能性が高い．そこで，すべてのマイクロホンアレイにおける分離音に共通する成分だけを抜き出すことで，スポット内に存在する目的音源を抽出する．次節ではこの考え方に基づき，BF 分離音に共通する音響特徴成分の抽出する方法について説明する．

3.2 NMF による共通成分抽出

本節では，BF 分離音 $Y_i^{(m)}(f,\tau)$ から共通成分を抽出する処理について説明する．NMF は振幅スペクトルがいくつかの基底となるスペクトルの重ね合わせで表現できると考え，非負制約下で低ランク近似を行う．

全 BF 分離音に共通する音響成分を推定するために，各マイクロホンアレイで得られた振幅スペクトログラムを $\boldsymbol{Y}_i^{(m)} \in \mathbb{R}_+^{F\times T_f}$，時間軸方向に結合したものを $\boldsymbol{Y}_i = [\boldsymbol{Y}_i^{(1)}, \ldots, \boldsymbol{Y}_i^{(M)}] \in \mathbb{R}_+^{F\times MT_f}$ とする．この $\boldsymbol{Y}_i$ に対して NMF を適用すると，$\boldsymbol{Y}_i$ を基底行列 $\boldsymbol{W_i}$ とアクティベーション行列 $\boldsymbol{H}_i$ の積に分解できる．

$$\boldsymbol{Y}_i \approx \boldsymbol{V}_i\boldsymbol{H}_i \tag{2}$$

ここで，$\boldsymbol{W}_i \in \mathbb{R}_+^{F\times K}$, $\boldsymbol{H}_i \in \mathbb{R}_+^{K\times MT_f}$ であり，K はあらかじめ決める基底数である．基底行列 $\boldsymbol{V}_i$ は K 種類の基底スペクトルを格納しており，$\boldsymbol{H}_i$ は各フレームにおける基底の強度を表している．NMF は，$\boldsymbol{Y}_i$ と $\boldsymbol{V}_i\boldsymbol{H}_i$ の誤差が小さくなるように基底行列とアクティベーション行列を学習する．学習法としては平均二乗誤差を目的関数とし，乗法更新アルゴリズム [1] を適用する．

続いて，$\boldsymbol{H}_i = [\boldsymbol{H}_i^{(1)}, \ldots, \boldsymbol{H}_i^{(M)}]$ のように $\boldsymbol{H}_i$ を時間フレーム方向に M 個に分割する．この分割により，式 (2) は次のように分解できる．

$$\begin{aligned} \boldsymbol{Y}_i^{(1)} &\approx \boldsymbol{V}_i\boldsymbol{H}_i^{(1)} \\ &\vdots \\ \boldsymbol{Y}_i^{(M)} &\approx \boldsymbol{V}_i\boldsymbol{H}_i^{(M)} \end{aligned} \tag{3}$$

式 (3) の各式は NMF による $\boldsymbol{Y}_i^{(m)}$ の分解になっている．基底行列 $\boldsymbol{V}_i$ はすべてのマイクロホンアレイに共通し，アクティベーション行列 $\boldsymbol{H}_i^{(m)}$ はマイクロホンアレイごとに得られる．これにより全てのマイクロホンアレイのスペクトログラムを通して学習された基底行列 $\boldsymbol{W}_i$ を得ることができる．

NMF では基底と音源の対応づけがされないため，目的音源に対応する基底を各フレームごとに指定する必要がある．本手法では，「各マイクロホンアレイの BF 分離音において，同時刻に出てくる共通の基底（音響成分）は目的音源である」という仮定を置く．類似する周波数構造を持つスペクトルは同じ基底で表される可能性が高くなる．例えば，目的音源はすべてのチャンネルに存在するため，音量の違いがあったとしても同じ基底が同時刻に推定されやすい．この仮定に基づき，基底の出現時刻がどの分離音でも一致している基

底は目的音源に対応すると考える．アクティベーション行列に対するバイナリマスク行列 $\boldsymbol{B}_i$ を次のように作成することで目的音源の基底を抽出する．

$$b_{i,k\tau} = \begin{cases} 1 & (\min(h^{(1)}_{i,k\tau}, h^{(2)}_{i,k\tau}, \cdots, h^{(M)}_{i,k\tau}) > \gamma) \\ 0 & (\text{otherwise}) \end{cases} \quad (4)$$

$h^{(m)}_{i,k\tau}$, $b_{i,k\tau}$ はそれぞれ各基底 k，各時間フレーム τ におけるアクティベーション行列 $\boldsymbol{H}^{(m)}_i$ とバイナリマスク行列 $\boldsymbol{B}_i$ の要素を表している．また，γ は各基底が稼働しているかどうかを決めるための閾値である．式 (4) では，各時刻，各基底ごとに最小値が閾値より大きい時は 1 に，小さい時は 0 にする．この処理はフレームごとに行われるため，フレーム単位での同期が必要である．ただし，ビームフォーミングの出力は同期されているとは限らないため，フレーム単位ではなく数フレーム単位に区切って判定を行うことで時間ずれの影響を緩和できる．また，時間ずれが大きい場合は，自身で時間を合わせる必要がある．この処理を各基底，各フレームごとに行うことでバイナリマスクを得ることができる．

得られたバイナリマスクによって推定信号のスペクトログラム $\hat{\boldsymbol{S}}^{(m)}_i$ は次のように計算できる．

$$\hat{\boldsymbol{S}}^{(m)}_i = \boldsymbol{V}_i(\boldsymbol{H}^{(m)}_i \odot \boldsymbol{B}_i) \quad (5)$$

ここで，$\odot$ は行列の要素積を表す．マイクロホンアレイごとに得られた推定信号に対して逆 STFT した後，時間ずれを相関関数を用いて補正し平均化することで，最終的な出力 $\hat{s}_i$ とした．

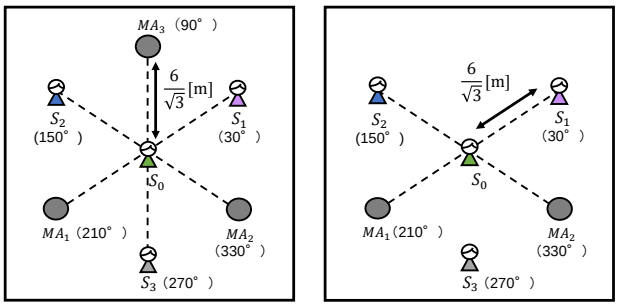
4 実験

本節では，提案手法についてシミュレーションと実環境での実験を行い，分離性能の評価を行う．

4.1 シミュレーション実験

シミュレーションには PyRoomAcoustics[1] という室内音響シミュレーションツールを利用した．図 3 のような 10m×10m の部屋を作成し，音源とマイクロホンアレイを配置した．マイクロホンアレイは円形，8ch，半径 3.65cm とした．また，残響時間 RT60 を 0s，0.3s，0.7s，1.1s に設定して，残響込みのシミュレーションを行った．音源は JNAS の新聞記事読み上げコーパス [7] から 7–10s 程度の 4 つの音源組を 20 パターン用意した．STFT の窓関数は Hamming 窓を用い，窓長 512，シフト幅 256 とし，サンプリング周波数は 16kHz とした．

[1] https://github.com/LCAV/pyroomacoustics



(a) three microphone arrays (b) two microphone arrays

図 3: シミュレーションの設定

表 1: 残響時間（RT60）の変化に伴う音源 S_0 のシミュレーション結果（SDR 改善値 [dB]）

method	アレイ数	0s	0.3s	0.7s	1.0s
NMF–SF	3	**14.3**	**11.3**	**9.7**	**9.1**
Delay–Sum	3	12.0	10.3	8.9	8.3
NMF–SF	2	12.6	9.7	8.2	7.7
Delay–Sum	2	10.0	9.1	7.7	7.1

ビームフォーミングには，ロボット聴覚用オープンソースソフトウェア HARK[9] に含まれる GHDSS [8] ノードを用いて計算した．音源方向は既知とし，GHDSS の入力に方向情報を与えた．NMF の基底数は 100 で統一し，アクティベーション行列の閾値 γ は 1.6×10^{-3} とした．

評価指標として，BssEval [10] に含まれる Source to Distortion Ratio（SDR）を使用する．SDR は推定信号に含まれる目的音源成分と目的音源以外のノイズ成分のパワー比によって定義され，分離音の品質を表す指標である．ソースコードは Bss Eval toolbox [11] を利用し，収録した状態の混合音からどれだけ SDR が改善するかを評価した．

また，比較として BF 分離音を遅延和（Delay–Sum）を適用した．

4.1.1 シミュレーション結果

シミュレーション環境で音源 S_0 に対して分離した結果を表 1 に示す．表から，各残響時間においてマイクロホンアレイが 2 台の場合と 3 台の場合ともに提案法の方が遅延和法よりも SDR 値が改善していることが分かる．また，音源 S_0 のように音源方向に別の音源が存在するような場合，どちらの手法もマイクロホンアレイ数が多い方が SDR 値が大きくなり，精度が向上していることが分かる．

次に，実際に推定された信号のスペクトログラムを図 4 に示す．スペクトログラムをみると，提案法は遅

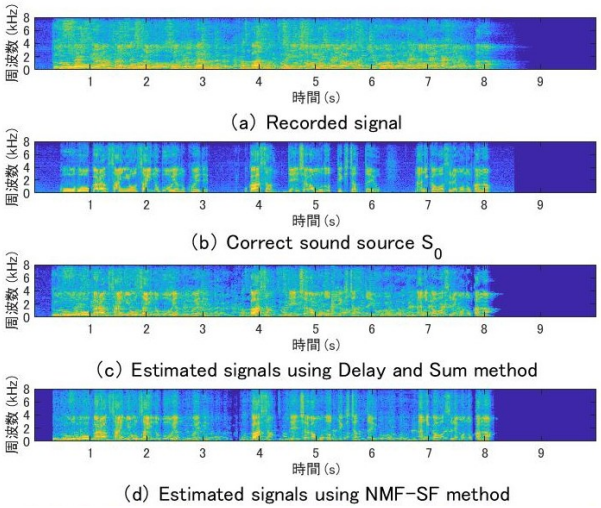


図 4: シミュレーション環境下で残響時間を約 0.7s とした際の分離結果を示す．上から収録信号，正解信号，遅延和による推定信号，提案法による推定信号を表す．横軸が時間縦軸が周波数，色が強度を表している．

延和法に比べノイズが除去され，正解信号に近いスペクトログラムになっている．例えば，3s〜4s の区間を見てみると遅延和法では残っていた雑音が低減されていることがわかる．遅延和法は目的音源を強調する一方で，目的音源以外の音を完全に除去することができない，これに対して，提案法は NMF により推定された基底スペクトルが目的音源かどうかを判定し，目的音源でないものに対してバイナリマスクをかけて除去するため，比較的他音源ノイズの除去に適していると考えられる．

4.2 実環境での実験

次に，実環境での実験を行った．音源はシミュレーションと同様に JNAS の音声コーパス [7] から 7–10s 程度の 4 つの音源組を 20 パターン用意し，GENELEC 8010APM スピーカーから音源を流した．収録装置は図 7 に示すような 16 チャンネルのマイクロホンアレイを使用し，サンプリング周波数 16kHz，量子化ビット数 24bit で収録した．各装置は図 5，6 のように配置し，高さは 1.2m で統一した．伝達関数は，HARKTool5 を利用し，幾何学的計算により作成した．NMF の基底数は 100 とし，アクティベーション行列の閾値 γ は 1.6×10^{-3} とした．比較手法として，遅延和法（Delay–Sum）と第 2 節で説明したクラスタリングベースのスポットフォーミング（LDA-SF）[5] を適用する．

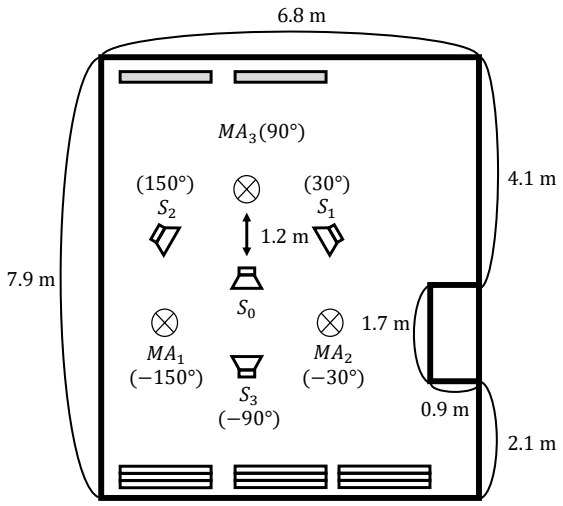


図 5: 実験場の見取り図

図 6: 実験の様子

4.2.1 実験結果と考察

実験結果を表 2 に示す．表は各音源に対する分離結果の SDR 改善値を表している．表を見ると，LDA–SF の性能が低くなっている．LDA–SF は第 2 節で説明したように同時発話音声を分離できないため，分離性能が低くなっていると考えられる．NMF–SF は，遅延和法とほぼ同等の性能になっており，あまり SDR 値が改善していないことがわかる．

この原因として考えられるのが残響音である．残響音が大きい環境では，ビームフォーミングの分離性能が低下してしまう．残響音の影響を見るためにスペクトログラムを図 8 に示す．図は上から，正解音源，各マイクロホンアレイ MA_1，MA_2，MA_3 でビームフォーミングして得られた分離音，提案手法による推定信号が表示されている．各マイクロホンアレイの BF 分離音を見ると，残響音が除去できておらず目的音源方向外の雑音が残っていることがわかる．各 BF 分離音に目的音源方向外の音が混入しているため，第 3.2 節で説明した共通基底選択が機能しなくなってしまってい

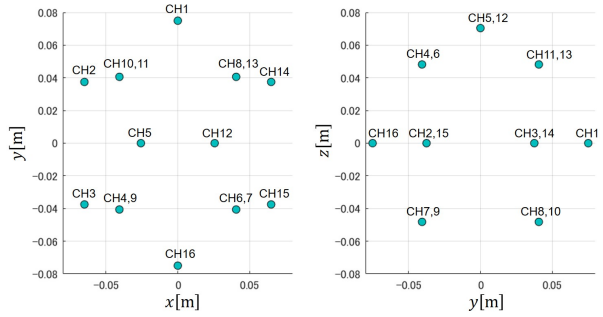


図 7: 使用したマイクロホンアレイの構成

表 2: 実環境での実験結果（SDR 改善値 [dB]）

method	アレイ数	S_0	S_1	S_2	S_3
NMF–SF	3	**3.4**	**5.0**	**11.8**	**12.5**
LDA–SF	3	0.73	1.0	1.3	0.83
Delay–sum	3	3.2	4.8	11.6	11.8
NMF–SF	2	2.6	2.9	2.1	5.3
Delay–sum	2	2.6	3.6	2.1	5.5
LDA–SF	2	0.80	-0.45	-1.2	2.3

ると考えられる．

その他に考えられる原因としては，スピーカーの指向性がある．シミュレーション環境では音源は等方的に広がっていると仮定される一方，実環境では音源に向きがあるため，マイクロホンアレイによって音が入りにくかったり，反射音の方が大きくなったりする可能性がある．現状では NMF の共通成分推定をアクティブ行列の閾値で判定しているため，誤検出が多くなると考えられる．

5　むすび

本稿では，複数台のマイクロホンアレイを用いた NMF による空間音源分離手法の実環境での性能評価を行った．シミュレーションにおいて，提案手法は既存手法よりも SDR がおよそ 0.8dB〜2.3dB 向上することが分かった．しかし，実環境では残響の影響を受けやすく，シミュレーション時に比べて分離精度が向上しにくいということが分かった．今回の実験はケーススタディであるため，様々な会場やシチュエーションで実験評価を進めていき，よりロバストな分離手法を構築することが今後の課題である．

謝辞

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

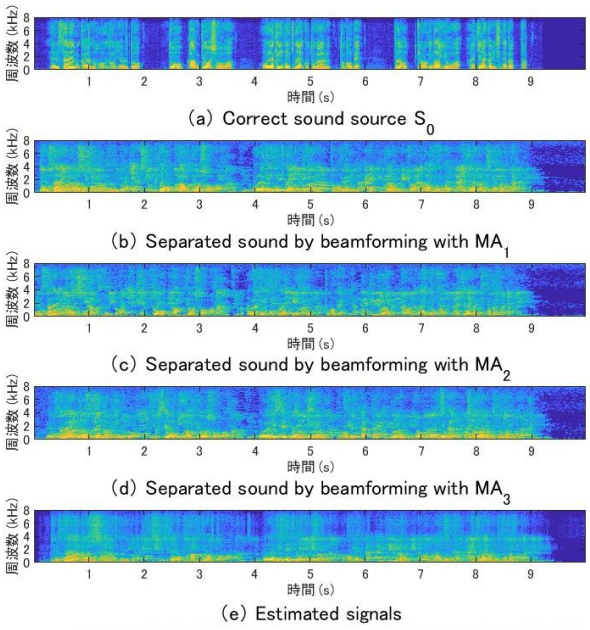


図 8: 分離結果のスペクトログラム．上から正解信号，マイクロホンアレイ MA_1，MA_2，MA_3 でビームフォーミングした分離音，最終的な推定信号を表す．

参考文献

[1] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca and R. J. Plemmons : “Algorithms and applications for approximate nonnegative matrix factorization”, *Computational Statistics and Data Analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[2] K. Sekiguchi, Y. Bando, K. Itoyama and K. Yoshii : “Layout optimization of cooperative distributed microphone arrays based on estimation of source separation performance”, *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 83–93, 2017.

[3] 鈴木基之, 本城剛士 : “2 本の超指向性マイクを用いたスポットフォーミング法の提案”, 研究報告音楽情報科学 (*MUS*), vol. 2014, no. 71, pp. 1–6, 2014.

[4] M. Taseska and E. A. P. Habets : “Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 7, pp. 1291–1304, 2016.

[5] 鍵本泰宏, 糸山克寿, 西田健次 and 中臺一博: “複数マイクロホンアレイを用いた LDA によるスポッ

トフォーミングの検討”, 第 20 回計測自動制御学会システムインテグレーション部門講演会 (SI2019)) 講演論文集, pp. 1505-1510, 2019.

[6] 鍵本泰宏, 糸山克寿, 西田健次 and 中臺一博: “複数マイクロホンアレイを用いた NMF による空間音源分離法の提案と評価”, 日本ロボット学会誌, vol. 39, no. 7, pp. 669–672, 2021.

[7] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi： “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research.” *The Journal of the Acoustical Society of Japan (E)* vol. 20, no. 3, pp. 199–206, 1999.

[8] H. Nakajima, K. Nakadai and 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. 1476–1485，2009.

[9] 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.

[10] E. Vincent, R. Gribonval and C. Févotte：“Performance measurement in blind audio source separation”，*IEEE Transactions on Audio, Speech, and Language Processing*，14(4)，pp. 1462–1469，2006.

[11] F.-R. Stöter, A. Liutkus and N. Ito：“The 2018 signal separation evaluation campaign”，*International Conference on Latent Variable Analysis and Signal Separation*，pp. 293–305，2018.

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-058-06

Speech-Classification in Edge-Computing for IoT Applications

Haris Gulzar*, Muhammad Shakeel*, Kenji Nishida*, Katsutoshi Itoyama*, Kazuhiro Nakadai* †
*Dept. of Systems and Control Engineering, Tokyo Institute of Technology, Tokyo, Japan
Email: {gulzar, shakeel, nishida, itoyama}@ra.sc.e.titech.ac.jp,
†Honda Research Institute Japan Co., Ltd., Saitama, Japan
Email: nakadai@jp.honda-ri.com

***Abstract*—Speech based interface for interacting with smart devices has recently gained traction due to significant improvement in Machine Learning (ML) based algorithms for speech recognition and classification. Edge computing has come into play to make cloud-computing scalable because number of low power AI-enabled Internet of Things (IoT) devices is increasing rapidly. Convolutional Neural Network (CNN) has proved its performance in image recognition and speech classification alike. In this paper, we have proposed an edge computing solution using System-on-Chip based device from perspective of speech-enabled IoT applications. Speech commands classification task is performed to demonstrate the acceleration of CNN network on SoC edge computing device. As IoT and edge devices have limited computational resources, ideally a smaller model with similar performance as state-of-the-art models is required to be deployed. We have taken a data-centric approach to elevate the performance of CNN for audio classification task with a very light CNN model. The comparison of proposed approach has shown that our CNN model achieved similar performance as large models with 6X smaller number of parameters and 14X smaller number of Floating-Point Operations (FLOPs). We have also implemented the acceleration for our CNN model on FPGA part of SoC processor to reduce latency in real time implementation. Our implementation demonstrated more than 6X acceleration factor as compared to base implementation which is also higher than other proposed approaches for CNN acceleration.**

***Index Terms*—Edge Computing, Machine Learning, Internet of Things, Hardware Accelerator, Speech Classification**

I. Introduction

Efficiency of various human machine interfaces like face expression, gestures and speech recognition has been improving with advancement and innovation in Machine Learning (ML) algorithms [1]. Many State-of-the-Art (SOTA) ML models have been proposed to improve the performance of speech recognition as it is becoming de-facto interface in Human Robot Interaction (HRI), which has already found its application in assistive robots for various indoor scenarios [2] and coworking robots in industrial environments [3]. A Convolutional Neural network (CNN) based ML approach has also been proposed to control the drones using different voice commands [4]. Another ML based approach is introduced to maneuver the airplane in a simulated setup which gives hint in further expansion of speech-based applications in diverse environments [5]. Currently high performing ML models are significantly large in terms of memory and computations and they are usually deployed on large cloud platforms with abundant resources available for high computations. Recently, the focus has been shifted to deploy ML models on low power IoT devices which are increasing in demand and in numbers in several daily life applications [6]. These low power IoT devices due to their small footprints have increased in number and with speech-enabled functionality they are particularly useful for handicapped people who have some mobility restrictions [7].

This work is supported by JST, CREST Grant No. JPMJCR19K1, Japan.

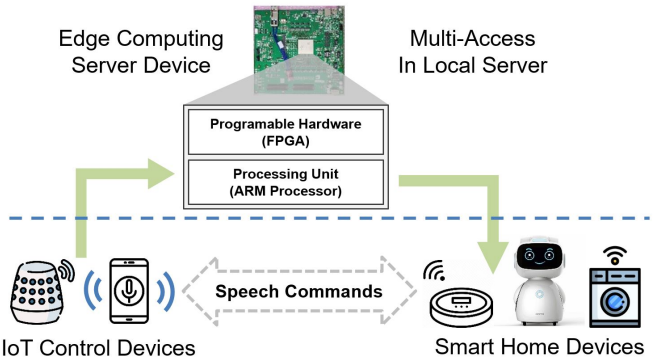


Fig. 1. Speech Interface for IoT devices enabled by Edge Computing.

Devices with smaller footprints would have small space for computational resources, thus the computational burden has to be unloaded to nearby computational resource and this is where edge computing comes into play. In addition to that transferring personal speech data to remote cloud is vulnerable to cyber-attacks and pose a privacy concern. Continues integration with cloud computing would also require a seamless internet connection and any disturbance in network connection may lead to unwanted scenario in terms of high latency or network interruption. Edge computing solves these problems by bringing the computational resources from cloud close the application node. A device enabled with speech recognition having additional computational resource in the local server in form of network edge is an ideal scenario for speech-enabled IoT applications. Fig. 1 shows that how small devices can utilize the local edge server to utilize its computation power to run the speech enabled applications. A multi-access edge computing approach also allows to elevate the fault tolerance of the overall system by using multiple

devices, where one device can take over the role of other device in case of device failure.

Edge devices are still inferior to cloud devices in terms of available resources due to their smaller footprint; they are basically scaled sown version of cloud nodes. ML algorithms have to be tailored according to the resource constraints of the targeted device in such a manner that there is not a significant performance degradation while reducing their computational cost and memory. Different types of devices have been considered for Edge applications including CPU, GPU and FPGA [8]. FPGA is one of the best candidates in this case due to its flexibility to be a reprogrammable hardware which allows faster delivery of the final product to the market. System on Chip (SoC) based processor has further enhanced the FPGA ability and performance by fabricating the CPU and FPGA on the same chip. As illustrated in Fig.1 SoC has two parts; in our case FPGA acts as an accelerator and CPU hosts the Operating System to provide a platform for real time application. Our work has made following contributions while keeping in mind the requirements of speech application in low power IoT devices:

- It proposed the lightest CNN model in terms of parameters and computational cost while maintaining the SOTA performance.
- It shows how taking a data-centric approach can help in achieving high performance as large models with a significantly smaller model.
- It demonstrates the deployment of FPGA accelerator for CNN for speech classification on Edge devices.

II. Related Work

A plethora of Recurrent Neural Network (RNN) and CNN based ML approaches have been proposed for speech commands classification, which are summarized in Table 1. An attention based RNN technique is also proposed for similar task [9-10], showing significant improvement in classification accuracy performance. But as seen from the Table 1, RNN models are large in terms of number of parameters which makes it harder to deploy them on small memory devices. CNN models on the other hands deliver good performance with comparatively less number of parameters. EdgeSpeechNet which is a CNN based approach [11] achieves similar performance as RNN based approaches with smaller number of parameters, but Floating-Point Operations (FLOPs) are very large. RES15 has also delivered a reasonable performance but number of parameters and FLOPs increase even further [14].

Depth wise Separable CNN (DS-CNN) approach [15] maintains good performance by exploiting depth wise convolution operations and achieves 95.4% accuracy for speech commands classification. CNN network can learn better from training data if the size of network is increased significantly but it leads to very high number of operations and parameters [16] making it impossible to be deployed on small devices. Type of network affects performance but efficiently utilizing useful information in the training data also leads to better performance with smaller network size.

In this paper, we have also shed some light on data-centric approach by demonstrating the high accuracy performance

TABLE I
Performance Metrics of Different Models on Google Speech-Commands Dataset V1

Model Type	Model Name	Accuracy (%)	Params. (K)	FLOPS (M)
RNN	MHAtt-RNN [9]	97.2	743	-
	EdgeRNN [10]	96.62	830	26.96
	EdgeRNN-G [10]	96.82	830	2.96
CNN	EdgeSpeechNetA[11]	96.8	107	343
	EdgeSpeechNetD[11]	95.8	80.3	24.5
	RES15 [14]	95.8	238	894
	DS-CNN [15]	95.4	161	56.9
	CNN [16]	96.19	1,488	-
	CNNv1 (Ours)	96.45	224	**23.7**
	CNNv1-L (Ours)	95.4	**67.7**	**6.5**
	CNNv2 (Ours)	95.11	**93.2**	**3.7**

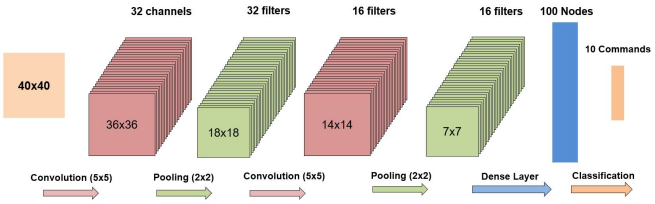


Fig. 2. CNN architecture of CNNv2.

with a significantly smaller model having small number of parameters and FLOPs. In second part of the paper, we have demonstrated the deployment of accelerator for CNN which achieved better acceleration factor as compared to proposed approaches so far. The following sections explain about the dataset used for this task, CNN network architecture, introduction of hardware and network deployment.

III. Dataset & Input Features

The dataset being used in this work is named as Speech Commands, developed by Google [23]. There are ten most common speech commands in first version of the dataset e.g., “Yes”, “No”, “Up”, “Down”, “Left”, “Right”, “On”, “Off”, “Stop” and “Go”. For developing speech-based applications, this data is a very reasonable starting point. The complete dataset is divided into two parts as 80% training data and 20% as validation data. Increasing the amount of training examples using data augmentation improves the overall performance of ML model [13]. In our work we have used random time shifting of audio signal as our data augmentation technique.

Speech commands audio files are sampled at 16kHz frequency. First of all we have taken Frequency Cepstral Coefficients (MFCC) of the signal which has 13 frequencies in one time bin equivalent to 25ms. MFFCs provide rich information of speech signal in lower dimension by extracting most of the phoneme information as speech signal is basically the continuous sequence of phonemes. In speech signal amplitude of consecutive phonemes vary significantly so we take the difference of amplitude of power spectrum for consecutive phonemes.

These features are known as delta features, which have same effect as derivative of the signal. After taking 13 delta features of the signal, we take another 13 double delta features

by taking the difference once again [24]. These features inherently make the features robust to white noise because it is suppressed by taking difference of signal. Lastly, the final feature of the signal is spectrogram power for each frame, making the final shape of features matrix to be 40x40. The CNN architecture, its accelerator design and hardware introduction are explained in the following sections.

TABLE II
OUR CNN PERFORMANCE WITH DIFFERENT INPUT FEATURES

Features Type	Features Dimension	Accuracy (%)
MFCC	13x40	94.11
MFCC, Delta	(13+13)x40	94.89
MFCC, Delta, Double-Delta, Delta-Energy	(13+13+13+1)x40	95.11

IV. OUR CNN ARCHITECTURE AND ITS COMPARISON WITH STATE OF THE ART

CNN along with its many variations has proved very high performance for image recognition task due to its ability to recognize local patterns. CNN can also perform equally well for sound classification task, given the speech data is first converted into a two-dimensional matrix like an image. In image recognition task, optimizing the network architecture for the given task is mostly the only way to improve performance, but in speech classification, we have another task to focus on, that is to efficiently extract the information from audio data. After efficient data extraction our CNN model has achieved good performance even though it is smaller in size.

TABLE III
CNN HYPERPARAMETERS COUNT

Layer	Parameters	CNNv1	CNNv2	CNNv1-L
Conv.1	Channels	64	32	32
	Kernel	5x5	5x5	5x5
Conv.2	Channels	128	16	64
	Kernel	3x3	5x5	3x3
Conv.3	Channels	64	-	32
	Kernel	3x3	-	3x3
Dense	Nodes	128	100	100
Total No. of Parameters		224,576	93,158	67,702

The model being used in this work is CNN with convolutional layers followed by a fully connected layer and classification layer. The input shape of spectrogram is a symmetric 2-dimensional matrix; hence first layer of the network is 2-dimensional convolution layer. First convolutional layer for both CNNv1 and CNNv2 has same size of symmetric kernel i.e. 5x5. The stride length of 1 for convolutional layer and 2 for pooling layers is used throughout the network. Hyperparameters like number of output channels, number of convolutional layers, kernel-size for following layers and number of nodes in fully-connected layer are varied in three versions of CNN models. Rectified Linear Unit (ReLU) activation function is used after each pooling layer. Activation function is relatively simpler to model in High Level Synthesis (HLS) as it only requires to change negative values to zero. Three variations of CNN model have been used in our experiment with different hyperparameters. The detail of hyperparameters for each

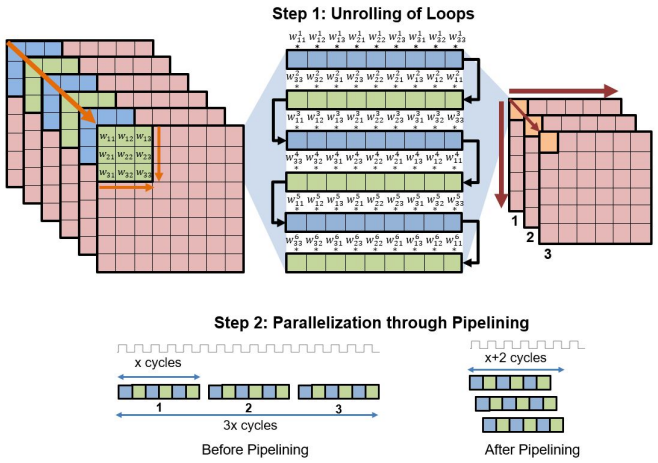


Fig. 3. Parallelization by loop unrolling followed by pipelining.

version of the model are listed in Table 3. Largest model is decided according to the constraints of the edge device which we have use in this experiment. For network training Adam optimizer is used with cross entropy loss. Training is performed for 300 epochs where validation loss and accuracy are stabilized.

CNNv1 which is the largest model in our case achieves SOTA performance with 6X smaller number of parameters than another CNN approach for a similar task [16]. In a similar way, our largest model has achieved better performance than RES15 []14] with enormously low number of operations e.g. 40X. We performed another experiment by reducing the size of CNNv1 up to 3X while maintaining performance up to 95%, and lighter version of the model is named as CNNv1-L. This model achieves similar performance as DS-CNN [15] with 2X smaller number of parameters and 9X smaller number of FLOPs. This model also has 9X smaller FLOPs than EdgeSpeechNetD [11] while achieving almost similar accuracy. We also tried to investigate the effect on accuracy by changing the number of convolutional layers and used a network with two convolutional layers, named as CNNv2. This model has large number of parameters than CNNv1-L still has lower accuracy. This clearly shows the supremacy of convolutional layers, which delivers better performance by doing more operations with less parameters. This also implies that deeper CNNs has better performance than shallow CNN networks.

Even our largest model CNNv1 has 7X smaller number of parameters than CNN [14] but still achieves higher performance. Similar, CNNv1 has 4X smaller size than EdgeSpeechNetD [I8] but still achives 0.65% better accuracy.

V. HARDWARE INTRODUCTION

The device used in this research is a custom-built Multiaccess Edge Computing (MEC) device named as M-KUBOS [18] which acts as edge server in our case. It has Xilinx Zynq Ultrascale+ XCZU19 System on Chip (SoC) processor on board. SoC on M-KUBOS has FPGA as programmable hardware and ARM processor as processing unit fabricated

on the same chip. Moreover, SoC has on-chip memory for storing weights of the neural network logic which cuts short the reading and writing time of data into a separate memory while computations take place. SoC delivers high end performance by leveraging the synergistic usage of processing unit and programable hardware which communicate to each other through Advanced eXtensible Interface (AXI) as illustrated in Fig. 4. In addition to the on-chip memory, M-KUBOS has additional Ultra-RAM on board for storing large data which can be accessed by ARM processor and FPGA through Direct Memory Access (DMA). M-KUBOS has Linux operating system with Python production for Zynq devices (PYNQ) functionality [22]. PYNQ framework in Linux enables easy development of applications based on Zynq devices by flexibly using developed hardware components in Python program. M-KUBOS acts as shh server and can be seamlessly accessed remotely.

High Level Synthesis (HLS) tool by Vivado is used to design the CNN Logic in C language. Neural network logic can be synthesized and exported as Intellectual Property (IP) core to be flexibly used in hardware integration on Vivado Design Suite. Vivado Design Suite provides immense amount of customization for SoC based boards like M-KUBOS, where available hardware resources can be flexibly put together using Function Block Diagram (FBD) based programming [17]. Finally integrated hardware design is synthesized and bitstream file is burned into FPGA through Python-based application program in Linux operating system. Weights from trained ML model are also transferred to on-chip memory and processing system uses FPGA as an accelerator for inference as shown in Fig. 3.

VI. Accelerator Implementation

A convolutional layer in neural network has six loops to compute activations for next layer from input channels. As described in Algorithm 1, two inner most loops compute output for one kernel operation. The loop on top of kernel operation iterates over each channel of input layer. Looping over these kernel operations for every input channel followed by a summation returns a value corresponding to one output channel as shown in Fig. 3. These kernel operations in convolutional network provides enormous amount of opportunity for parallelization. As each channel in output layer has different weights for all input channels, computations corresponding to output values for each output channel can be performed concurrently.

Convolutional operations in CNN network can be parallelized by first unrolling the loops and then performing pipelining. For example, in Fig. 3, three inner most loops are unrolled into a long vector which is responsible for outputting one value in an output channel. If the output channels are three, conventional approach would compute output activation corresponding to each channel in series manner. The simplest operation in CNN takes 3 clock cycles i.e., reading, multiplying or adding and writing. If computing a single value in one output channel takes x number of clock cycles, then three output values corresponding to each output channel would take 3x cycles. However, these independent operations can be scheduled to be executed in parallel with just one cycle delay for scheduling them as illustrated in Fig. 3.

Algorithm 1: Convolutional Function

```
   Input: input, weight, bias
   Output: activation output
1  Convolution(input, weight, bias, output)
2  for w ∈ {1, ..., output_channel_width} do
3      for h ∈ {1, ..., output_channel_height} do
4          #pragma HLS PIPELINE
5          for n ∈ {1, ..., output_channels} do
6              #pragma HLS UNROLL
7              for i ∈ {1, ..., input_channels} do
8                  for j ∈ {1, ..., kernel_width} do
9                      for k ∈ {1, ..., kernel_height} do
10                         output_whn+ = w_ijk * input_ijk
11                         +bias_ijk
12             Above three loops unrolled into a vector.
13         Above loop is pipelined for parallel processing.
14 end Convolution
```

Vivado HLS tool allows to flexibly select the loops to be unrolled and parallelized. Algorithm 1 shows that we have unrolled three inner most loops of convolution function. The loop on top of three unrolled loops is pipelined and will execute its operation in parallel. Similarly, more number of loops can be unrolled and pipelining operation can be used for even higher loops. As we can on unrolling the loops, the size of one operations will increase i.e. multiplication and as we do pipelining size of parallel operations will increase i.e. number of operations. Hence total resource utilization will increase after optimization. Theoretically all loops in convolution layer can be parallelized, given that there is no limitation on available resources. However, finding the optimal balance between amount of parallelization and meeting the resource constraints is an important task [17]. Given the availability of computational resources in M-KUBOS, three to four loops are unrolled followed by pipelining in different convolutional layers of CNN network.

VII. Experimental Setup & Results

Firstly, optimized CNN network i.e. CNNv is trained on Google Speech-Commands training dataset in Pytorch framework, and weights are obtained to transfer to chip memory before execution on M-KUBOS. Then HLS logic is synthesized on Vivado HLS and a hardware bitstream file of neural network logic is transferred to M-KUBOS device for programming the FPGA hardware. PYNQ programming framework is used for making and executing application program on peocessing system part of SoC device. After programming FPGA hardware and transferring network weights to on-chip memory Python-based application program sequentially takes the inputs and computes them by utilizing the on-chip FPGA accelerator. End-to-end latency is measured by executing base

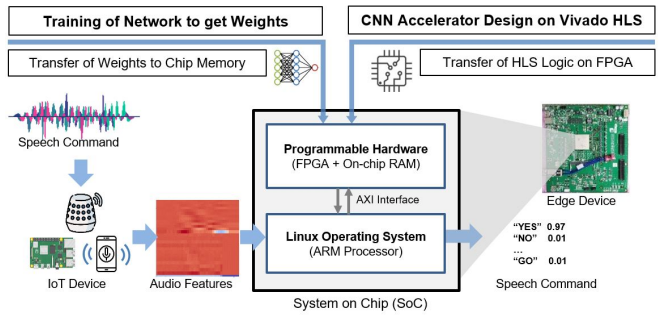


Fig. 4. Application framework for sound classification on edge device.

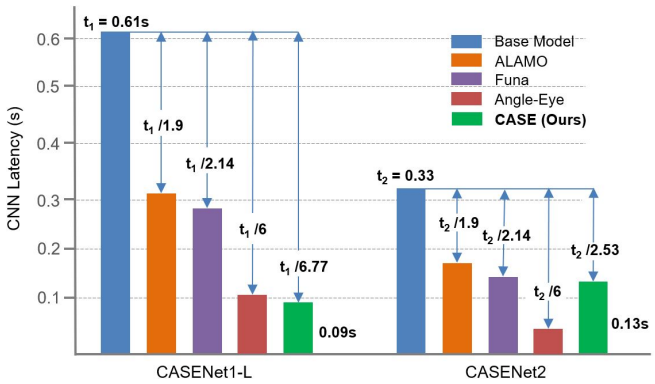


Fig. 5. Improvement in latnecy after parallelization in comparison to other methods.

implementation of our CNN logic followed by its optimized version for CNNv1-L and CNNv2. Improvement in latency after using optimized version is illustrated in Fig. 5. We used Python-based time library to measure the end-to-end execution latency for speech classification inference.

In addition to our approach, some other works are being proposed for CNN acceleration on image classification tasks, where CNN implementation remains the same but audio features input matrix is replaced by image matrix. Fune [19], a CNN accelerator tool achieves up to 2.14X speed-up of CNN by providing optimal parameters search for hardware synthesis. ALAMO [20] is another approach which also explores CNN acceleration for image recognition task by using an approach of loop unrolling followed by sequential computation, resulting in 1.9X speed-up factor. Another proposal of CNN acceleration, Angle-Eye [21] achieved 6X as their best speed-up factor by exploiting the potential of parallelization in CNN. Our approach in case of CNNv1-L has achieved an acceleration factor of 6.7X, which is higher than the previously mentioned approaches. An important observation in our experiment is that CNNv2 has lower latency without parallelization but still achieves lower acceleration factor than CNNv1. The reason is that 3 Convolutional Layers in CNNv1 provide more opportunity of parallelization than two convolutions in CNNv2. As convolution reduces size of input, so lesser number of convolutional layers in network means more parameters before fully connected layer which has relatively lower potential of parallelization.

VIII. Conclusion & Future Work

This work is focused on implementing a CNN accelerator on SoC based edge device for classification of speech commands. The CNN proposed in this paper is a very light network which delivered SOTA accuracy performance while keeping the computational cost and memory significantly lower. We demonstrated that how keeping data centric approach helps to achieve high performance with small model by efficiently extracting useful information from training data. The comparison of our model shows that it delivers high performace same as large SOTA models. Second half of the paper focused on accelerator design for CNN network which is deployed on FPGA part of SoC processor. After using the accelerated design, the end-to-end latency of the speech coammand classification was reduced by 6X. The future work related to this on-going research is as follows:

- To extend the implementation to Automatic Speech Recognition (ASR) from speech commands classification.
- To make an integrated framework for sound sepration, source localization and speech recognition.
- To measure and improve the power consumption of FPGA based device and compare it with GPU and CPU.

References

[1] R. Supreeth & Kamath Ajay & N. Srinidhi & Kumaraswamy Ramaswamy. (2021). Fully Responsive Image and Speech Detection Artificial Yankee (FRIDAY): Human Assistant. SN Computer Science. 2. 10.1007/s42979-021-00630-8.

[2] José Novoa, Rodrigo Mahu, Jorge Wuth, Juan Pablo Escudero, Josué Fredes, and Néstor Becerra Yoma. 2021. Automatic Speech Recognition for Indoor HRI Scenarios. J. Hum.-Robot Interact. 10, 2, Article 17 (May 2021), 30 pages.

[3] Mustafa Can Bingol, Omur Aydogmus, “Performing predefined tasks using the human–robot interaction on speech recognition for an industrial robot”, Engineering Applications of Artificial Intelligence, Volume 95, 2020,103903, ISSN 0952-1976.

[4] C. M. J. Galangque and S. A. Guirnaldo, ”Speech Recognition Engine using ConvNet for the development of a Voice Command Controller for Fixed Wing Unmanned Aerial Vehicle (UAV),” 2019 12th International Conference on Information & Communication Technology and System (ICTS), 2019, pp. 93-97.

[5] M. Vukovic, M. Stolar and M. Lech, ”Cognitive Load Estimation From Speech Commands to Simulated Aircraft,” in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1011-1022, 2021.

[6] C. Zonios and V. Tenentes, ”Energy Efficient Speech Command Recognition for Private Smart Home IoT Applications,” 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST), 2021, pp. 1-4.

[7] Mayer, J.. “IoT Architecture for Home Automation by Speech Control Aimed to Assist People with Mobility Restrictions.” (2017).

[8] G. Dinelli, G. Meoni, E. Rapuano and L. Fanucci, ”Advantages and Limitations of Fully on-Chip CNN FPGA-Based Hardware Accelerator,” 2020 IEEE International Symposium on Circuits and Systems (ISCAS), 2020, pp. 1-5.

[9] Rybakov, Oleg & Kononenko, Natasha & Subrahmanya, Niranjan & Visontai, Mirko & Laurenzo, Stella. (2020). Streaming keyword spotting on mobile devices. arxiv.org/abs/2005.06720.

[10] S. Yang, Z. Gong, K. Ye, Y. Wei, Z. Huang and Z. Huang, ”EdgeRNN: A Compact Speech Recognition Network With Spatio-Temporal Features for Edge Computing,” in IEEE Access, vol. 8, pp. 81468-81478, 2020.

[11] Lin, Zhong & Chung, Audrey & Wong, Alexander. (2018). EdgeSpeechNets: Highly Efficient Deep Neural Networks for Speech Recognition on the Edge. arxiv.org/abs/1810.08559.

[12] J. Xu, S. Li, J. Jiang and Y. Dou, "A Simplified Speaker Recognition System Based on FPGA Platform," in IEEE Access, vol. 8, pp. 1507-1516, 2020.
[13] A. N. Çayır and T. S. Navruz, "Effect of Dataset Size on Deep Learning in Voice Recognition," 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2021, pp. 1-5.
[14] Tang, Raphael & Lin, Jimmy. (2017). Deep Residual Learning for Small-Footprint Keyword Spotting. arxiv.org/abs/1710.10361.
[15] Zhang, Yundong, Naveen Suda, Liangzhen Lai and V. Chandra. "Hello Edge: Keyword Spotting on Microcontrollers." ArXiv abs/1711.07128 (2017).
[16] A. Soliman, S. Mohamed and I. A. Abdelrahman, "Isolated Word Speech Recognition Using Convolutional Neural Network," 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2021.
[17] L. Stornaiuolo, M. Santambrogio and D. Sciuto, "On How to Efficiently Implement Deep Learning Algorithms on PYNQ Platform," 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2018, pp. 587-590.
[18] "FPGA computing platform M-KUBOS", https://www.paltek.co.jp/design/original/m-kubos/index.html
[19] Q. Xiao and Y. Liang, "Fune: An FPGA Tuning Framework for CNN Acceleration," in IEEE Design & Test, vol. 37, no. 1, pp. 46-55, Feb. 2020.
[20] Yufei Ma, Naveen Suda, Yu Cao, Sarma Vrudhula, Jae-sun Seo, "ALAMO: FPGA acceleration of deep learning algorithms with a modularized RTL compiler Integration", Volume 62, 2018, Pages 14-23.
[21] K. Guo et al., "Angel-Eye: A Complete Design Flow for Mapping CNN Onto Embedded FPGA," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 37, no. 1, pp. 35-47, Jan. 2018.
[22] Python Productivity for Zynq (PYNQ), http://www.pynq.io/
[23] Warden, Pete. "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition." ArXiv abs/1804.03209 (2018).
[24] Delta and Delta-delta audio features, Introduction to Speech Processing, https://wiki.aalto.fi/display/ITSP/Deltas+and+Delta-deltas

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

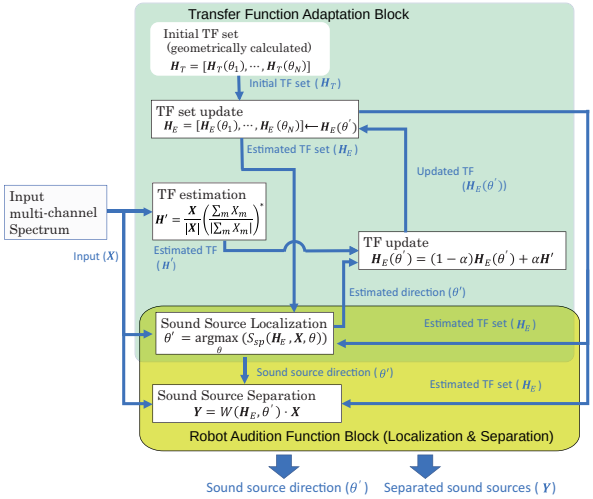


図 1: Proposed Framework for fully-online always-adaptation of TFs with sound source localization and separation

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．

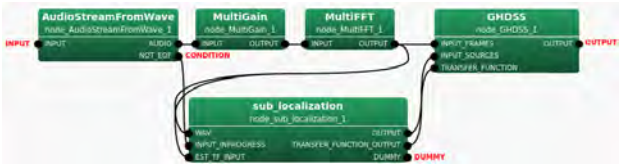
$$\boldsymbol{Y} = \boldsymbol{W}(\boldsymbol{H}_E, \theta') \cdot \boldsymbol{X}, \tag{5}$$

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．

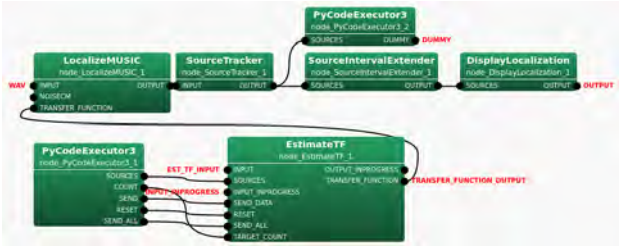
4 実装

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．

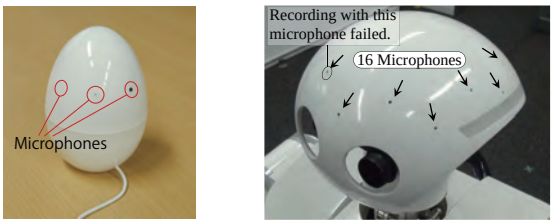
TF_T^L: 計測伝達関数セット（スピーカ高: 低）
TF_T^M: 計測伝達関数セット（スピーカ高: 中）
W_T: Tamago の周りを移動しながら録音した白色雑音
S_T: Tamago の周りを移動しながら録音した音声
M_T: 音源分離用の同時発話音声

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

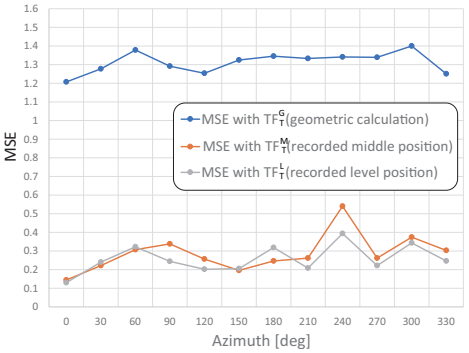


図 4: MSE in TF estimation

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/

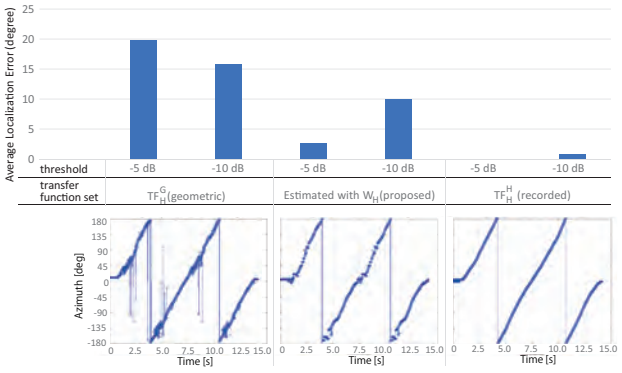


図 5: Sound source localization results: the upper panel shows the average localization errors and the lower panel shows localization results at the threshold of -10 dB.

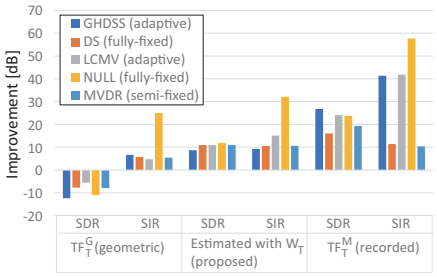


図 6: Sound source separation results

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

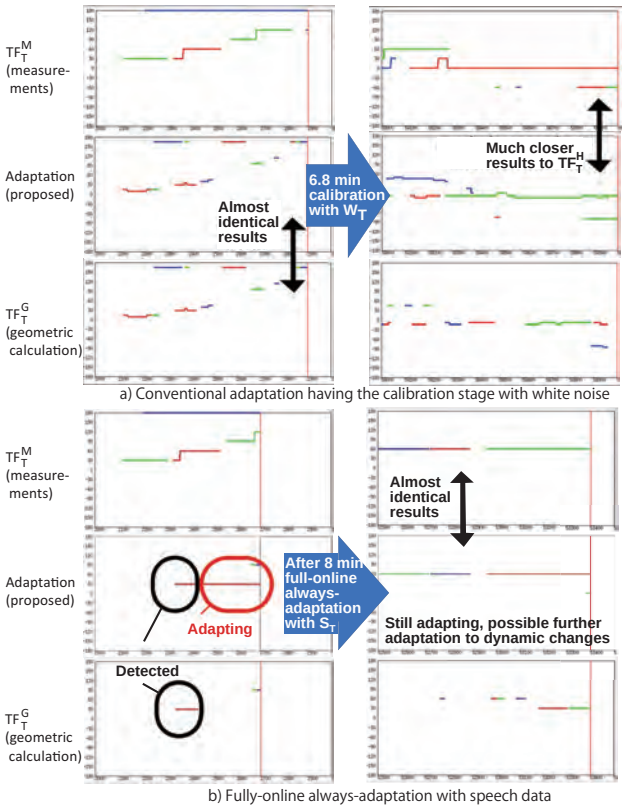


図 7: Two Online Demonstrations. The vertical and horizontal axes are the azimuth in degree and time frame in 10 ms of each subplot.

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.

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-058-08

雑踏環境における音源地図の生成

坂東 宜昭[1*]　升山 義紀[2,1]　佐々木 洋子[1]　大西 正輝[1]

[1] 産業技術総合研究所　[2] 東京都立大学

概要: 本稿では，非負値行列因子分解 (NMF) に基づく，雑踏音環境のマッピングについて述べる．移動ロボットを用いての音源の空間的な配置を推定する音環境マッピングは，知的システムが周辺環境を認識し適切な行動とるために不可欠である．従来の枠組みは，初段の音源定位に強く依存した構成となっており，残響や拡散性雑音の強い雑踏環境下では性能を発揮できなかった．そこで本研究では，音源定位の代わりに NMF を用いた音源分離に基づく音環境マッピングを提案する．本手法は，まずスペクトル特徴に基づき観測信号を分解したあと，個別の音源に対して位置推定するので，雑踏環境下でも安定してマッピングできる．数値シミュレーションにより提案法の有効性を評価した．

1　はじめに

自律移動ロボットが，環境中の音源の位置や分布を推定する音環境マッピングは，ロボットが周囲の音環境を空間的に認識し適切に応答するために不可欠な基盤技術である [1–6]．特に，展示会場や市街地など雑踏環境でも頑健に動作する枠組みを確立できれば，周囲の音響情景に合わせた適応的な接客や，都市規模の環境モニタリングなど様々な応用技術を実現できる．

従来の音環境マッピングの多くは，まずマイクロホンアレイを用いた音源定位 [7–10] により音源到来方向を推定し，異なる観測点での推定結果から三角測量することで，音源地図を生成していた [1–4]．この枠組では，時間フレームごとに動作する音源定位をオンライン処理として実装できるため，実時間で地図を生成できる．一方，最終出力である地図の推定性能が，初段である音源定位に強く依存しており，残響や拡散性雑音が強い雑踏環境では大きく性能が劣化する課題があった．

本研究では，収録した録音信号全体を直接モデル化する音環境マッピングを提案する．本手法の初段では，フレーム独立に処理していた音源定位の代わりに，録音信号全体を単一の生成モデルで音源分離する非負値行列因子分解 (NMF) [11, 12] を適用する．NMF で分解された個別の音源信号に対し，音源定位と三角測量によりその位置を推定する．初段で NMF を用いて観測信号を予め分解することで，アレイ信号処理では性能劣化しやすい残響や拡散性雑音が含まれる観測信号でも頑健な動作を実現する．また NMF は，観測信号全体を一挙に分解する枠組みであるため，同じ音源を複数回観測した場合など，長い録音信号全体の共起関係をモデル化でき，安定して複数の音源を弁別できる．数値混合音を用いた評価実験でその有効性を評価した．

*連絡先：国立研究開発法人 産業技術総合研究所
〒 135-0064 東京都江東区青海 2-4-7 産総研 臨海副都心センター 8F
E-mail: y.bando@aist.go.jp

2　NMF に基づく音環境マッピング

提案法では，移動ロボットを用いて収録した観測混合音 $\mathbf{y}_{\tau,ft} \in \mathbb{C}^M$ およびロボットの自己位置 $\mathbf{x}_\tau \in \mathbb{R}^2$ から，音源 $n = 1, \ldots, N$ の座標 $\mathbf{s}_n \in \mathbb{R}^2$ を推定する．ここで，$\tau = 1, \ldots, \mathcal{T}$, $f = 1, \ldots, F$ および $t = 1, \ldots, T$ は，それぞれミニバッチ，周波数ビンおよび時間フレームインデックスを表す．本手法では，観測信号を T フレームのミニバッチに分割し，バッチ τ 内ではロボットの移動を無視できると仮定する．以降で述べる通り，提案法は，1) NMF による観測信号分解，2) 基底クラスタリングによる音源パワースペクトル密度の推定，3) 三角測量による音源位置推定の 3 つの処理で構成される．

2.1　ガンマ過程非負値行列因子分解

提案法ではまず，スペクトル情報に基づき観測信号を分解するため，NMF を適用する．具体的には，多チャネル観測信号 $\mathbf{y}_{\tau,ft}$ の平均パワースペクトログラム $\bar{y}_{\tau,ft} \triangleq \frac{1}{M}\|\mathbf{y}_{\tau,ft}\|_2^2 \in \mathbb{R}_+$ を，K 本のスペクトル基底 $\mathbf{w}_k = [w_{k1}, \ldots, w_{kF} \in \mathbb{R}_+^F$ とそれらの時間アクティベーション $\mathbf{h}_{\tau,k} = [h_{\tau,k1}, \ldots, h_{\tau,kT}] \in \mathbb{R}_+^T$ の積へ分解する．このとき，基底数 K は，観測信号の複雑さに合わせて最適な値が変動するため，事前に決定することが難しい．そこで本手法では，基底数を事前に定めないガンマ過程 NMF [12] を用いて観測 $\bar{y}_{\tau,ft}$ を表現する．

$$\bar{y}_{\tau,ft} \sim \mathrm{Exp}\left(\sum_{k=1}^{K} g_{\tau,k} w_{kf} h_{\tau,kt}\right) \tag{1}$$

ここで，$g_{\tau,k} \in \mathbb{R}_+$ は，各基底の観測に対する寄与度合いを表すゲイン変数である．このゲイン変数にスパースな事前分布を仮定することで，十分大きい K を設定すれば，観測の複雑さに合わせて必要な本数の基底の

みを用いた分解が行われる．ゲイン変数 $g_{\tau,k}$ には，共役事前分布である以下のガンマ分布を仮定する．

$$g_{\tau,k} \sim \mathcal{G}\left(\frac{a_0}{K}, a_0\right) \tag{2}$$

ここで，$a_0 \in \mathbb{R}_+$ は $g_{\tau,k}$ のスパース度合いを制御するハイパーパラメータである．本モデルは，$K \to \infty$ で $g_{\tau,k}$ がガンマ過程に従う近似モデルとなっている．また，基底 w_{kf} とアクティベーション $h_{\tau,kt}$ には，それぞれ期待値が 1 となる以下のガンマ事前分布を仮定する．

$$w_{kf} \sim \mathcal{G}\left(a_1, a_1\right), \quad h_{\tau,kt} \sim \mathcal{G}\left(a_2, a_2\right) \tag{3}$$

ただし $a_1 \in \mathbb{R}_+$ と $a_2 \in \mathbb{R}_+$ は，それぞれ w_{kf} と $h_{\tau,kt}$ のスパース度合いを制御するハイパーパラメータである．

NMF の推論では，真の事後分布 $p(\mathbf{G}, \mathbf{W}, \mathbf{H} \mid \bar{\mathbf{Y}})$ を近似する変分事後分布 $q(\mathbf{G}, \mathbf{W}, \mathbf{H}) \triangleq q(\mathbf{G})q(\mathbf{W})q(\mathbf{H})$ を求める．この推論は，文献 [12] と同様に導出した以下の更新則を反復することで行われる．

$$q(g_{\tau,k}) \leftarrow \mathcal{GIG}\Bigg(\frac{a_0}{K}, a_0 + \sum_{f,t} \frac{\langle w_{kf} h_{\tau,kt} \rangle}{\omega_{\tau,tf}}, \sum_{f,t} \bar{y}_{\tau,ft} \phi^2_{\tau,tfk} \left\langle w^{-1}_{kf} h^{-1}_{\tau,kt} \right\rangle \Bigg) \tag{4}$$

$$q(w_{kf}) \leftarrow \mathcal{GIG}\Bigg(a_1, a_1 + \sum_{\tau,t} \frac{\langle g_{\tau,k} h_{\tau,kt} \rangle}{\omega_{\tau,tf}}, \sum_{\tau,t} \bar{y}_{\tau,ft} \phi^2_{\tau,tfk} \left\langle g^{-1}_{\tau,k} h^{-1}_{\tau,kt} \right\rangle \Bigg) \tag{5}$$

$$q(h_{\tau,kt}) \leftarrow \mathcal{GIG}\Bigg(a_2, a_2 + \sum_{f} \frac{\langle g_{\tau,k} w_{kf} \rangle}{\omega_{\tau,tf}}, \sum_{f} \bar{y}_{\tau,ft} \phi^2_{\tau,tfk} \left\langle g^{-1}_{\tau,k} w^{-1}_{kf} \right\rangle \Bigg) \tag{6}$$

ここで，$\mathcal{GIG}$ は一般化ガウス分布 [12] を表し，$\omega_{\tau,tf} \in \mathbb{R}_+$ と $\phi_{\tau,tfk} \in \mathbb{R}_+$ ($\sum_k \phi_{\tau,tfk} = 1$) は以下の値をとる補助変数である．

$$\omega_{\tau,tf} = \langle g_{\tau,k} w_{kf} h_{\tau,kt} \rangle \tag{7}$$

$$\phi_{\tau,tfk} \propto \left\langle \frac{1}{g_{\tau,k} w_{kf} h_{\tau,kt}} \right\rangle^{-1} \tag{8}$$

この更新則は，対数周辺尤度 $\log p(\bar{\mathbf{Y}})$ の変分下限を最大化するように導出されている．変分下限の最大化は，変分事後分布と真の事後分布との Kullback–Leibler ダイバージェンス $\mathcal{D}_{\mathrm{KL}}[q(\mathbf{G}, \mathbf{W}, \mathbf{H}) \mid p(\mathbf{G}, \mathbf{W}, \mathbf{H} \mid \bar{\mathbf{Y}})]$ の最小化に対応する [13]．

2.2 基底のクラスタリング

初段の NMF で得た K 個の基底は，次段のベイズ混合ガウスモデル (GMM) により，N 個の音源にクラスタリングされる．NMF は，観測信号を K 個の基底とアクティベーションの組に分解するが，これらは N 個の音源と 1 対 1 対応しない．本稿では，NMF の各基底が複数の音源を表現せず，1 つの音源のみと対応すると仮定し，基底を N 個のクラスタに分割する．具体的には，バッチごとの基底ゲインの期待値 $\langle g_{\tau,k} \rangle$ を特徴量としてベイズ GMM によりクラスタリングする．ベイズ GMM は，そのディリクレ事前分布の効果により，十分な音源数 N を準備しておけば，不要な音源クラスを自動的に縮退させることができる．

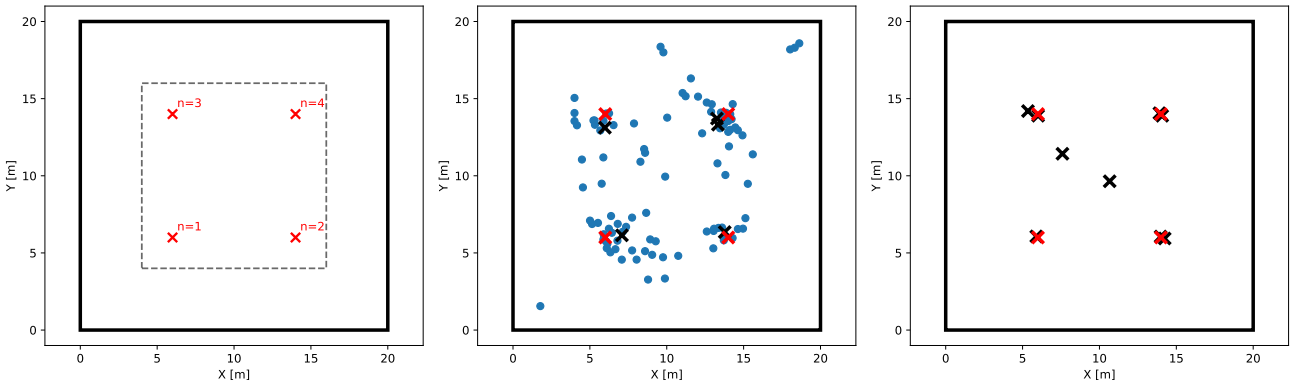
2.3 音源位置の推定

再終段では，ガンマ過程 NMF とベイズ GMM により得た N 個の音源パワースペクトル密度を用いて，多チャネル観測混合音 $\mathbf{y}_{\tau,ft}$ から音源位置 s_n を推定する．具体的にはまず，観測信号を Wiener フィルタリングすることで，音源像 $\hat{\mathbf{y}}_{\tau,nft} \in \mathbb{C}^M$ を得る．次に，遅延和ビームフォーマを用いて音源像の空間スペクトルを計算し，バッチ τ での音源 n の到来方向 $d_{\tau,n}$ を推定する．得られた音源到来方向 $d_{\tau,n}$ と自己位置 $\mathbf{x}_\tau$ を用いて三角測量し，音源位置 $\mathbf{s}_n$ を得る．

2.4 視覚情報に基づく枠組みとの関連

活発に研究されている隣接技術の 1 つである，単眼カメラによる環境マッピングについて概観しながら，音環境マッピングとの関連および提案法の位置づけについて議論する．視覚情報を用いた代表的なマッピング技術として，既知の自己位置から地図を推定する多視点ステレオ [14] と，自己位置と地図を同時推定する visual SLAM (simultaneous localization and mapping) [15, 16] や SfM (structure from motion) [17] がある．従来の音源定位に基づくマッピングは，音源定位により得た音源到来方向から地図上の音源位置を推定する点で，多視点ステレオに近い枠組みである．一方，本稿で扱う音環境マッピングでは，自己位置推定しないものの，音源位置だけでなく混合音から音源信号を推定する点で，マッピングのみを解く多視点ステレオより visual SLAM の方が類似点が多い．Visual SLAM は，画像上の疎な特徴点に対して幾何的な誤差を最小化する特徴点法 [15] と，画像間の輝度値の誤差を直接最小化する直接法 [16] に大別できる．提案法の観測モデルである式 (1) は，多チャネル録音信号 $\mathbf{y}_{\tau,ft}$ に対するゼロ平均多変量複素ガウス尤度に対応している[1]．観測信号に対する誤差 (尤度) を直接モデル化している点で，提案法は直接法に近い．一方提案法は，NMF，基底クラスタリングおよび三角測量の 3 つの処理をカスケード接続

[1] 完全な一致には，対数尤度に $1/M$ を乗じる必要がある．



(a) 音源位置 (赤) とロボット軌跡 (灰)　(b) 音源定位に基づくマッピング　(c) NMF に基づくマッピング (提案法)

図 1: (a) に実験設定を，(b) と (c) にマッピング結果を示す．灰色線はロボットの軌跡を，赤 × は正解の音源位置を表す．黒 × は推定された音源位置で，青点は HARK により検出された各音源の三角測量結果である．

して構築されており，visual SLAM や SfM のような全体最適化にはなっておらず，後段の処理へ誤差が蓄積されていくと考えられる．音源地図と音源分離の同時最適化は今後の課題とする．

3　評価実験

本節では，移動ロボットで観測した録音信号を模した，数値混合音を用いた評価実験について報告する．

3.1　実験設定

屋内を移動ロボットが巡回して収集した 10 分間の $M = 4$ チャネル混合音を生成して評価に用いた．この実験では，残響時間 (RT_{60}) 600 ms を持ち高さ 4 m で 20 m 四方の室内を想定し，4 個の音源を室内に配置した．各音源は，FSDKaggle2018 データセット [18] から選んだそれぞれ 20 秒以上の長さを持つクリップを繰り返し発している．これらのクリップには，犬の鳴き声やチャイム音など，非定常な信号を選んだ．ロボットは，図 1 のように，この室内を 2 分間で 1 週する速度で移動した．混合音は鏡像法 [19] を用いて生成した．また，観測信号には拡散性雑音として，WSJ0 [20] から選んだ英語音声によるバブルノイズを信号対雑音比 15 dB で重畳している．混合音は，サンプリング周波数 16 kHz として生成した．

音源定位に基づく音環境マッピングを実装し，提案法と比較した．この手法は，ロボット聴覚オープンソースソフトウェア HARK [21] が提供している MUSIC (multiple signal classification) 法 [9, 10] による音源定位と音源トラッキング法から音源到来方向軌跡を推定し，三角測量により音源位置を推定する．HARK による音源定位では，音源信号の無音区間やロボットが音源から定期的に遠ざかることを要因として，別の時刻に観測された同じ音源を別の音源として検出してしまう．そこで，得られた複数の音源位置をベイズ GMM でクラスタリングして，最終的な音源位置の出力とした．

表 1: 正解音源位置から最近傍の推定位置の平均誤差．

手法	誤差 [m]
音源定位に基づくマッピング	0.797
NMF に基づくマッピング (提案法)	0.104

3.2　実験結果

図 1-(b) と-(c) に推定結果を，表 1 に位置推定誤差を示す．音源定位 (MUSIC) に基づくマッピング (図 1-(b)) では，103 個の音源とその位置 (青点) が検出され，これらをクラスタリングして 5 点の音源位置が推定された．この枠組では，音源信号の無音区間により音源追跡に失敗し，短い音源として検出されているため，三角測量の精度が低く，真値から大きく離れた位置を持つ音源が多く検出されている．これらをクラスタリングした結果，位置推定誤差は 0.797 m に留まっている．提案法による結果 (図 1-(c)) では，9 個の音源位置が推定され，2 個を除き，4 つの目的音源から 1 m 以内に定位されている．位置推定誤差は 0.104 m であった．部屋の中心付近に，音源位置から大きく外れた音源が 2 つ検出されているが，これらは拡散性雑音や目的音源の残渣成分と考えられる．提案法は，録音信号全体を一挙に直接分解するため，より安定して音源位置を推定できている．一方で提案法においても，音源 $n = 3$ の正解位置に複数個の推定結果 (黒 ×) が重なっ

ているように，単一の音源を複数の音源として検出してしまっている．これは，同じ音源でも大きく異なるアクティベーションパターンを持つ基底を，単一の音源としてクラスタリングできていないためと考えられる．本問題は，NMF による音源分離と後段の音源位置推定を一挙に最適化する拡張により，互いの情報を相補的に利用することで解決できると考えられる．

4 おわりに

本稿では，NMF による信号分解を基盤とした音環境マッピングについて述べた．従来法の多くは，混合音から各音源の到来方向を推定する音源定位に強く依存した構成で，性能劣化の原因になっていた．そこで本研究では，初段で NMF による音源分離を適用することで，拡散性雑音や残響のある環境でも安定した動作を実現する．数値混合音を用いた評価実験では，HARK により構築した音源定位に基づくマッピングより高い精度で音源位置を推定できていることを確認した．今後は，音源分離と音源位置推定の同時最適化による音源クラスタリングの性能向上や，オンライン推論によるリアルタイムマッピングへの拡張などを進める．

謝辞 本研究の一部は，JST ACT-X 数理・情報のフロンティア JPMJAX200N の支援を受けた．

参考文献

[1] Y. Sasaki, S. Kagami, and H. Mizoguchi, “Multiple sound source mapping for a mobile robot by self-motion triangulation,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006, pp. 380–385.

[2] Y. Sasaki, R. Tanabe, and H. Takernura, “Online spatial sound perception using microphone array on mobile robot,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2478–2484.

[3] J. Even, Y. Morales, N. Kallakuri, J. Furrer, C. T. Ishi, and N. Hagita, “Mapping sound emitting structures in 3D,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 677–682.

[4] S. Michaud, S. Faucher, F. Grondin, J.-S. Lauzon, M. Labbé, D. Létourneau, F. Ferland, and F. Michaud, “3D localization of a sound source using mobile microphone arrays referenced by SLAM,” *arXiv preprint arXiv:2007.11079*, 2020.

[5] C. Schymura and D. Kolossa, “Potential-field-based active exploration for acoustic simultaneous localization and mapping,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 76–80.

[6] D. Su, T. Vidal-Calleja, and J. V. Miro, “Towards real-time 3D sound sources mapping with linear microphone arrays,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1662–1668.

[7] F. Grondin and J. Glass, “SVD-PHAT: A fast sound source localization method,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 4140–4144.

[8] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in *Microphone arrays*. Springer, 2001, pp. 157–180.

[9] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, “Intelligent sound source localization for dynamic environments,” in *IEEE/RSJ international conference on Intelligent Robots and Systems (IROS)*, 2009, pp. 664–669.

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

[11] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[12] M. D. Hoffman, D. M. Blei, and P. R. Cook, “Bayesian nonparametric matrix factorization for recorded music,” in *International Conference on Machine Learning (ICML)*, 2010, pp. 1–8.

[13] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[14] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, “A comparison and evaluation of multi-view stereo reconstruction algorithms,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 519–528.

[15] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM,” *IEEE Transactions on Robotics*, pp. 1–17, 2021.

[16] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[17] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[18] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, “General-purpose tagging of Freesound audio with audioset labels: Task description, dataset, and baseline,” in *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2018, pp. 69–73.

[19] 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.

[20] J. Garofalo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) complete,” *Linguistic Data Consortium, Philadelphia*, 2007.

[21] 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.

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-058-09

Detecting earthquakes: a novel deep learning-based approach for effective disaster response

Muhammad Shakeel[*1] Katsutoshi Itoyama[*1] Kenji Nishida[*1] Kazuhiro Nakadai[*1*2]

[*1] Tokyo Institute of Technology [*2] Honda Research Institute Japan, Co., Ltd.

This research provides an efficient earthquake event classifier that aims to aid robots in automating the conventional disaster response process. Additional sensors and automation are constantly required to react efficiently to a crisis scenario. Deep learning has shown effectiveness in a wide range of applications having a low signal-to-noise ratio, which encouraged us to present a unique 3-dimensional convolutional recurrent network-based earthquake detection method to demonstrate its efficacy in real-time implementation. We train the network using a publicly available earthquake dataset and perform ablations on real-time collected event samples. We preprocess the raw earthquake signals using Log-Mel-based features extraction to retrieve spatial and temporal information. The model extracts the feature information from the low-frequency seismic signals. Furthermore, we propose implementing the model in real-time to distinguish major and minor tremors from seismic signals with an accuracy, sensitivity, and specificity of 98%, 97.7%, and 99.79%, respectively, and a probability threshold of 0.7. Additionally, we develop and validate the model using a two-month continuous data stream from a laboratory-based personal seismometer. The method reliably detects all 63 strong earthquakes recorded by the Meteorological department in Japan from November to December 2019.

1. INTRODUCTION

Over the past decade, tremendous progress has been made in Search, Rescue, and Disaster Robotics[1], and several revolutionary technologies have been developed to enable efficient disaster response. However, considerable effort has to be made in this field to mitigate the effect and destruction caused by natural catastrophes that go beyond human perception. Earthquake recognition continues to be a key and significant aspect for successful crisis response, and the proposed study is an important step forward in the area of "Disaster Robotics".

Identifying earthquakes is a challenging research subject, and a reliable classification method is required to distinguish earthquake waveform from seismological noise. There is a strong potential in using deep learning models in earth observation to describe and identify earthquakes (e.g. [2, 3]) effectively. Effective implementation of these methods is relatively dependent on the availability of high-quality datasets. To address this issue and expedite exploration in this discipline, a worldwide collection of seismic data for machine learning applications has been provided recently. Forming a new dataset, such as STEAD[4], is used in various ways to assist in the development of technologies for the seismological industry. It can also benefit the robotics field in the coming future for efficient disaster mitigation.

Deep learning emerged as a noteworthy area in the field of artificial intelligence and has evolved in various disciplines, including speech recognition, computer vision, and computational linguistics. Availability of enormous processing capabilities[5, 6, 7, 8, 9, 10, 11], deep learning models, implemented as convolutional neural networks (CNN), have demonstrated considerable advances in various classification-related tasks and obtained promising outcomes in a variety of tasks. Thus, an effective learning algorithm

Contact: Muhammad Shakeel, Tokyo Institute of Technology, 2-12-1-W8-30 Ookayama, Meguro-ku, Tokyo, 152-8552, JAPAN, E-mail: shakeel@ra.sc.e.titech.ac.jp

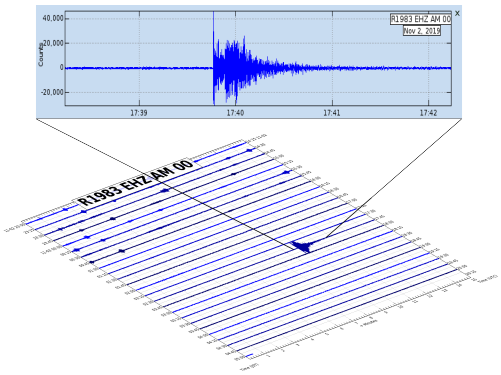


Figure 1: Earthquake detected, by a model learned using 3D-CNN-RNN architecture, within a stream of real-time data available from personal seismometer. The earthquake was reported having a Magnitude of 3.9 in the Japan Meteorological Agency database on November 02, 2019 17:39:29 UTC

is intended to recognize poor signal-to-noise ratio occurrences in a seismograph compared to traditional techniques to automate and enhance the earthquake identification method. With a significant fraction of historical raw earthquake waveforms availability and its promising application area in terms of "Disaster Robotics," we decided to employ deep learning-based algorithms such as convolutional and recurrent neural networks to identify and classify earthquakes. It is possible to convert the seismic signal into a Log-Mel spectrogram. These time-frequency transitions are used as a two-dimensional input to the CNN to extract suitable features from the data.

The three major components of our suggested technique are as follows: Firstly, we generate our custom dataset using the 1.2 million waveforms currently accessible in STEAD. Secondly, we train a model to identify seismic events using a 3-dimensional convolutional recurrent architecture to enhance its accuracy further. Thirdly, we implement and test the system on real-time data trans-

mitted via a personal seismometer to analyze the performance of our classification network. This approach will result in the first implementation of an earthquake monitoring system powered by artificial intelligence for emergency preparedness in the robotics domain. The primary benefit of this technology is that it may be deployed in an environment where robots can be engaged in real-time in the event of a potentially major seismic event. In summary, we assert that our technique has made the following contributions:

1. Propose a three-dimensional (3D) convolution framework for the identification of seismic signals. In a conventional convolutional recurrent network, feature maps are layered to capture the optimum amount of information from the input data; however, we combine separate RNNs on every filter of the last convolutional layer in this study;

2. This is the first study to use a feature extraction technique based on Log-Mel spectrograms to seismic waveforms to the best of our knowledge;

3. Performance evaluation of the system using triangular-shaped filters set to 60 in Log-Mel spectrograms;

4. Extensive analysis of real-time data obtained through a personal seismometer: an earthquake detecting gadget[12].

In this section, we briefly review closely related approaches.

Conventional Methods. Due to their simplicity, STA/LTA[13, 14] and template matching[15, 16] are often used approaches for event identification in seismology. STA/LTA detects earthquakes by comparing short-term average energy to long-term average energy. However, in difficult circumstances with a poor signal-to-noise ratio and time-varying background noise, it generates many false detections. Template matching is a technique for detecting anomalies in candidate waveform data that needs previous knowledge of the candidate waveform data. Cross-correlation algorithms employed in template matching are inefficient and lack generality to handle data in real-time. Both of these systems have a low signal-to-noise ratio and a high false-positive rate, making them impractical for real-time applications, particularly disaster response, where high accuracy is the primary goal.

ConvNetQuake[17]. Convolutional neural networks (CNNs) for seismic data have lately gained popularity as a means of overcoming the limits of traditional techniques. Even though ConvNetQuake is built on the deep topology of a convolutional neural network[9], it is learned on unprocessed waveforms. It does not include feature engineering, which is critical for extracting spatiotemporal information from seismic data. The 2D-CNN framework[9] used in ConvNetQuake serves as a feature extractor in various classification-related tasks, and it is widely recognized as being superior to hand-crafted features. Because of this, a spectrogram of a seismic signal (an intermediary representation of the seismic signal) is required as a two-dimensional input in leveraging the high-dimensional information. In this study, dense CNN models demonstrate good detection accuracy since simulated noisy data was combined with actual data to improve accuracy; however, we demonstrate that state-of-the-art performance may also be obtained if convolutional networks are utilized sensibly alongside recurrent networks on entire real data.

CRED[18]. In this latest research, earthquake detection is treated as a sequence-to-sequence learning problem[19], and two-dimensional convolutional layers are organized in residual blocks, as described in [6], to optimize feature extraction. The authors conducted a comprehensive review of CRED and compared it to established techniques such as STA/LTA and template matching. Their method recognized three orders of magnitude more events than STA/LTA and reduced the false positive rate, demonstrating the deep learning architecture's efficacy and reliability. However, in this research, two-dimensional convolutional neural networks are layered with RNN(Bi-LSTM) layers to extract local features and model hidden temporal relationships, respectively. In general, super deep CNN architectures[6],[9] outperform regular CNN models. However, expanding the receptive area of a 3-dimensional convolutional recurrent network by extracting spectral and temporal feature maps may also give state-of-the-art results.

2. PROPOSED METHOD

We propose a three-dimensional (3D) convolutional architecture for earthquake detection and expand the usage of Log-Mel spectrograms to extract features from seismic signals to attain greater temporal and spatial resolution. The 3D-CNN architecture is employed in various fields, including human action detection[20] for video processing applications, audio-visual recognition[21], and, more lately, speaker verification tasks that do not need text[22]. We introduce a 3D-CNN-RNN framework in this study to leverage the temporal and spatial characteristics of seismic waveforms. In comparison, a traditional CNN-RNN architecture stacks feature maps together. In contrast, we implement distinct RNNs to every filter of the final convolution operation to capture most temporal information from the seismic waves. Similar to ConvNetQuake and CRED, we use Log-Mel energies to balance frequency and temporal characteristics.

2.1 3D CONVOLUTIONAL NEURAL NETWORKS

In principle, the 3D-CNN is the expansion of 2D-CNN. When 2D-CNNs are employed on 2D feature maps we can only extract the information in spatial domain. In the case of seismic events, if two events occur at the same time it is most desirable to extract the temporal information related to actual earthquake signal. The said information can only be inferred in temporal domain to capture the changing behaviour of the signal. To tackle the aforementioned issue, we propose to perform 3D convolutions in convolution stages which is desirable in computing the features from spatial and temporal perspective. In theory, 3D convolution is performed by convolving a 3D kernel (filter) and stacking multiple adjacent frames in a form of cube. In this topology, stacked frames in the previous layer are connected to feature maps in the convolution layer, thereby capturing temporal information. In formulation, the value of any unit at position (x,y,z) in the jth feature map in the ith layer, denoted as $u_{ij}^{x,y,z}$, is given by

$$u_{ij}^{x,y,z} = g\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{p,q,r}\, u_{(i-1)m}^{(x+p)(y+q)(z+r)}\right), \quad (1)$$

where g is the activation function, b_{ij} is the bias for the feature map, R_i is the size of the 3D kernel along the time axis, $w_{ijm}^{p,q,r}$ is the (p,q,r)th value of the kernel connected to mth feature in the previous layer.

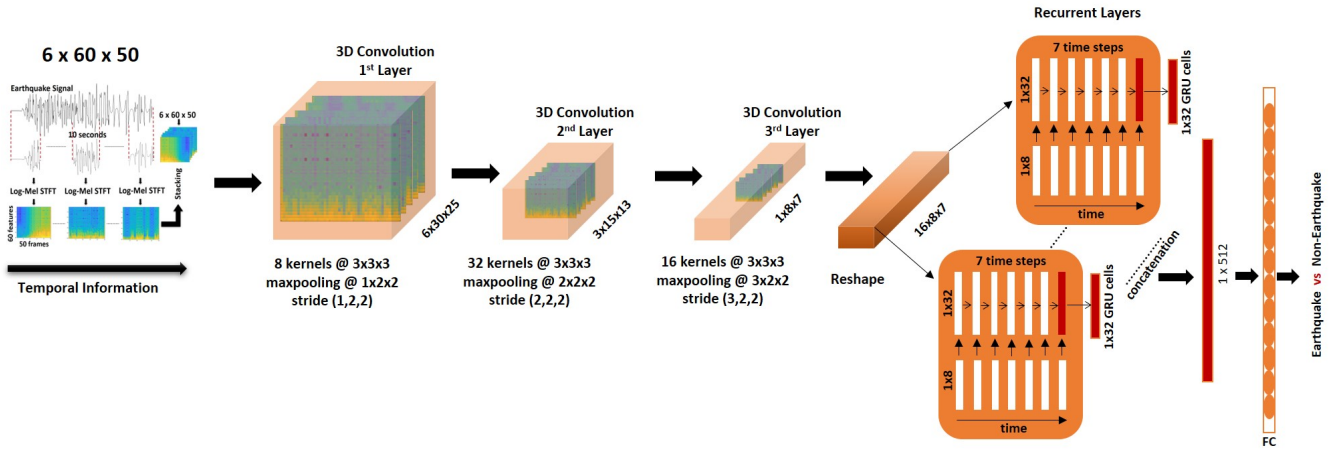


Figure 2: 3D-CNN-RNN combination for earthquake detection. Combination of three convolutional layers and sixteen separate GRUs for each filter in the final convolutional layers is used in above architecture. Each feature map of the last layer is fed to 32 GRU cells in the sixteen recurrent layers. Softmax output layer acts as a fully connected (FC) last layer to classify the events in to earthquake and not-earthquake. Input is a stack of 10-second ground motion clips.

2.2 RECURRENT NEURAL NETWORK (RNN)

The RNNs are important in sequence-to-sequence learning tasks as they retain relations among inputs while training. In practical applications[23] *gated* RNNs, also known as gated recurrent units or GRUs, are used most effectively because they have the derivatives that neither vanish nor explode while creating paths through time. In many sequential tasks GRUs are used because they have the capability of simultaneously controlling the forgetting factor and the decision to update the state unit with a single gating unit. The update equations[24] for GRUs are as follows:

$$h_i^{(t)} = u_i^{(t-1)} h_i^{(t-1)} + (1-u_i^{(t-1)})\sigma\left(b_i + \sum_j U_{i,j} x_j^{(t-1)} + \sum_j W_{i,j} r_j^{(t-1)} h_j^{(t-1)}\right), \quad (2)$$

where u stands for the update gate and r for the "reset" gate. Their value is separately defined as:

$$u_i^{(t)} = \sigma\left(b_i^u + \sum_j U_{i,j}^u x_j^{(t)} + \sum_j W_{i,j}^u h_j^{(t)}\right), \quad (3)$$

$$r_i^{(t)} = \sigma\left(b_i^r + \sum_j U_{i,j}^r x_j^{(t)} + \sum_j W_{i,j}^r h_j^{(t)}\right), \quad (4)$$

In GRUs reset and update gates can independently "ignore" some parts of the state vector making it dynamically control the time scale and forgetting behaviour of different units.

2.3 3D CNN-RNN ARCHITECTURE

A range of CNN-RNN topologies may be designed using the 3D convolution and Recurrent Neural Network methods described earlier. We explain a 3D-CNN-RNN framework that we constructed for an earthquake classification task in the next section. Three convolutional layers are used in this design, as seen in Fig.2. We suggest a $3 \times 3 \times 3$ field of view (3×3 in the space, 3 in the time axis). Furthermore, we apply max pooling operation in each convolutional layer, such as: $(1 \times 2 \times 2)$ for the first, $(2 \times 2 \times 2)$ for the second, and $(3 \times 2 \times 2)$ for the final layer. Strided convolutions combined with a max-pooling layer operation enable us to downsample the signals through each dimension, lowering the computational complexity while rapidly increasing the field of view across the original signal. To retrieve temporal and spatial details, indirect connections are employed. Moreover, we add the ReLU (rectified linear unit) activation function $g(.) = max(0,.)$ to each convolutional layer, using its backpropagation rule to cancel out any gradient elements that are smaller than zero. To optimize the learning rate, batch normalization[25] is performed individually in each layer. To prevent overfitting, a dropout rate of 0.5 is employed. The Xavier initialization[26] technique is adopted for randomly initializing the training weights. We apply several GRUs to the final convolutional layer's feature maps to extract temporal information from seismic waves. Since the final layer has sixteen filters (kernels), each filter is represented by a distinct GRU, resulting in sixteen GRUs. We built seven recurrent layers among each feature map, where seven is the number of time steps mapped from the 50 timestamps in the original spectrogram. The recurrent network comprises 32 GRU cells per layer. Each recurrent layer employs a many-to-one structure, and the output of all layers is concatenated and then fed into a fully linked layer. Using the backpropagation technique, optimization is performed simultaneously on 3D-CNN and RNN architectures. As a final layer, a fully connected softmax layer comprising two nodes is implemented to categorize occurrences as earthquakes or non-earthquake.

3. DATA AND METHODS

3.1 PROPERTIES OF DATASET

Earthquakes occur when rapid movements across active faults release stored elastic energy in the rocks, generating shock waves that flow through the ground. Each day, thousands of earthquake events occur worldwide, of which fifty are intense enough to be experienced (magnitude > 2.5)[27]. These seismic signals are recorded at local seismic stations, and there was a need for a single

universal database. To anticipate the potential challenge, Stanford researchers have released a database featuring seismic waves from throughout the world from January 1984 to August 2018. STanford Earthquake Dataset (STEAD)[4] is a publicly accessible online database for research purposes. It is classified into two types: localized earthquakes and seismic noise (free of earthquake signals). It comprises *sim* 1,050,000 three-component seismograms, with 6000 samples per waveform in the east-west, north-south, and vertical directions. However, we selected the vertical component of the waveform since our model would be validated using a real-time personal seismometer that can only monitor the vertical component. Each earthquake event contains 32 properties, one of which is 'source magnitude,' which is critical for balancing the dataset. The majority of earthquake waveforms presented in the database have a magnitude of < 2.5. We utilized only manually selected waveforms, i.e., those provided by seismic stations, and ignored any waveforms determined by computerized algorithms. We picked distinct waveforms for the training and test sets. All waveforms (earthquake and non-earthquake) are classified according to their stated properties. The waveforms in the given dataset have been detrended (i.e., the mean has been removed), resampled at 100 Hz, and then filtered using a 1-45 Hz bandpass filter. To address data disparity and increase generalization, we omitted specific waveforms and created our dataset. There are 108,680 and 46,561 waveforms used to train and test the model, respectively. The dataset for training and test set makes up a composition of 70% and 30% independently. A one-hot encoding of the training and test set is performed, and each waveform is labeled with 1 if an earthquake event is present and 0 if no earthquake event (seismic noise). The statistics for the development and evaluation sets are presented in Table 1 and 2.

Table 1: Dataset orientation for Earthquake Waveforms.

Earthquake Magnitudes	**Earthquake Waveforms** (Training Set)	**Earthquake Waveforms** (Test Set)
> 0	10868	4656
> 1	10868	4656
> 2	10868	4656
> 3	10868	4656
> 4	9923	4252
> 5	883	378
> 6	61	26
> 7	1	0
Total	54340	23280

3.2 DATA REPRESENTATION: FEATURE EXTRACTION

We propose that Mel Spectrograms be employed as a data representation of seismic waveforms at the frame level. Mel-spectrograms are constructed by incorporating linearly spaced triangular-shape filters in the Mel scale. Further, we obtain the log-energies in the Mel scale. The method is very similar to MFCCs, except that the Discrete Continuous Transform (DCT) is not used in this case. We apply this approach to seismic data to extract frequency components by applying triangular-shape filters while maintaining the maximum temporal information. We divided 60-second ground motion data into six 10-second segments. Overlapping windowed signals represent the temporal features. Using the sequence of these window signals, a single spectrogram of a ten-second clip is generated. The signal is framed using a 400ms window length. Using a Fast Fourier Transform (FFT) with 64 bins (zero padded) and a hamming window with a 50% signal overlap, we can calculate a Short-Time Fourier Transform (STFT) with zero padding. The complex spectrum of a seismic signal $s(t)$ may be expressed as follows:

Table 2: Dataset orientation for seismic noise waveforms.

Non-Earthquake Waveforms (Training Set)	**Non-Earthquake Waveforms** (Test Set)
54340	23281

$$S(n,f) = |S(n,f)|e^{j\theta(n,f)} \tag{5}$$

where $|S(n,f)|$ is the magnitude and $\theta(n,f)$ as the phase spectrum for frequency f in frame n.

Mel-scale is extensively used in speech recognition tasks as one of the feature extraction method. However, we propose this scale can also be utilized for seismic signals. Several analytical expressions exist to convert Hertz-scale frequencies to Mel-scale and one of the common relation as given by D. O'Shaughnessy is used in our study to extract features for network training.

$$m = 2595\log_{10}(1 + f/700) \tag{6}$$

and filter bandwidths computed using,

$$f = 700(10^{m/2595} - 1) \tag{7}$$

Linearly spaced triangular-shape filters in Mel-scale are constructed using the aforementioned equation. The number of filters are set to 60, that act as spectral features for our problem. Finally the magnitude values are then converted into log magnitudes and were normalized as input to the network.

$$S(n,f) = \log(|S(n,f)|) \tag{8}$$

Each feature map has the dimensionality of $\delta \times 60 \times 50$. δ is the number of seismic signal clips, 60 are the number of filters in Mel-scale and 50 is the window frames used to calculate the STFT. Finally the input feature for 3D-CNN RNN architecture is $6 \times 60 \times 50$ and feature extraction process as employed on raw seismic waveform is shown in Fig.3.

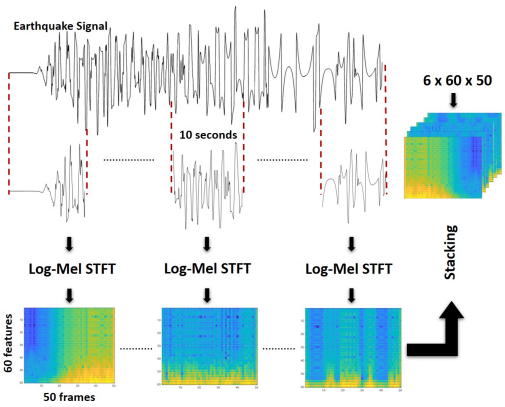


Figure 3: Data input process for 3D-CNN RNN network: Feature extraction using Log-Mel spectrograms

4. EXPERIMENTS

4.1 EVALUATION METRICS

We report the performance evaluation metric of our earthquake detector classifier in terms of Precision (sensitivity), Recall (specificity) and F-score (accuracy) using (9), (10) and (11) respectively. Sensitivity is defined as the number of earthquakes predictions that are accurate, Specificity is defined as the number of instances that are accurately predicted, and the F-score is the harmonic mean of sensitivity and specificity. We calculate these scores using a decision threshold value (thresh) for output probabilities.

$$Precision = TP(thresh)/(TP(thresh)+FP(thresh)) \quad (9)$$

$$Recall = TP(thresh)/(TP(thresh)+FN(thresh)) \quad (10)$$

$$Fscore = 2 \times Precision \times Recall/(Precision+Recall) \quad (11)$$

where, TP denotes true positives, FP denotes false positives, and FN are false negatives. TP=1 and FP=0 in a perfect classifier.

4.2 TRAINING

We used 3D-CNN architecture in all it's essence i.e. 3-dimensional convolution and three CNN layers. We trained the model using a drop-out rate of 0.5. In binary classification problem, as in our study, we employed binary cross-entropy loss on the softmax function. We employed RMSProp optimizer[24] with initial learning rate of 10^{-3} and a momentum decay of 0.9 to avoid the gradient vanishing/exploding issues, and to increase the learning rate. We use batch size of 64 i.e. 64 training examples to train our models. Furthermore, as a training policy, we split the training set into 97% of the total training examples whereas the remainig 3% of the examples are used as a validation set to monitor the validation loss and observe the training process. The data in the training set is shuffled randomly and to overcome the overfitting problem and maintain generalization we stopped the training after 100 epochs. Data augmentation strategy is not applied because of the availability of large amount of data. During testing, we selected the best model having highest accuracy on the validation set and calculated the predictions. Tensorflow is used to implement the model. We train our networks on a single NVIDIA V100 GPU. The learning time for our proposed architecture is 24 hours that includes feature extraction, training, testing and predicting the probabilities.

4.3 DETECTION ACCURACY

The detection accuracy of the algorithm is the percentage of waveforms correctly classified as earthquake or seismic noise. We selected the best model based on the tuned hyperparameters to detect an earthquake event. Regardless of the threshold choice, our earthquake detector successfully detects 22380 earthquake events as catalogued in STEAD and misclassifies 23 of the earthquake waveforms as seismic noise, whereas it correctly classifies 23258 noise events and misclassifies 900 as earthquakes. In summary (see Fig.4), our algorithm predicts 22380 true positives, 23 false negatives, 900 false positives, and 23258 true negatives. Therefore, the sensitivity (fraction of earthquake events that are true events) is 96%, and specificity (fraction of true events correctly detected) is 99.99%. However, with a probability threshold of 0.7, sensitivity of the network is increased to 97.70% and specificity decreased to 99.79%.

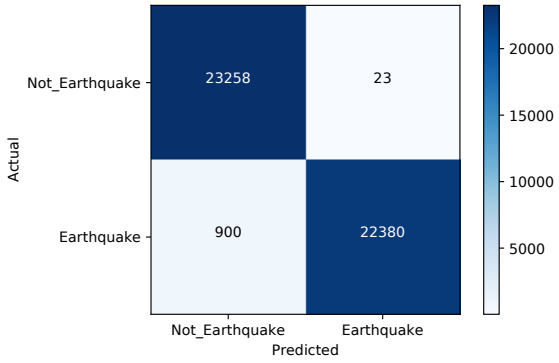


Figure 4: Confusion Matrix Regardless of Threshold Value

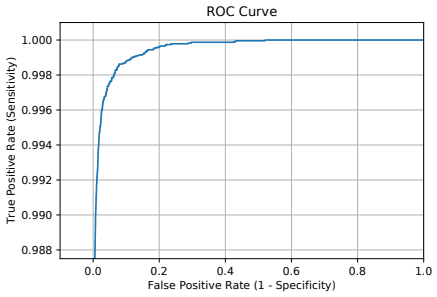


Figure 5: Receiver Operating Characteristics (ROC)

Building a robust deep learning model typically requires a large amount of labelled training data as discussed in **ConvNetQuake**[17] and **CRED**[18]. In general our classifier has superior performance as compared to other two methods. In **ConvNetQuake** authors report the classifier has a precision of 94% and recall of 100% while in **CRED** the model has a precision of 96% and recall of 99%. In **CRED** model is trained using 500000 seismographs (250000 as earthquake events and 250000 as noise waveforms) and with a denser network, whereas in **ConvNetQuake** the network was trained using 702748 waveforms (2709 events and 700,039 noise windows) whereas noise windows were synthetically generated. However, we have demonstrated that state-of-the art results can also be achieved by using a smaller dataset and with a less denser network i.e. only 3 convolutional

layers. Our model is learned using a real and smaller dataset, making it to generalize better in real-time scenarios. The superior performance of our method is due to its reliance on both spectral and temporal feature extraction of the signal rather than the waveform and spectral features only. Hence, denoising the signal as input to the learned model can help reduce the false positive rate as large amount of seismic noise is present in the real-time environment. Furthermore, for fair comparison in general, the datasets for the baseline and proposed methods should have been same but due to non-availibility of datasets used by ConvNetQuake and CRED; we had to train our model using an efficient dataset i.e. STEAD and is publicly available for an acceptable comparison in future studies.

CONCLUSION

We presented a unique autonomous system capable of detecting earthquakes using a learned system based on deep neural networks and a personal seismometer. This breakthrough significantly expands the application fields for artificial intelligence systems, including seismology and disaster robotics. The described method is not dependent on an alert center and may function effectively as an earthquake detection tool in metropolitan areas on its own. While the current study concentrated on detection, future work will examine how artificial intelligence-based algorithms may be utilized to improve the reaction time of an earthquake warning system.

References

[1] S. Tadokoro, Ed., *Disaster Robotics*. Springer International Publishing, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-05321-5

[2] W. Zhu and G. C. Beroza, "PhaseNet: a deep-neural-network-based seismic arrival-time picking method," *Geophysical Journal International*, vol. 216, no. 1, pp. 261–273, 10 2018. [Online]. Available: https://doi.org/10.1093/gji/ggy423

[3] S. Qu, Z. Guan, E. Verschuur, and Y. Chen, "Automatic high-resolution microseismic event detection via supervised machine learning," *Geophysical Journal International*, vol. 218, no. 3, pp. 2106–2121, 06 2019. [Online]. Available: https://doi.org/10.1093/gji/ggz273

[4] S. M. Mousavi, Y. Sheng, W. Zhu, and G. C. Beroza, "Stanford earthquake dataset (stead): A global data set of seismic signals for ai," *IEEE Access*, vol. 7, pp. 179 464–179 476, 2019.

[5] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[10] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[11] R. B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.08083, 2015. [Online]. Available: http://arxiv.org/abs/1504.08083

[12] J. Diaz, M. Schimmel, M. Ruiz, and R. Carbonell, "Seismometers within cities: A tool to connect earth sciences and society," *Frontiers in Earth Science*, vol. 8, Feb. 2020. [Online]. Available: https://doi.org/10.3389/feart.2020.00009

[13] R. Allen, "Automatic phase pickers: Their present use and future prospects," *Bulletin of the Seismological Society of America*, vol. 72, no. 6B, pp. S225–S242, 12 1982.

[14] M. Withers, R. Aster, C. Young, J. Beiriger, M. Harris, S. Moore, and J. Trujillo, "A comparison of select trigger algorithms for automated global seismic phase and event detection," *Bulletin of the Seismological Society of America*, vol. 88, no. 1, pp. 95–106, 02 1998.

[15] S. J. Gibbons and F. Ringdal, "The detection of low magnitude seismic events using array-based waveform correlation," *Geophysical Journal International*, vol. 165, no. 1, pp. 149–166, 04 2006. [Online]. Available: https://doi.org/10.1111/j.1365-246X.2006.02865.x

[16] D. R. Shelly, G. C. Beroza, and S. Ide, "Non-volcanic tremor and low-frequency earthquake swarms," *Nature*, vol. 446, no. 7133, pp. 305–307, 2007. [Online]. Available: https://doi.org/10.1038/nature05666

[17] T. Perol, M. Gharbi, and M. Denolle, "Convolutional neural network for earthquake detection and location," *Science Advances*, vol. 4, no. 2, 2018. [Online]. Available: https://advances.sciencemag.org/content/4/2/e1700578

[18] S. M. Mousavi, W. Zhu, Y. Sheng, and G. C. Beroza, "Cred: A deep residual network of convolutional and recurrent units for earthquake signal detection," *Scientific reports*, vol. 9, no. 1, pp. 10 267–10 267, Jul 2019, 31311942[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31311942

[19] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *CoRR*, vol. abs/1409.3215, 2014. [Online]. Available: http://arxiv.org/abs/1409.3215

[20] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[21] A. Torfi, S. M. Iranmanesh, N. M. Nasrabadi, and J. M. Dawson, "Coupled 3d convolutional neural networks for audio-visual recognition," *CoRR*, vol. abs/1706.05739, 2017. [Online]. Available: http://arxiv.org/abs/1706.05739

[22] A. Torfi, N. M. Nasrabadi, and J. M. Dawson, "Text-independent speaker verification using 3d convolutional neural networks," *CoRR*, vol. abs/1705.09422, 2017. [Online]. Available: http://arxiv.org/abs/1705.09422

[23] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 2146–2153.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. [Online]. Available: http://proceedings.mlr.press/v9/glorot10a.html

[27] P. M. Shearer, *Introduction to Seismology*. Cambridge Univ. Press, 2009.

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-058-10

距離スペクトルを用いた平板形状の認識について

Planar Object Recognition based on Distance Spectrum

公文 誠[1*]　福永 陸人[1] 中妻 啓[1]
Makoto Kumon[1],　Rikuto Fukunaga[1],　Kei Nakatsuma[1]

[1] 熊本大学
[1] Kumamoto University

Abstract: 本研究では送受信波の干渉による定在波に基づく距離スペクトルを利用して環境中の物体表面形状を推定する手法を考える. 距離スペクトルは対象物体を点としてモデル化し, 点までの距離を示す情報を与えるものとして提案されたものであるが, 実際の大きさのある対象では反射表面が影響し, 対象物体の形状に関する情報が埋め込まれている. 本研究では距離スペクトル波形と物体表面の端点との関係を示し, 対象物体表面の大きさの推定法, ならびにこのコンセプトに基づいて複数地点からの観測を統合することで簡単な形状の物体である平板の形状推定へと拡張する. また, 数値実験ならびに屋外実験によって提案法の妥当性を検討・考察する.

1　緒言

ロボットは多様な環境での活動が期待されており, それぞれの置かれた状況を応じ自律的に決定を下す必要がある. このため環境を適切に認識する能力はロボットに必須の基本機能である.

カメラや LiDAR (light detection and ranging) のような光学センサ, 超音波の近接センサやソナー (sound navigaion and ranging; SONAR) などの音響センサなどがロボットの環境情報の認識に活用されている. 音響センサは光学式のそれと比べて分解能に制限があることが多いものの, 悪条件の環境であってもロバストな認識が可能で, 光学センサと組み合わせる相補的な応用が期待される.

超音波センサは波長が短く指向性の高い信号によって精度良く距離測定出来る一方, 減衰が大きく数 m 程度の測定レンジに限られる. 一方, 可聴音 (帯域 20〜20kHz 程度) は超音波に比べ波長が長く減衰が小さいため, より長い伝搬距離を得ることが出来る. 可聴音を用いて環境認識の好例として,「クリック音」（舌打ち音）を用いてヒトが環境中の物体を認識する能力が挙げられる [1]. 熟練した「クリック音」使用者は 30m 以上 [2] の認識距離があると言われ, 単に物体までの距離だけでなくその形状を知覚する [3].

このような可聴音を用いた認識技術の一つとして, Uebo[4] は, 送出信号と対象からの反射信号の干渉による定在波のパワースペクトル（距離スペクトル）が周波数領域において周期的であることを利用した対象との距離推定法を提案している. また, 岸波 [5] は距離スペクトルのアプローチを位相情報と組み合わせる [6] ことで複数物体の認識に拡張している.

ところで環境認識には対象物体までの距離だけではなく, 物体の形状も重要な情報である. 例えば超音波距離センサでは反射波の波形から, 対象の平面や角など形状を推定出来る [7]. 可聴音では広がった音波が広い範囲で反射することを考えれば, 反射波あるいは干渉波もある程度広範囲の環境中の対象物体の形状情報を含んでいると考えられる. 本研究では距離スペクトル波形に着目し, 実際に対象表面の縁までの距離が得られることを明らかにする. また複数地点で計測した情報を統合することで対象の形状を推定する方法へと応用する.

本論文の構成は次の通りである. Uebo[4] の距離スペクトルを用いた対象物体までの距離推定法について 2 節で簡単に紹介した後距離スペクトルと物体表面の形状の関係について 3 節で述べ, 形状推定法を提案する. 提案法の妥当性について, 数値例と実験を通じて議論し (4 節), 最後に 5 節でまとめる

2　可聴音の定在波に基づく距離測定

本節では Uebo[4] の提案する可聴音による距離推定の方法を簡単にまとめる.

受音点（マイクロホン）, 音源（スピーカ）, 静止物体の配置を Fig.1 に示す. L は静止物体までの距離である. 簡単のため信号送出点と受信点は同じ位置であるとする.　時刻 t で距離 x でのスピーカから送出された信号を $v_{\mathrm{Tr}}(t,x)$ と記し, 送出信号を線形チャープ信号

*連絡先： 熊本大学
熊本市中央区黒髪 2-39-1
E-mail: *kumon@gpo.kumamoto-u.ac.jp*

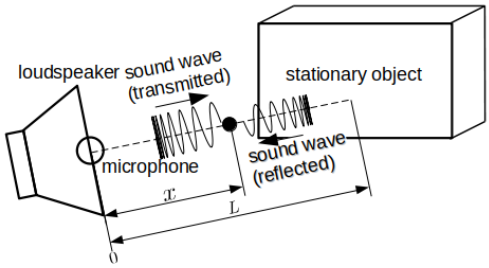


図 1: 定在波を用いた距離推定

であるとすると,

$$v_{\mathrm{Tr}}(t,x) = Ae^{j(2\pi\int_0^{t-|x|/c} f(\tau)\mathrm{d}\tau+\theta)}, \tag{1}$$

と表わせる. ここで $f(\tau)$ は以下で与えられる瞬時周波数である:

$$f(\tau) = \frac{f_w}{T}\tau + f_1, \tag{2}$$

また, A, c, θ は信号の振幅, 音速と初期位相, T, f_1, f_N, f_w は信号のパラメータで信号長, 最低周波数, 最高周波数で帯域は $f_w = f_N - f_1$ の関係がある. 簡単のため音源は原点にあるものとする.

今, 静止物体からの反射波 $v_{\mathrm{Ref}}(t,x)$ は

$$v_{\mathrm{Ref}}(t,x) = A\gamma e^{j\phi}e^{j(2\pi\int_0^{t-(2L-|x|)/c} f(\tau)\mathrm{d}\tau+\theta)}, \tag{3}$$

と表せるとする. ここで, γ と ϕ は対象での反射係数と位相変化であり, 周波数によらず一定とする. 受音点位置 $x=0$ において観測される合成波 $v_{\mathrm{C}}(t,0) = v_{\mathrm{Tr}}(t,0) + v_{\mathrm{R}}(t,0)$ のパワーは

$$|v_{\mathrm{C}}(t,0)|^2 = |A|^2\left\{1+\gamma^2+ 2\gamma\cos\left(2\pi\frac{f_W}{T}\frac{2L}{c}t - 2\pi\frac{f_W}{2T}\left(\frac{2L}{c}\right)^2 + 2\pi f_1\frac{2L}{c} - \phi\right)\right\}. \tag{4}$$

となる. ここで, 右辺の第 1 項は送信波の成分, 第 2 項は反射波の成分, 第 3 項の cos 項は干渉によって生じた成分に対応する. 干渉成分は周期関数であり, 干渉成分の周波数が静止体の位置 L に比例する. (2) を (4) に代入すれば, パワー v_{C} は以下で与えられる.

$$p(f,0) = |A|^2\left\{1+\gamma^2+2\gamma\cos\left(\frac{4\pi L}{c}f + C_0\right)\right\}, \tag{5}$$

ここで C_0 は干渉項での定数成分を表わす.

$p(f,0)$ から定数成分を取り除いた信号の周波数領域でのパワー密度を距離スペクトルと呼ぶ. 距離スペクトルは (5) の cos の周波数に対応するピークを示し, 対象までの距離 L の情報を与える.

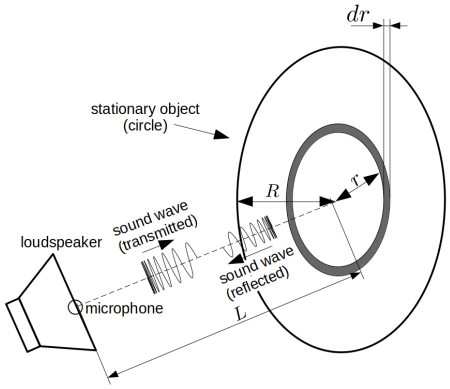


図 2: 円板からの反射

3 距離スペクトルを用いた物体形状推定

3.1 円板モデル

前節で考えた距離スペクトルによる物体検知は対象の最接近点での反射のみ, つまり物体を点として対象を捉えていた. 実際の物体は空間的に広がりを持っており, 表面での反射の影響が距離スペクトルにも影響する. この影響を考察するため, 本節では物体の表面が簡単な円板でモデル化した場合で考察する.

Fig.2 に示す距離 L にある半径 R の円板に向け音波を送出し反射する場合を考える. 簡単のため, 円板の中心を通り, 円板に垂直な直線上に音源と観測点があると仮定する. 角周波数 ω の信号が半径 r, 幅 $\mathrm{d}r$ の円環で反射した際, $x=0$ での信号を $v_{\mathrm{Ref},r,\omega}(t,0)\mathrm{d}r$ と表せば

$$\begin{aligned} v_{\mathrm{Ref},r,\omega}(t,0)\mathrm{d}r &= \gamma\sin\left(\omega(t-\frac{\sqrt{L^2+r^2}}{c})+\phi\right) \\ &\quad\times\frac{L}{\sqrt{L^2+r^2}}\frac{2\pi r}{\pi R^2}\frac{1}{\sqrt{L^2+r^2}}\mathrm{d}r. \end{aligned} \tag{6}$$

と書ける. 円板にわたって (6) の $v_{\mathrm{Ref},r,\omega}(t,0)\mathrm{d}r$ を積分すれば

$$\begin{aligned} v_{\mathrm{Ref},\omega}(t,0) &= \int_0^R v_{\mathrm{Ref},r,\omega}(t,0)\mathrm{d}r \\ &= \frac{2\gamma L}{R^2}\left\{C_1\sin(\omega t-\phi)+C_2\cos(\omega t-\phi)\right\}, \end{aligned} \tag{7}$$

となる. ここで

$$\begin{aligned} C_1 &= \int_0^R \frac{r}{L^2+r^2}\cos(\frac{\omega}{c}\sqrt{L^2+r^2})\mathrm{d}r \\ &= \mathrm{C_i}\left(\frac{\omega}{c}\sqrt{L^2+R^2}\right) - \mathrm{C_i}\left(\frac{\omega}{c}L\right) \end{aligned}$$

$$C_2 = \mathrm{S_i}\left(\frac{\omega}{c}\sqrt{L^2+R^2}\right) - \mathrm{S_i}\left(\frac{\omega}{c}L\right).$$

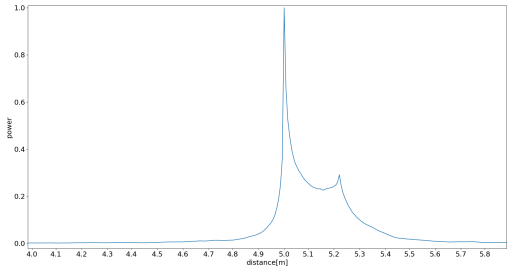


図 3: 距離スペクトル (1.5m radius disk at 5.0m)

であり, $\mathrm{C_i}(\cdot)$ と $\mathrm{S_i}(\cdot)$ はそれぞれ余弦積分と正弦積分を表わす [8].

受信点での送信波 $\sin\omega t$ と反射波 $v_{\mathrm{Ref}}(t)$ を合わせた信号のパワー $P(\omega)$ は,

$$\begin{aligned} p(\omega,0) &= \left[1+\frac{2\gamma L}{R^2}\{C_1\cos\phi+C_2\sin\phi\}\right]^2 \\ &\quad +\left[\frac{2\gamma L}{R^2}\{-C_1\sin\phi+C_2\cos\phi\}\right]^2 \\ &= 1+\alpha^2({C_1}^2+{C_2}^2)+2\alpha(C_1\cos\phi+C_2\sin\phi), \end{aligned}$$

となる. ただし $\alpha=\frac{2\gamma L}{R^2}$.

$\mathrm{C_i}(x)$ と $\mathrm{S_i}(x)$ は $|x|\gg 1$ において

$$\mathrm{C_i}(x)\approx\frac{\sin x}{x}-\frac{\cos x}{x^2},\quad \mathrm{S_i}(x)\approx\frac{\pi}{2}-\frac{\cos x}{x}-\frac{\sin x}{x^2}. \quad (8)$$

と近似できる [9]. $a=\frac{\sqrt{L^2+R^2}}{c}$ と $b=\frac{L}{c}$ $(a>b)$ とおき,

$$\begin{cases} {C_1}^2+{C_2}^2\approx 0 \\ C_1\cos\phi+C_2\sin\phi\approx\frac{\sin(a\omega-\phi)}{a\omega}-\frac{\sin(b\omega-\phi)}{b\omega}. \end{cases} \quad (9)$$

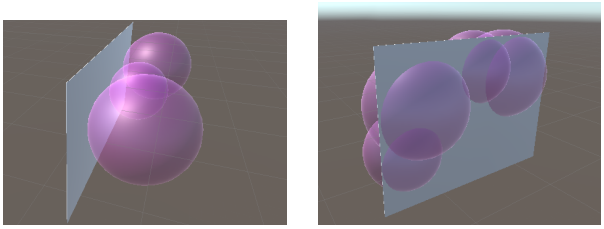
の近似から

$$p(\omega,0)\approx 1+2\alpha\left(\frac{\sin(a\omega-\phi)}{a\omega}-\frac{\sin(b\omega-\phi)}{b\omega}\right). \quad (10)$$

という近似モデルを得る (詳細は付録参照).

(10) より $p(\omega,0)$ はそれぞれ係数 a と b を持つ 2 つの sinc 関数で近似される. このことから距離スペクトルは a と b つまり円板の縁 (a) と距離 (b) に対応する点にピークを持つ.

この近似モデルの妥当性を確認するため, 簡単な音響シミュレーションの結果を示す. シミュレーションでは, 円板から $L=5.0$m 離れた点から半径 $R=1.5$m の円板に音波を放射し, 円板は $\gamma=0.05$, $\phi=\pi$rad の反射係数を持つものとする. この時の距離スペクトルを Fig. 3 に示す. 距離スペクトルは 5.0m と 5.22m にピークがあり, これらは円板までの距離と円板の縁までの距離 $(\sqrt{5.0^2+1.5^2}\approx 5.22)$ に対応している. 詳細は割愛するが, 四角板での数値シミュレーションでも板までの距離と二辺に対応する 3 つのピークが確認されている.



(a) Tangent plane (b) Intersecting spheres

図 4: 距離スペクトルを半径に有する球の共通接平面

3.2 複数観測の統合による形状推定

1 回の計測から求めた距離スペクトルより, 対象までの距離とその辺縁までの距離を推定出来る. 次に, 計測を複数地点で行い, 観測情報を統合することで対象物体の表面形状を推定する方法を考える.

3.2.1 表面検出

距離スペクトルの第 1 ピークは対象までの最短距離を与える. このことから, 対象物体の表面は各計測で得られる距離スペクトルの第 1 ピークで示される距離を半径に持つ球の共通接平面に含まれることになる (Fig. 4(a)).

この共通接平面は以下のようにして求められる. 球 i の中心と半径をそれぞれ O_i と r_i とし, 平面の単位法線ベクトルを n とする. 平面と中心の距離を d と表わすと以下の関係がある.

$$\|O_i-x\|=r_i, \quad (11)$$

$$n^Tx+d=0. \quad (12)$$

接点 x_i は (11) と (12) を満足するので, $\pm r_i+n^TO_i+d=0$ が成立する. この関係を N 個の観測についてまとめると

$$\begin{bmatrix} O_1^T & 1 \\ O_2^T & 1 \\ \cdots & \\ O_N^T & 1 \end{bmatrix}\begin{bmatrix} n \\ d \end{bmatrix}+\begin{bmatrix} \pm r_1 \\ \pm r_2 \\ \vdots \\ \pm r_N \end{bmatrix}=0,\quad |n|=1. \quad (13)$$

となる.

観測雑音などの影響で半径 r_i が正しく得られないことがあり, この結果 (13) は唯一解を持たない可能性がある. そこで, random sample consensus (RANSAC)[10] に倣って N 個の観測からランダムに選んだ観測情報を用いて (13) の部分問題を解き, それらのうちもっとも適合したものを (13) の解として採用する . また上式中の符号の組み合わせも, この過程で解消出来る. なお, 適合の程度には (13) の残差を用いることとした.

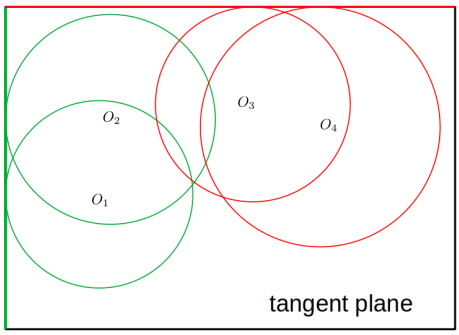


図 5: 接平面と球の交叉円の例

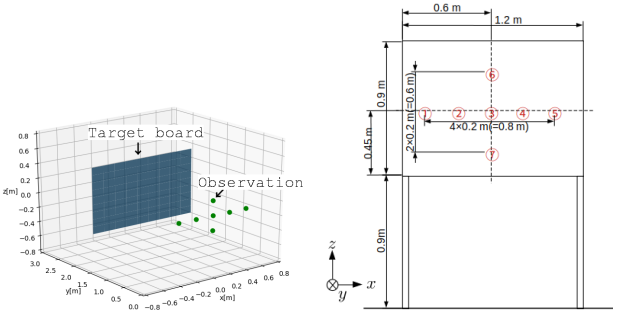


図 6: 対象 (板) と 観測

3.2.2 辺の検出

Fig. 4(b) のように, 距離スペクトルの第 2 ピーク以降に対応する長さを半径とする球が対象表面と交差する円を考えると, これらは上述の対象表面上の辺と接点を持つ (Fig. 5). このことを利用して, 本節では距離スペクトルの第 2 ピーク以降から対象物体表面の辺を推定する方法を考える. 円と辺の接点は唯一に求められないので, この問題は単純には解くことはできないが, 対象平面の辺が十分な数の複数の観測から得られる交叉円の共通接線として得られると期待することは妥当と考え, 本論文ではある直線に対して交叉円の接線とみなせる頻度を解としての適合度とする方法を提案する. 具体的には Hough 変換 [11] のアルゴリズムを利用した以下の方法による.

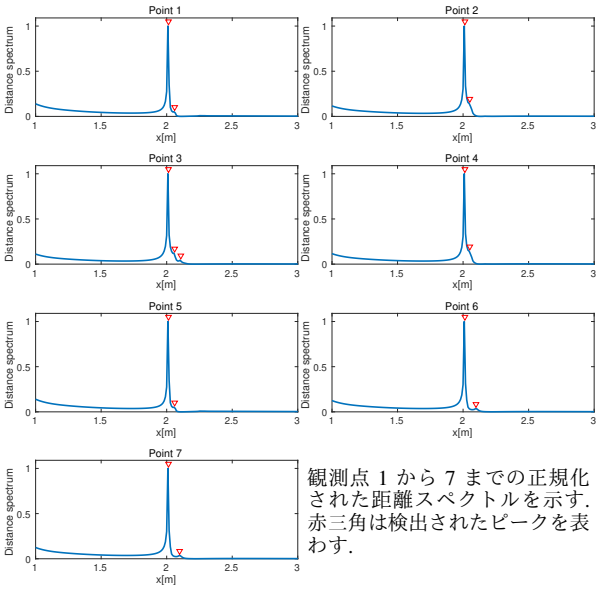
計測系の原点の垂線の足を接平面上の座標系の原点と考え, 交叉円 i の中心を ${}^{t}O_i$, 半径 ${}^{t}r_i$ とし, 接平面上の接線の単位法線ベクトルを ${}^{t}n$ とする. 考えている直線と接平面原点との距離を ${}^{t}d$ とすれば, 接点の関係から ${}^{t}d = {}^{t}O_i^T {}^{t}n + {}^{t}r_i$ を得る. なお, ${}^{t}n$ は接平面内の向き ($\xi \in [0, 2\pi\mathrm{rad})$ とする) で特徴づけることが出来, ${}^{t}n = {}^{t}n(\xi)$ と書けるので ${}^{t}d = {}^{t}d(\xi; {}^{t}O_i, {}^{t}r_i)$ と表わせる.

検出された交叉円の個数を M とし, ξ と d について次の尤度関数を導入する.

$$l(\xi, d) = \sum_{i=1}^{M} \exp\left(-\beta(d - {}^{t}d(\xi; {}^{t}O_i, {}^{t}r_i)^2\right), \qquad (14)$$

ここで β は適当な正定数である. $l(\xi, d)$ のピークのうち閾値 l_{th} を越えたものを考え, これらのピークに対応する直線のパラメータ $\{(\xi, d) | l(\xi, d) \geq l_{\mathrm{th}}, \mathrm{grad} l(\xi, d) = 0\}$ を物体表面の辺を表わす直線の候補として扱うことにする.

得られた接線候補のうち実際の辺に対応する有効区間を特定するため, 交叉円との接点を用いて次にように考える. ξ_j と d_j を直線 j のパラメータとすると, 交叉円 i との接点 ${}^{t}p_{i,j}$ は ${}^{t}p_{i,j} = {}^{t}O_i^T {}^{t}n^{\perp}(\xi_j) n^{\perp}(\xi_j) + d_j {}^{t}n(\xi_j)$ である. ここで ${}^{t}n^{\perp}$ は ${}^{t}n$ に直交する接平面上の単位ベクトルである. このようにして求めた接点のある区間近辺が対象表面の辺の存在する区間とみなす.



観測点 1 から 7 までの正規化された距離スペクトルを示す. 赤三角は検出されたピークを表わす.

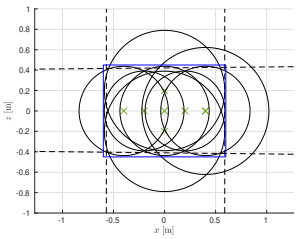
図 7: 距離スペクトル

4 検証

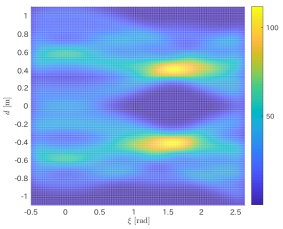
4.1 数値シミュレーション

本節では音響シミュレーションによって提案法の有効性を検証する. 対象物体は Fig. 6 に示す 1.2m × 0.9m の板とし, 20kHz までの帯域を有するチャープ信号を送波子から発し, 48kHz サンプリングで音信号を収録する. 送波子・受波子はともに対象物体から 2.0m の同じ位置にあり, 図に示すように 7 点 での計測を行う. なお, 距離スペクトルのピーク検出には, 最大ピーク（第 1 ピーク）以降の変曲点を用いて極値を探索する方法とした.

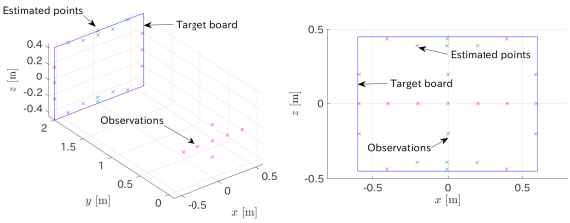
検出された距離スペクトルを Fig. 7 に示す. 図から第 1 ピークが対象までの距離に対応している検出され



(a) Circles on the estimated target surface

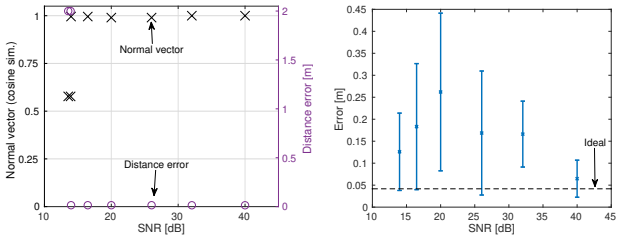


(b) Likelihood of tangent lines



(c) 推定形状; 3D 表示 (左) と 推定平面上への投影 (右). 十字 (赤) が観測点を表わす.

図 8: 物体推定

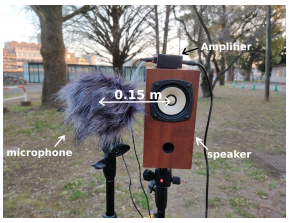


(a) Plane estimation

(b) Shape estimation

図 9: 推定性能

(a) Setup



(b) Sensor device

図 10: 実験環境

ており, それに続いて辺に対応するピークが得られていることが分かる.

3.1 節の手法を適用し, 距離スペクトルの第 1 ピークより対象とする板を含む平面を推定したところ

$$(0.000, 1.000, 0.000)x - 2.014 = 0$$

を得た. さらに, この平面情報を用いて複数観測を 3.2 節の方法によって統合し, 交叉円 (Fig. 8(a)) の共通接線の尤度 (14) を求めた (Fig. 8(b)). 続いて, 尤度の高い接線 (Fig. 8(a) 中の点線) を抽出し, 交叉円との接点を求めた (Fig. 8(c) 中の十字). これらの図より, 提案法が対象の板の辺上の点を検出していることが分かる. 推定された 18 点と辺との距離を誤差として求めたところ, 平均誤差は 0.0417m でその標準偏差は 0.0301m であった.

より現実的な状況を想定し, 測定信号に雑音として一様乱数を加法的に観測信号に重畳した状況でのシミュレーションも行った. あわせて同期加算による雑音抑制機能も導入した.

推定性能の評価として, 推定された対象平面の法線ベクトルの余弦距離と平面と原点との距離の推定誤差を用いた (Fig. 9(a)). Signal-to-noise ratio (SNR) が 14dB より良い条件では法線方向を精度良く推定出来ているため, 距離スペクトルの第 1 ピークが適切に求められていることを示唆している. SNR が 14dB 未満となると, いくつかの第 1 ピークが誤った距離に表れるようになり, 対象表面の法線推定の失敗を引き起こしている.

平面法線が正しく推定できた場合を対象に形状推定誤差を求めた結果を Fig. 9(b) に示す. このことから, 対象平面の法線が求まる範囲では, 対象の形状も適切に推定できていることが分かる.

4.2 実験

次に提案法の実現可能性を検証するために実験を行った. 対象物体はシミュレーションと同じ大きさの板を用いた (実験風景を Fig. 10 に示す). スピーカでチャープ信号を送出し, 携帯型レコーダ (Zoom, H4n Pro) で音信号を収録した. スピーカ・マイク間の距離は 0.15m となるよう設置した (Fig. 10(b)). このスピーカ・マイクの組を用いて複数地点で計測を行い, 提案法を適用した. 環境中の騒音に対応するため, 各地点で 100 回の計測信号を同期加算した.

実験で得られた距離スペクトルを Fig. 11 に示す. 推定された平面は

$$(0.1408, 0.9889, -0.0469)x - 2.009 = 0$$

であり平面の法線ベクトルの真値との余弦距離は 0.9889, 推定距離 d の誤差は 0.009m であった. また, 対象表面の辺として推定された点 (Fig. 12) と辺からの平均誤差, 標準偏差はそれぞれ 0.1009m と 0.0647m であった. 横辺を適切に推定できなかったものの, 上下の辺は良く推定できており, その辺の範囲 (x の線分長さ) は実際の対象板に相当するものであった.

4.3 考察

シミュレーションならびに実験より, 複数地点で計測した距離スペクトルを統合し対象物体表面の形状を求める方法は距離スペクトルの第 1 ピークがある程度良く推定できれば対象表面が精度よく求められること, な

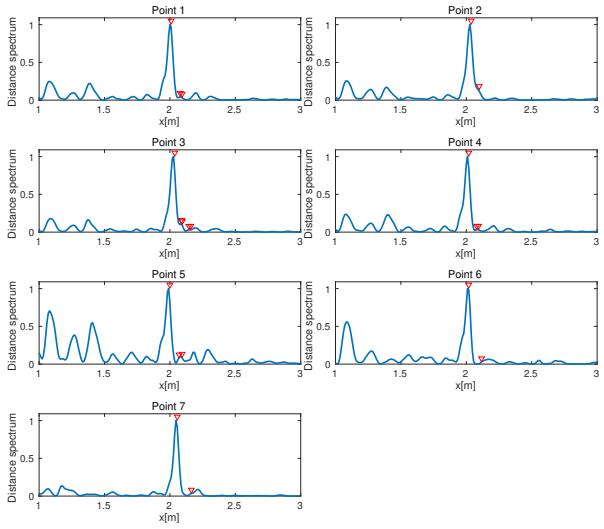
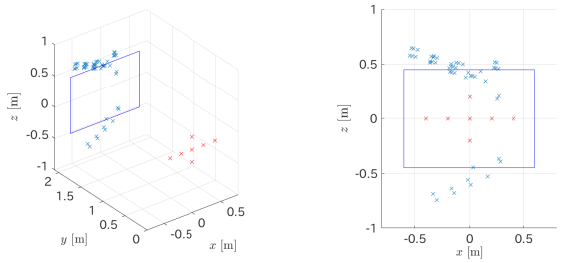


図 11: 距離スペクトル



(a) 3D-view (b) Front-view

図 12: 推定された物体形状

らびに平面が適切に推定されればある程度対象物体表面の周辺形状を推定できることが示された. 一方, 雑音に対しては敏感で同期加算等の音響信号処理によって雑音の影響を低減させることが必要である.

Uebo[4] によれば距離スペクトルの最小分解能は $\frac{c}{2f_w}$ なので, ピーク検出アルゴリズムを考えれば対象までの距離の分解能は $d_{\mathrm{res}} = 2\frac{c}{2f_w}$ となる. 従って, 距離 L にある物体表面の辺までの最小距離が q の場合, 距離スペクトルの L と $\sqrt{L^2+q^2}$ にある二つのピークがこの分解能で検出されるためには $\sqrt{L^2+q^2}-L > d_{\mathrm{res}}$ を満足しなければならない. 可聴域の帯域がおよそ 20kHz であることから, d_{res} はおよそ 8.5×10^{-3}m 以上である. 例えば $L=2$m の距離にある物体の場合なら, q は 0.262m より大きいことが求められ, 今回の実験で対象とした物体はその条件を満足していた. 提案手法はより大きな対象については有効と期待されるが, そのような大きな対象物体については音響系の指向性の影響も考慮することになるだろう.

5 結言

本論文では, 距離スペクトルの複数のピークが対象物体までの距離だけでなく, その表面形状の情報を含むことを示し, この情報を取り出して形状推定を行う一つの方法を提案した. 数値シミュレーションによって提案手法の妥当性を示すとともに, 実験においてある程度の対象の形状を推測できることを確認した.

雑音下での距離スペクトルのロバストな算出法とこのような不確かさを含む距離スペクトルから対象の特徴に対応するピークを検出する方法についてはより詳細な検討が必要である. また, 室内のように複雑で多くの反射を含む環境への拡張は将来の課題の一つである.

本研究は科学研究費補助金 19H00750 の支援を受けたものです.

付録

余弦積分と正弦積分の級数展開による近似 $\overline{\mathrm{C_i}}(x)$ と $\overline{\mathrm{S_i}}(x)$ を考える.

$$\overline{\mathrm{C_i}}(x) = \frac{\sin x}{x} - \frac{\cos x}{x^2}, \quad \overline{\mathrm{S_i}}(x) = \frac{\pi}{2} - \frac{\cos x}{x} - \frac{\sin x}{x^2}.$$

従って

$$\begin{aligned}
\overline{\mathrm{C_i}}^2(x) + \overline{\mathrm{S_i}}^2(x) &= \frac{\pi^2}{4} + \frac{1}{x^2} + \frac{1}{x^4} - \pi\left(\frac{\cos x}{x} + \frac{\sin x}{x^2}\right) \\
\overline{\mathrm{C_i}}(x)\overline{\mathrm{C_i}}(y) + \overline{\mathrm{S_i}}(x)\overline{\mathrm{S_i}}(y) &= \frac{\pi^2}{4} \\
&\quad -\frac{\pi}{2}\left(\frac{\cos y}{y} + \frac{\sin y}{y^2} + \frac{\cos x}{x} + \frac{\sin x}{x^2}\right).
\end{aligned}$$

である. これから

$$\begin{aligned}
{C_1}^2 + {C_2}^2 &\approx \frac{1}{(a\omega)^2} + \frac{1}{(b\omega)^2} + \frac{1}{(a\omega)^4} + \frac{1}{(b\omega)^4} \\
&\quad -\frac{2}{ab\omega^2}\sqrt{1 + \frac{a^2+b^2}{(ab\omega)^2} + \frac{1}{(ab\omega)^2\omega^2}}\sin\left((b-a)\omega + \phi_0\right) \\
&\qquad \left(\tan\phi_0 = \frac{1+\frac{1}{ab\omega^2}}{\frac{1}{\omega}\left(\frac{1}{a}-\frac{1}{b}\right)} = \frac{ab\omega^2+1}{(b-a)\omega}\right)
\end{aligned}$$

であり

$$\begin{aligned}
C_1\cos\phi + C_2\sin\phi &\approx \frac{\sin(a\omega-\phi)}{a\omega} - \frac{\sin(b\omega-\phi)}{b\omega} \\
&\quad -\frac{\cos(a\omega-\phi)}{(a\omega)^2} + \frac{\cos(b\omega-\phi)}{(b\omega)^2}.
\end{aligned}$$

となる. 十分に大きな ω では, $\frac{1}{\omega^2} \ll 1$ となるため (9) が得られる.

参考文献

[1] A.J. Kolarik, et al, “A Summary of Research Investigating Echolocation Abilities of Blind and Sighted Humans,” Hearing Research 310, pp.60-68, 2014.

[2] D. Pelegrin-Garcia, et al., “Localization of a Virtual Wall by Means of Active Echolocation by Untrained Sighted Persons,” Applied Acoustics 139, pp.82-92, 2018.

[3] L. Thaler, et al., “Human Echolocators Adjust Loudness and Number of Clicks for Detection of Reflectors at Various Azimuth Angles,” Proc. R. Soc. B, 285.1873, 2018.

[4] T. Uebo, N. Nakasako, N. Ohmata and A. Mori, “Distance Measurement based on Standing Wave for Band-limited Audible Sound with Random Phase,” *Acoustical Science and Technology*, **30**(1), pp. 18-24, 2009.

[5] 岸波華彦, 糸山克寿, 西田健次, 中臺一博, 重み付け尤度関数と定在波を用いた可聴音による二次元環境認識, 日本ロボット学会誌,**39**-3, pp.271-274, 2021.

[6] 高尾麻衣子, 干場功太郎, 中臺一博, 可聴音を用いた周波数選択に基づく距離推定法の実環境利用に向けた評価, 第 49 回人工知能学会 AI チャレンジ研究会予稿集, pp.29-34, 2017.

[7] P. P. Smith, “Active Sensors for Local Planning in Mobile Robotics,” World Scientific Series in Robotics and Intelligent Systems: Vol. 26, World Scientific, 2001.

[8] W.E. Weisstein, “Sine Integral,” Wolfram Web Resource, `https://mathworld.wolfram.com/SineIntegral.html`. (閲覧 2021/02/25)

[9] J. R. Airey, “The converging factor in asymptotic series and the calculation of Bessel, laguerre and other functions,” The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 24.162, pp.522-553, 1937.

[10] M. A. Fischler and R. C. Bolles, “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,” Communication of ACM, 24 (6), pp. 381–395, 1981.

[11] P.V.C. Hough, “Method and Means for Recognizing Complex Patterns,” US Patent US3069654A, 1962.