人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-01

# 遠隔マイクによる日常会話音声認識に向けた取り組み

## Advances on Distant Conversational Speech Recognition

神田 直之 *
Naoyuki Kanda

マイクロソフト
Microsoft Corporation

## 概要

遠隔マイクによって録音された音声から「誰が」「何を」「いつ」話したかを正確に認識する技術は，自動議事録作成や対話ロボットなどの実現のために必須である。しかし遠隔収録された会話音声には雑音や残響が重畳しているだけでなく，複数話者の音声の重複が頻繁に観測されるため，単独話者向けもしくは近接マイク向けに開発された従来の音声認識モデルは適切に動作しない。本講演では，遠隔マイクもしくはマイクアレイを用いてリアルタイムに複数話者の会話を音声認識し話者決定するための各種の取り組みについて紹介する。

*連絡先：Microsoft Corporation
1 Microsoft Way, Redmond, WA, USA
E-mail: Naoyuki.Kanda@microsoft.com

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-02

# Parallel Adapter Model と Near-Identity 初期化を用いた音声認識の雑音耐性向上

# Improving Noise Robustness of Automatic Speech Recognition based on a Parallel Adapter Model with Near-Identity Initialization

大﨑 崇博[1] *　周藤 唯[2]　糸山 克寿[1,2]　西田 健次[1]　中臺 一博[1]

[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:** 本論文では, 音声認識の雑音耐性を少量のトレーニングで改善するために Parallel Adappter Model (PAM) を提案する．音声認識の雑音耐性を向上するために音声強調を使用すると，強調音声による歪みや目的音の情報除去により，認識精度が十分に向上しない．提案手法の PAM では，学習済み音声認識と音声強調に加え，アダプターと呼ばれる小規模ネットワークを適切な初期化手法でモデルに組み込み少量の再学習を行うことで，音声強調と音声認識の親和性を高め，認識精度を向上する．検証実験では，わずか 10 epoch の再学習で，音声認識の精度が 26.1 ポイント向上した．

## 1　はじめに

音声認識（Automatic Speech Recognition, ASR）は, 人間の発話音声をテキストに変換するシステムである．これまでの音声認識は，GMM(Gaussian mixture model) と HMM(Hidden Malkov model) を組み合わせた GMM-HMM モデル [1] や，GMM を DNN(Deep neural network) に置き換えた DNN-HMM モデル [2, 3] が主流であった．しかし，計算機の性能向上や大規模言語コーパスの拡充に伴い，近年では音響モデルと言語モデルを 1 つのネットワークで構成した End-to-end ASR が盛んに研究されている [4, 5, 6, 7]．このような ASR は，一般的に雑音が少ない環境で録音された教師用音声が用いられる．それに対して，ASR を実環境で用いる場合，入力音声には目的音声に加えて背景ノイズが混ざるため, 性能が低下する．特に，工場内やレストラン等の騒がしい場所で ASR を用いる場合には，音声認識の雑音耐性を向上する手法が必要不可欠となる．

ASR の雑音耐性を高める研究はこれまで行われており，代表的なものに音声強調 (Speech Enhancement, SE) を ASR フロントエンドとして用いる手法がある．SE はノイズが混ざった音声から，ノイズのパワーを弱め目的音声のみを強調する技術である．SE 手法にはマイクロホンアレイ処理 [8, 9] や深層学習 [10, 11] を用いたものがあり，近年では単一チャネル入力への適用が可能である深層学習ベースの手法が広く用いられている [12, 13]．しかし，音声強調がノイズ音だけではなく話し声の情報まで除去することや，出力音声に歪みが生じたりすることが原因となり，SE と ASR の単純な組み合わせでは，期待した性能向上が得られないことが知られている．

この歪みによる ASR と SE 間のミスマッチを解消するには 2 つのアプローチがある．1 つ目は，再トレーニングを行わない手法である．Takeda らの研究 [14, 15] では，モンテカルロ推定に基づく特徴量推定により，再学習なしで認識精度を向上している．しかし，推定には，潜在特徴量の繰返し計算が必要であり，時間を要する割に性能向上が小さい．2 つ目は，モデルの再トレーニングによるミスマッチの緩和である．モデル全体を再学習する手法や SE 部のみを再学習する手法が提案されている [16, 17, 18] が，いずれの手法も膨大なパラメータ更新が必要であり，学習コストが大きい．

そこで本研究では，アダプターを使用した少量の再トレーニング手法を提案し，ミスマッチ問題と再学習コスト問題を解決する．アダプターは小規模のネットワークであり，学習済みのモデルと同時に用いる．アダプターが組み込まれたモデルを学習する際は，ASR や SE のパラメータは固定し，アダプターのパラメータのみ更新することで，更新パラメータ数を抑えることができる．また，アダプター学習にはノイズを付与

*連絡先：東京工業大学
〒 152-8552 東京都目黒区大岡山 2-12-1
osaki@ra.sc.e.titech.ac.jp

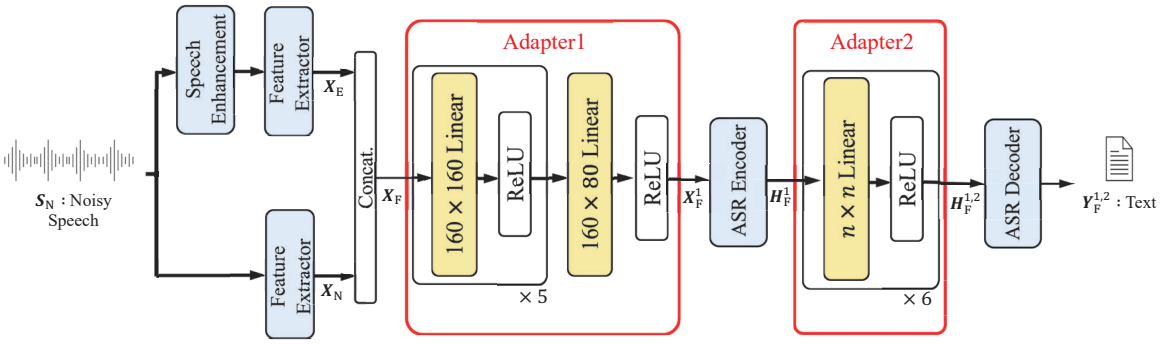



図 1: Parallel Adapter Model (PAM)．Adapter2 の次元数 $n$ は潜在特徴量の次元数に準拠し，100 時間の学習用音声を用いる場合は 256，960 時間の学習用音声を用いる場合は 512 に設定．

した音声を用いることで，ノイズ耐性の高いモデルが学習できる．以上より，少ない再学習量でモデルの耐雑音性向上が期待できる．

## 2 提案手法

本章では，提案手法である Parallel Adapter Model (PAM) とその初期化手法を，ASR や SE と共に説明する．

### 2.1 音声認識 (ASR)，音声強調 (SE)

ASR には，Encoder, Decoder からなる End-to-End 音声認識モデルを用いるものとする．ASR への音声入力を $\boldsymbol{S}$ とすると，まず，$\boldsymbol{S}$ を特徴抽出器 (Feature Extractor, FE) によって音声特徴量 $\boldsymbol{X}$ に変換する．次に，変換した音声特徴量 $\boldsymbol{X}$ を Encoder 部に送出し，潜在特徴量 $\boldsymbol{H}$ に変換，その後，Decoder 部で，テキストに変換する．音声認識モデル学習では，音声 $\boldsymbol{S}$ が入力されたときに正しいテキスト $\boldsymbol{Y}$ が出力される確率が最大になるように，Encoder, Decoder のパラメータ $\boldsymbol{\theta}_{enc}, \boldsymbol{\theta}_{dec}$ の推定を行う．

SE は，雑音が混入した音声 $\boldsymbol{S}_{\mathrm{N}}$ から，雑音を抑圧した音声 $\boldsymbol{S}_{\mathrm{E}}$ に変換する．本稿では，入力音声が単チャネルであることを想定しているため，深層学習ベースの SE 手法を用いる．SE のパラメータ $\boldsymbol{\theta}_{\mathrm{se}}$ は，ASR と組み合わせて用いる前に，クリーン音声と強調音声の類似度が高くなるように事前学習を行うものとする．

### 2.2 Parallel Adapter Model

図 1 に，提案モデルである Parallel Adapter Model (PAM) を示す．このモデルは，SE を適用するフローに加えて，SE を行わない未処理音声を用いるフローを持つ．それぞれのアダプターは，学習可能なパラメータ $\boldsymbol{\theta}_{\mathrm{adp1}}$，$\boldsymbol{\theta}_{\mathrm{adp2}}$ を持った，ASR や SE よりも十分に小さいネットワークである．

アダプター 1 への入力 $\mathrm{Cat}(\boldsymbol{X}_E, \boldsymbol{X}_N)$ は，強調音声と未処理音声の特徴量を結合した特徴量であり，$\boldsymbol{X}_{\mathrm{F}}^1$ が出力される．

$$\boldsymbol{X}_{\mathrm{E}} = \mathrm{FE}(\boldsymbol{S}_{\mathrm{E}}) \quad (1)$$

$$\boldsymbol{X}_{\mathrm{N}} = \mathrm{FE}(\boldsymbol{S}_{\mathrm{N}}) \quad (2)$$

$$\boldsymbol{X}_{\mathrm{F}}^1 = \mathrm{Adapter1}\left(\mathrm{Cat}(\boldsymbol{X}_E, \boldsymbol{X}_N); \boldsymbol{\theta}_{\mathrm{adp1}}\right) \quad (3)$$

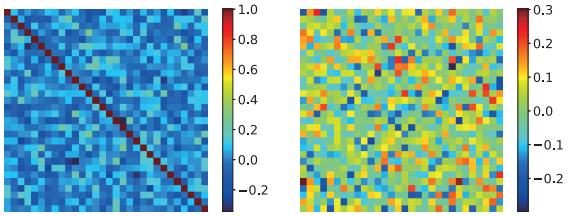
アダプター 2 への入力は Encoder の出力 $\boldsymbol{H}_{\mathrm{F}}^1$ であり，Decoder への入力 $\boldsymbol{H}_{\mathrm{F}}^{1,2}$ を出力する．

$$\boldsymbol{H}_{\mathrm{F}}^{1,2} = \mathrm{Adapter2}\left(\boldsymbol{H}_{\mathrm{F}}^1; \boldsymbol{\theta}_{\mathrm{adp2}}\right) \quad (4)$$

モデルを学習する際は，アダプターのパラメータのみ更新し，ASR や SE のパラメータは固定した状態で，認識精度が最も良くなるようなアダプターのパラメータを学習する．

PAM を構築する上で，アダプターには 3 つの要件，すなわち「選択」「復元」「混合」が求められる．「選択」は，音響特徴量から音声成分を抽出し，雑音成分を除去するとともに，SE による歪みを緩和し，ミスマッチ問題に対処する機能である．「復元」は，SE によって過度に欠落した情報を復元する機能である．例えば，倍音成分同士が持つ強い相関関係を基に，アダプターが損なわれた特徴量成分を復元できると考えられる．「混合」は，アダプター 2 に必要とされる，2 つの特徴量から認識に適した特徴量を生成することで，ASR の性能向上を実現する機能である．

これらのうち，「復元」と「混合」は線形層で，「選択」は活性化関数として ReLU を用いることで実現する．またアダプターの表現力を向上させるために，本稿では，ReLU 関数をもつ線形層を 6 層スタックすることで，アダプターを形成する (図 1)．これにより，軽量なモデルでアダプターを表現できる．アダプター 1




(a) near-identity 初期化　(b) ランダム 初期化

図 2: アダプターの初期化方法

により，SE と ASR のミスマッチ問題を緩和するとともに，SE によって過度に失われる音声情報を未処理音声や他の周波数帯の音声情報で補うことが期待できる．アダプター 2 により，アダプター 1 で緩和しきれなかった雑音や歪み成分に由来する潜在特徴量のミスマッチが緩和されることが期待される．

### 2.3 Near-identity 初期化 (NI 初期化)

一般的に，ニューラルネットワークのパラメータはランダム初期化を行うことが多い (図 2(b)) が，パラメータ値が最適値と大きく異なるため，学習に時間がかかる傾向がある．この問題に対し，アダプターが恒等写像に近くなるような near-Identity 初期化 (NI 初期化) が有効であることが報告されている [19]．そこで，NI 初期化をアダプターに適用する．この場合，図 2(a) に示すように，線形層の対角成分を 1 に，他の重みとバイアスを 0 に近い値とすることで，アダプタの入力と出力がほぼ同じになるように初期化することができる．なお，アダプター 1 の 最終線形層の NI 初期化では，未処理音声特徴量と強調音声特徴量の平均が出力となるように，対角成分が 0.5 の対角行列を 2 つ横に並べた値で初期化する．

## 3 実験設定

本章では，提案モデルである PAM と NI 初期化の有効性を確認するために，LibriSpeech 音声コーパス [20] を用いたモデル学習を行い，認識性能に関する比較実験を行う．アダプター層の学習では，train-clean-100 subset（100 時間）を音声を用いる実験と，train-clean-100, train-clean-360, train-other-500（合計 960 時間）の音声を用いる実験を行う．学習を行う際には，自動車製造工場で録音した雑音を SNR (信号対雑音比) が 0 dB になるよう足し合わせた学習データを作成し，25 epoch 分の学習を行った．なお，使用した雑音は，定常的な低周波成分と不定期な高周波を含む機械音が含まれる．評価用音声には，LibriSpeech データセットの dev-clean と test-clean を用いた．雑音ロバスト性を評価するために，これらの評価音声に上記工場で別途収録した雑音を SNR が，15 dB，10 dB，5 dB, 0 dB になるように足し合わせたものを用意し，評価用音声とした．

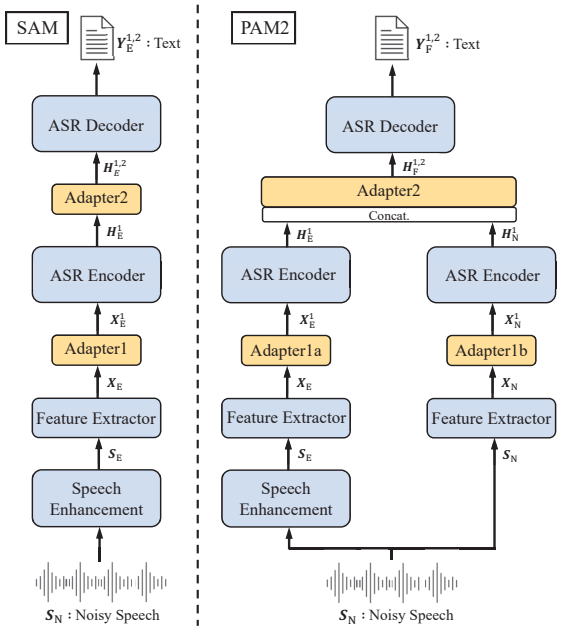



図 3: 比較モデル ; SAM,PAM2

### 3.1 比較モデル

PAM の有効性を確かめるために，他の 2 種のアダプターモデルである，Serial Adapter Model (SAM) と，Parallel Adapter Model 2 (PAM2) を比較に用いた．図 3 にこれらの比較モデルの構造を示す．SAM では，SE と ASR の直列モデルに対して，2 箇所にアダプターを挿入する．このモデルは他のモデルに比べて軽量であるが，未処理音声を推論に用いないモデルである．PAM2 は，提案手法である PAM と同様に，SE を適用しない音声を用いる．2 つの音響特徴量を，別々にアダプタ 1a とアダプタ 1b に送出し，共通の学習済みパラメータをもつエンコーダーによって潜在特徴量に変換し，これらの特徴量をアダプター 2 によって結合・混合する．PAM2 の Adapter2 の構造は，図 1 の Adapter1 と同様に，入力層は潜在特徴量の 2 倍の次元数を持ち，最後の線形層でもとの次元に戻される．

また，これらの比較モデルとは別に，PAM において片方のアダプターのみを用いる場合の性能も検証する．

### 3.2 モデル設定

入力音声は，16 bit, 16 kHz サンプリングの信号であり，FE は，これを窓長 512，シフト長 160 でフレーム化したのち，80 次元のメルフィルタバンクを適用した結果得られる音響特徴量を出力する．

表 1: 100 時間の教師用音声を用いてモデル事前学習・アダプター学習を行った場合の音声認識精度 (WER). 「H」は「hybrid CTC/Attention ASR [4]」, 「C」は「Conv-Tasnet [12]」, 「N」は「NI 初期化」, 「R」は「ランダム初期化」, 「Tunable Param.」は, アダプターのパラメータ数の合計を表す.

| No. | Model | | | | Tunable | clean | | 15dB | | 10dB | | 5dB | | 0dB | | Ave. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR | SE | Type | Init. | Param. | dev | test | dev | test | dev | test | dev | test | dev | test | |
| 1 | H | - | - | - | - | **8.3** | **8.6** | 15.8 | 15.0 | 29.4 | 26.6 | 57.4 | 52.6 | 85.5 | 81.1 | 38.0 |
| 2 | H | C | - | - | - | 9.0 | 9.2 | 14.4 | 13.6 | 21.8 | 19.5 | 38.2 | 33.9 | 68.9 | 63.6 | 29.2 |
| 3 | H | C | Takeda [14] | - | - | 9.3 | 9.5 | 14.3 | 13.8 | 22.1 | 20.1 | 39.1 | 34.6 | 68.0 | 62.4 | 29.3 |
| 4 | H | C | PAM (Proposed) | N | 536K | 12.3 | 12.0 | **13.1** | **12.9** | **15.9** | 15.7 | **23.8** | **21.7** | **42.8** | **38.6** | **20.9** |
| 5 | | | PAM | R | | 25.2 | 23.6 | 18.9 | 18.3 | 22.8 | 21.7 | 32.2 | 30.2 | 54.3 | 50.4 | 29.8 |
| 6 | H | C | SAM | N | 434K | 11.1 | 11.2 | 14.5 | 14.2 | 19.1 | 17.9 | 30.9 | 27.8 | 55.5 | 50.3 | 25.3 |
| 7 | | | SAM | R | | 53.7 | 53.0 | 54.2 | 53.8 | 58.2 | 57.4 | 67.6 | 66.0 | 82.9 | 80.4 | 62.7 |
| 8 | H | C | PAM2 | N | 1.52M | 11.6 | 11.5 | **13.1** | **12.9** | 16.0 | **15.6** | 24.2 | 24.2 | 44.6 | 44.6 | 21.8 |
| 9 | | | PAM2 | R | | 92.9 | 93.0 | 87.9 | 88.5 | 88.3 | 88.3 | 90.5 | 89.9 | 95.6 | 94.9 | 91.0 |

表 2: 960 時間の教師用音声を用いてモデル事前学習・アダプター学習を行った場合の音声認識精度 (WER). 「H」は「hybrid CTC/Attention ASR [4]」, 「C」は「Conv-Tasnet [12]」, 「N」は「NI 初期化」, 「R」は「ランダム初期化」, 「Tunable Param.」は, アダプターのパラメータ数の合計を表す.

| No. | Model | | | Tunable | clean | | 15dB | | 10dB | | 5dB | | 0dB | | Ave. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR | SE | Type | Param. | dev | test | dev | test | dev | test | dev | test | dev | test | |
| 10 | H | - | - | - | **2.6** | **2.6** | **3.8** | **3.5** | 7.0 | 6.2 | 24.8 | 20.2 | 69.1 | 62.3 | 20.2 |
| 11 | H | C | - | - | 3.8 | 4.0 | 5.2 | 4.8 | 7.6 | 6.9 | 15.1 | 12.9 | 36.9 | 31.9 | 12.9 |
| 12 | H | C | Takeda [14] | - | 2.9 | 2.9 | 4.3 | 4.0 | 6.6 | 6.0 | 15.7 | 13.1 | 43.9 | 37.4 | 13.7 |
| 13 | H | C | PAM (Proposed) | 1.72M | 4.2 | 4.3 | 4.6 | 4.6 | **6.0** | **5.7** | **9.9** | **8.9** | **24.2** | **20.3** | **9.3** |
| 14 | H | C | SAM | 1.61M | 3.6 | 3.6 | 5.1 | 4.6 | 7.3 | 6.6 | 14.9 | 12.6 | 36.3 | 31.7 | 12.6 |
| 15 | H | C | PAM2 | 5.85M | 4.5 | 4.6 | 5.0 | 5.2 | 6.5 | 6.3 | 10.5 | 9.6 | 26.3 | 22.1 | 10.1 |

ASR には, ESPnet [21], の hybrid CTC/attention モデル [4] を用いる. このモデルは, 12 個の Conformer ブロックで構成されるエンコーダーと, 6 層の Transformer ブロックと線形層で構成されるデコーダーをもつ. エンコーダーが出力する潜在特徴量は, 100 時間の音声で学習する場合は 256 次元に, 960 時間の音声で学習を行う際には 512 次元にした. 損失関数には, CTC 損失と Attention デコーダーの損失の重み付き和を用い, 重みはそれぞれ 0.3, 0.7 である.

SE には, 深層学習ベースの手法である Conv-TasNet [12] を用いた. このモデルは, エンコーダー, セパレーター, デコーダーから構成される. エンコーダーは 1 層の畳み込み層, デコーダーは 1 層の転置畳み込み層からなる. セパレーターはマスク推定用のネットワークで, それぞれ 8 個の畳み込みブロックを含む, 4 個の Temporal Convolutional Network [22] からなる. SE は SI-SNR [23] を最大化するように, CHiME-4 データセット [24] で事前学習されたモデルを用いた.

PAM と PAM2 の学習では SpecAug を用いず, SAM の学習では SpecAug を用いた. NI 初期化を用いる場合, 対角成分は 1, それ以外の重みとバイアスは平均 0, 標準偏差 0.01 の正規分布で初期化した. またランダム初期化を用いる場合, 線形層の入力次元数を $m$ とすると, 一様分布 $[-\sqrt{1/m}, \sqrt{1/m}]$ で初期化を行った.

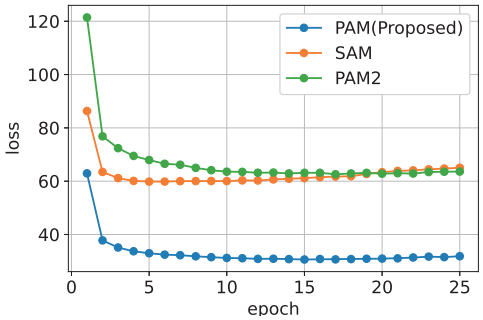


図 4: PAM,SAM,PAM2 の学習損失推移

## 4 実験結果

本章では, 実験結果について, 初期化手法, モデルの妥当性, アダプターの寄与の 3 点から考察する.

### 4.1 初期化手法について

表 1 に, 100 時間の教師用音声を用いてアダプター学習を行った結果を示す. これより, NI 初期化を用いると, 音声強調のみを用いた場合 (No. 2, 3) に比べて, すべてのアダプターモデル (No. 4, 6, 8) で雑音環境での認識性能が向上した. NI 初期化を用いることで, もとのモデルの性質を大きく崩さない初期値から学習が

表 3: 100 時間の教師用音声を用いてモデル事前学習・アダプター学習を行った場合の音声認識精度 (WER).「H」は「hybrid CTC/Attention ASR [4]」,「C」は「Conv-Tasnet [12]」,「Adapter」ではそれぞれのアダプターの使用有無を示している．No. 16, 19 はそれぞれ表 1 の No. 2, 4 と同一の実験結果である．

| No. | Model | | | Adapter | | clean | | 15dB | | 10dB | | 5dB | | 0dB | | Ave. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR | SE | Type | 1 | 2 | dev | test | dev | test | dev | test | dev | test | dev | test | |
| 16 | H | C | - | - | - | **9.0** | **9.2** | 14.4 | 13.6 | 21.8 | 19.5 | 38.2 | 33.9 | 68.9 | 63.6 | 29.2 |
| 17 | H | C | PAM (Proposed) | ✓ | - | 13.0 | 12.9 | 14.2 | 13.8 | 17.1 | 16.9 | 26.3 | 23.7 | 47.0 | 42.4 | 22.7 |
| 18 | | | | - | ✓ | 10.5 | 10.8 | 14.0 | 13.8 | 19.3 | 18.2 | 32.1 | 28.9 | 58.8 | 53.1 | 26.0 |
| 19 | | | | ✓ | ✓ | 12.3 | 12.0 | **13.1** | **12.9** | **15.9** | **15.7** | **23.8** | **21.7** | **42.8** | **38.6** | **20.9** |

開始でき，少ない学習量でも，パラメータが良好に学習できていることが確認できた．その一方で，ランダム初期化を用いた場合，性能が向上するケース (No. 5) と，劣化するケース (No. 7, 9) が見られた．ランダム初期化を用いて性能が向上したケースでも，NI 初期化を用いると，さらに性能が向上した．このため，ランダム初期化は，学習不足で十分な最適化ができていないことが推測される．以上の結果より，アダプター学習における NI 初期化の有効性が示された．以降では，NI 初期化を用いる場合のみ考える．

### 4.2 モデルごとの性能について

表 1 より，提案手法である PAM が，ほとんどの雑音環境での認識性能，および平均性能で最も良好であるという結果となった．SNR が 0 dB のとき，SE のみを適用した場合に比べて，dev-clean は 26.1 ポイント，test-clean は 25.0 ポイントの性能向上が確認できた．さらに，平均性能では，8.3 ポイント性能が向上した．SAM では，どのケースも PAM に比べて明らかに性能が低いことがわかる．PAM2 については，15 dB や 10 dB の雑音環境において PAM と同等の性能が確認できた．一方で，高雑音環境では PAM の方が性能が高い．更新パラメータ数では，PAM のほうが少ないことを考慮すれば，PAM がより優秀なモデルであると言える．また，図 4 に学習損失の推移を示す．これより，ほとんどのモデルで，10 epoch 程度で学習損失が収束しており，その中でも PAM が最も学習損失が小さい結果となっている．学習損失の点からも，PAM 有効性が確認できた．

表 2 は，960 時間の教師用音声を用いてアダプター学習を行った結果を示している．こちらの実験でも，提案手法である PAM が，雑音の多い環境での認識性能，および平均性能で最も良好であるという結果となった．SNR が，0 dB のとき，SE のみを適用した場合に比べて，dev-clean は 12.7 ポイント，test-clean は 11.6 ポイントの性能向上が確認できた．さらに，平均性能では，3.6 ポイント性能が向上した．

### 4.3 アダプターの性能への寄与について

表 3 より，SE のみ用いた場合に比べ，いずれかのアダプターを用いた場合にも性能の向上が確認された．Adapter1 と Adapter2 のそれぞれ一方のみを用いた場合 (No. 17, 18) では，Adapter1 のみを用いたほう (No. 17) が大きな性能向上がみられた．Adapter1 のみ用いた場合，SNR が 0 dB の環境において，dev-clean では 21.9 ポイント，test-clean では 21.2 ポイントの向上であったのに対し，Adapter2 のみ用いた場合，dev-clean では 10.1 ポイント，test-clean では 10.5 ポイントの向上であった．また平均性能も，Adapter2 よりも Adapter1 のほうが 3.3 ポイント上回っている．Adapter1 では音響特徴量を混合しており，特徴量の変換がより有効に機能したと考えられる．

その一方で，アダプターを 2 つ用いる PAM(No. 19) では，さらに精度が向上している．以上より，2 つのアダプターが性能に寄与し，音声認識性能を改善していることが確認された．

## 5 むすび

本研究では，音声認識の雑音耐性を向上するために，音声強調を音声認識のフロントエンドとして用いる場合に問題となる音声認識と音声強調のミスマッチ問題を緩和するために，Parallel Adapter Model とその初期化に Near-Identity 初期化法を利用することを提案した．Parallel Adapter Model は，強調音声に加えて，未処理音声を用いてミスマッチ問題の緩和をはかる一つ目のアダプターと，一つ目のアダプターで緩和しきれなかった潜在特徴量に残存するミスマッチを緩和する二つ目のアダプターの 2 つのアダプターを用いる手法である．また，Near-Identity 初期化法は，恒等写像に近い値で初期化することで，効率的な学習を可能にする手法である．実験の結果，train-clean-100 を用い提案手法で学習を行ったモデルは，アダプターを用いない場合と比較して，性能が最大で 25.2 ポイント，平均で 8.4 ポイント性能が向上した．また，PAM においては，2 つのアダプターを用いることで性能がより

向上することが確認された．今後の課題として，アダプター構造のさらなる検討や，他の ASR/SE モデルへの提案手法の適用があげられる．

## 謝辞

本研究は JSPS 科研費 JP19KK0260, JP20H00475 および JP23K11160 の助成を受けた．

## 参考文献

[1] B. H. Juang and L. R. Rabiner. Hidden Markov models for speech recognition. *Technometrics*, Vol. 33, pp. 251–272, 1991.

[2] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82–97, 2012.

[3] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, 2013.

[4] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 11, No. 8, pp. 1240–1253, 2017.

[5] Yui Sudo, Muhammad Shakeel, Brian Yan, Jiatong Shi, and Shinji Watanabe. 4D ASR: Joint modeling of CTC, attention, transducer, and mask-predict decoders. In *in Proc. Interspeech*.

[6] Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, et al. Reproducing whisper-style training using an open-source toolkit and publicly available data. *arXiv preprint arXiv:2309.13876*, 2023.

[7] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888, 2018.

[8] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani. Robust mvdr beamforming using time-frequency masks for online/offline asr in noise. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5210–5214, 2016.

[9] Jahn Heymann, Lukas Drude, Christoph Boeddeker, Patrick Hanebrink, and Reinhold Haeb-Umbach. Beamnet: End-to-end training of a beamformer-supported multi-channel asr system. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329, 2017.

[10] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. Complex spectral mapping for single- and multi-channel speech enhancement and robust asr. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 1778–1787, 2020.

[11] Panagiotis Tzirakis, Anurag Kumar, and Jacob Donley. Multi-channel speech enhancement using graph neural networks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3415–3419, 2021.

[12] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time – frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 8, pp. 1256–1266, 2019.

[13] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models, 2022.

[14] Ryu Takeda, Yui Sudo, Kazuhiro Nakadai, and Kazunori Komatani. Empirical Sampling from Latent Utterance-wise Evidence Model for Missing Data ASR based on Neural Encoder-Decoder Model. In *Proc. Interspeech*, pp. 3789–3793, 2022.

[15] Ryu Takeda, Yui Sudo, and Kazunori Komatani. Flexible Evidence Model to Reduce Uncertainty Mismatch Between Speech Enhancement and ASR Based on Encoder-Decoder Architecture. In *Proc. APSIPA*, 2023.

[16] Jisi Zhang, Catalin Zorila, Rama Doddipatla, and Jon Barker. On monoaural speech enhancement for automatic recognition of real noisy speech using mixture invariant training. In *Interspeech 2022*. ISCA, sep 2022.

[17] Xuankai Chang, Takashi Maekaku, Yuya Fujita, and Shinji Watanabe. End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation, 2022.

[18] Yuchen Hu, Nana Hou, Chen Chen, and Eng Siong Chng. Dual-Path Style Learning for End-to-End Noise-Robust Speech Recognition. In *Proc. INTERSPEECH 2023*, pp. 2918–2922, 2023.

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019.

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[21] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. Espnet: End-to-end speech processing toolkit, 2018.

[22] Colin Lea, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pp. 47–54, Cham, 2016. Springer International Publishing.

[23] Yi Luo and Nima Mesgarani. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, 2018.

[24] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, Vol. 46, pp. 535–557, 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-063-03

# End-to-end integration of online and offline encoders using auxiliary losses for automatic speech recognition

Muhammad Shakeel[1*] Yui Sudo[1] Yifan Peng[2] Shinji Watanabe[2]

[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] カーネギーメロン大学
[2] Carnegie Mellon University

**Abstract:** End-to-end (E2E) automatic speech recognition (ASR) models have two desirable properties: online and offline modes. The online ASR mode, which operates under strict latency constraints, processes speech frames in real-time to provide transcription. Conversely, the offline ASR mode waits for the complete utterance of speech frames before generating a transcription. Recently, the integration of online and offline ASR for recurrent neural network transducers (RNN-T) can be achieved through the joint training of online and offline encoders with a shared decoder. However, this integration comes at the cost of performance degradation in the offline ASR mode, as the shared decoder must handle features of varying contexts. Namely, with E2E integration framework of online and offline encoders, we explore two approaches to enhance the performance of both the ASR modes. First, we introduce separate RNN-T decoders for each ASR mode while maintaining shared encoders, thereby effectively managing features of different contexts. Second, we explore multiple auxiliary loss criteria to introduce additional regularization, thereby enhancing the overall stability and performance of the framework. Overall, evaluation results show 1.8%-2.5% relative character error rate reductions (CERR) on corpora of spontaneous Japanese (CSJ) for online ASR, and 4.4%-6.3% relative CERRs for offline ASR within a single model compared to separate online and offline models.

## 1 Introduction

End-to-end automatic speech recognition (E2E-ASR) systems [1] strive to achieve low latency and high performance for a variety of tasks, including online [2] and offline [3] ASR. However, the creation of distinct architectures for each task is neither scalable nor flexible. Therefore, a single E2E-ASR framework, capable of handling multiple tasks with high accuracy and adaptability, is preferable. One potential solution is the joint training of online and offline E2E-ASR tasks using shared weights. However, most of the existing methods suffer from negative transfer, a phenomenon where the performance of one task interferes with another and degrades the performance of either of them. For example, in the case of shared weights between online and offline E2E-ASR, the limited contextual features of online encoder could lead to a conflict with full-context offline encoder. This is because the model might not be able to effectively differentiate between the features relevant to each encoder, leading to the inclusion of irrelevant negative samples. To address this issue, we propose an E2E-ASR framework that integrates online and offline ASR and combines the unique capabilities of both the ASR models.

In this study, our primary objective is to optimize both online and offline ASR modes. This framework employs multiple encoders, with one designated for online ASR and another for offline ASR, and separate decoders for each. The framework extracts the hidden states of the online encoder and uses them as input for the offline encoder. This method allows us to integrate the functionalities of both encoders while maintaining separate decoders for online and offline modes. As a general approach rather than relying only on the cascaded integration [4] for performance optimization, our method introduces sepa-

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

rate online and offline recurrent neural network transducer (RNN-T) [5] decoders to leverage varying contextual information from both online and offline encoders. Additionally, we employ connectionist temporal classification (CTC) [6], attention mechanism [7], and masked language model (MLM) [8] based auxiliary losses to bring more regularization and refinement to the E2E-ASR framework. The CTC loss helps in aligning the input features with the output labels, the attention mechanism provides a dynamic alignment between the input and output sequences, and the MLM loss aids in predicting the masked input features.

Through extensive experimentation, we have been able to demonstrate the effectiveness of each auxiliary loss in improving the performance of the offline ASR model. Our results indicate a significant improvement in the accuracy and efficiency of both online and offline ASR models, validating the effectiveness of our integrated framework and auxiliary loss techniques.

## 2 Related work

One of the current challenges in E2E-ASR is to develop a unified model [9,10] that can handle both online and offline scenarios. Online ASR is suitable for applications that require low latency and real-time feedback, such as voice assistants and online meetings. However, online ASR can only use limited context information from the past and present frames, which may limit its accuracy and robustness. On the other hand, offline ASR is suitable for applications that do not have strict latency constraints, such as offline transcription and speech analysis. Offline ASR can exploit full context information from the whole utterance, which may improve its performance and generalization. Therefore, different context information may require different acoustic features and network architectures, making it difficult to jointly optimize a single model for both scenarios.

In recent studies [11], authors have explored the unification of online and offline encoders by using the same decoder for different input features, or by using the output of the online encoder as the input of the offline encoder [12]. However, integrating online and offline encoders often face challenges in offline scenarios, particularly when the online mode is prioritized during optimization. A common approach involves using a single shared decoder based on RNN-T to handle features of varying contexts. However, the shared

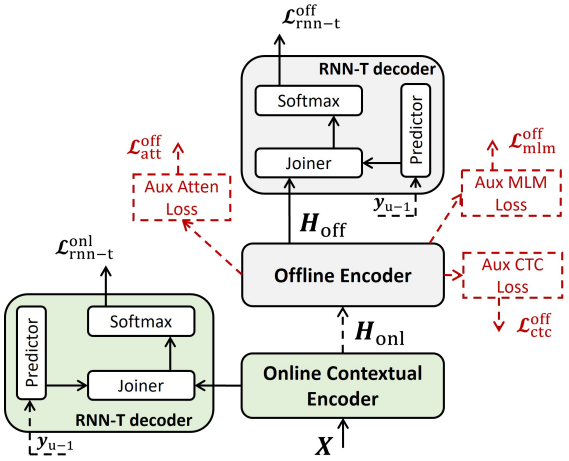



図 1: An end-to-end architecture of online and offline encoders integration with auxiliary losses: In black, the original losses. In red the new auxiliary losses.

decoder may not effectively differentiate between the limited and full context features, leading to the inclusion of more negative samples; thus, degrading the ASR performance. While multitask learning strives to enhance the performance of multiple tasks concurrently, the selective use of auxiliary losses [13] can support the primary task and improve the generalization ability of the framework. Therefore, we propose an end-to-end integration of online and offline ASR using separate RNN-T decoders and auxiliary losses, such as CTC, attention, and MLM loss, each of which brings its own strength in a single model. We expect that our proposed model can achieve better performance than the existing methods.

## 3 Framework

This section introduces the proposed end-to-end integration of online and offline ASR framework, followed by detailed descriptions of each of our design modules.

### 3.1 Online and offline encoder

We propose a joint architecture that combines an online encoder and an offline encoder, each with its own RNN-T decoder, to handle both online and offline application scenarios. The online encoder is based on block processing [2], which preserves the previous context information using context embeddings. The offline encoder is based on conformer [3], which captures the full context information from the whole utterance. The RNN-T decoders are based on recurrent

neural network transducer, which models the sequential nature of speech and text. Figure 1 shows the overview of our proposed architecture. The online encoder consists of $M$ encoder layers, each of which has a context inheritance mechanism. The context inheritance mechanism computes a context embedding for each block at each sublayer, and passes it to the next sublayer. The context embedding encodes the past and present context information of the block. The block size and hop length are denoted by $L_{\text{block}}$ and $L_{\text{hop}}$, respectively. The b-th block of the input audio feature sequence $\boldsymbol{X_b}$ is defined as:

$$\boldsymbol{X}^b = (\boldsymbol{X}_t | t = (b-1)L_{\text{hop}} + 1, ..., (b-1)L_{\text{hop}} + L_{\text{block}} + 1) \quad (1)$$

The hidden state for each block, labeled as the $b$-th block, is encoded whereas each block contains a series of hidden states of $L_{\text{block}}$- *length*, i.e., $\boldsymbol{H}^b = (\boldsymbol{h}^b_1, \ldots, \boldsymbol{h}^b_{L_{block}})$. The encoding process is carried out sequentially, resulting in a series of hidden states with a length of $T$. These features are then input into the offline encoder, where they are transformed into a subsampled sequence of hidden states, also of length $T$, as given in the Eq.(3).

$$\boldsymbol{H}_{\text{onl}} = \text{OnlineEncoder}(\boldsymbol{X}). \quad (2)$$

$$\boldsymbol{H}_{\text{off}} = \text{OfflineEncoder}(\boldsymbol{H}_{\text{onl}}). \quad (3)$$

In this study, we have the online contextual conformer encoder functioning as an independent online encoder-decoder module. This is linked to an offline encoder-decoder module via an output derived from the online encoder.

### 3.2 Online encoder-decoder loss

The acoustic features, denoted as $\boldsymbol{X} = (x_1, \ldots, x_T)$ are initially processed by the online module. This module employs a contextual block conformer as an encoder and the RNN-T as an online decoder, as described in [2] . The online RNN-T decoder computes the marginal likelihood of the output $y$ over all possible alignments, as shown in Eq.(4):

$$P_{\text{onl}}(\boldsymbol{y} \mid \boldsymbol{X}) = \sum_{\boldsymbol{o} \in \mathcal{O}(\boldsymbol{y})} P_{\text{onl}}(\boldsymbol{o}|\boldsymbol{X}) = \sum_{\boldsymbol{o} \in \mathcal{O}(\boldsymbol{y})} \left[ \prod_{i=1}^{T+S} P\left(\boldsymbol{o}_i \mid \boldsymbol{h}_{t_i}, g_{s_i}\right) \right], \quad (4)$$

The RNN-T model optimizes its parameters by minimizing the negative log-likelihood, as defined in Eq. (5):

$$\mathcal{L}^{\text{rnn-t}}_{\text{onl}} = - \sum_{(\boldsymbol{X} \rightarrow \boldsymbol{H}_{\text{onl}}, \boldsymbol{y})} \log P_{\text{onl}}(\boldsymbol{y} \mid \boldsymbol{H}_{\text{onl}}). \quad (5)$$

This loss ensures that the model is continually optimized to learn the online contextual features during training.

### 3.3 Offline encoder-decoder loss

The offline encoder-decoder modules receives the processed hidden sequences from the online encoder and employs full-context conformer as an encoder and separate RNN-T as an offline decoder. Here, the offline RNN-T decoder computes the marginal likelihood of the output $y$ over all possible alignments, as shown in Eq.(6):

$$P_{\text{off}}(\boldsymbol{y} \mid \boldsymbol{H}_{\text{onl}}) = \sum_{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{y})} P_{\text{off}}(\boldsymbol{a}|\boldsymbol{H}_{\text{onl}}) = \sum_{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{y})} \left[ \prod_{i=1}^{T+S} P\left(\boldsymbol{a}_i \mid \boldsymbol{h}_{t_i}, g_{s_i}\right) \right], \quad (6)$$

where the offline RNN-T loss refines the model parameters by minimizing the negative log-likelihood, as shown below:

$$\mathcal{L}^{\text{rnn-t}}_{\text{off}} = - \sum_{(\boldsymbol{H}_{\text{onl}} \rightarrow \boldsymbol{H}_{\text{off}}, \boldsymbol{y})} \log P_{\text{off}}(\boldsymbol{y} \mid \boldsymbol{H}_{\text{off}}), \quad (7)$$

### 3.4 Auxiliary losses

In this study, we enhance the performance of the offline ASR by optimizing the offline encoder-decoder module. This optimization is achieved through a multi-task approach, where a full-context Conformer encoder is shared with the CTC, attention, and MLM-based auxiliary losses. The offline encoder, which relies on the limited context hidden state features processed by the online encoder, often experiences performance degradation due to the restricted context information. Our approach mitigates this issue and enhances the robustness of the framework by incorporating additional auxiliary losses into the offline RNN-T decoder. The inclusion of these losses serves a dual purpose. Firstly, they contribute to the regularization of the framework, ensuring stability during the learning process. Secondly, they aid in the optimization of the model by providing additional signals for error correction during training.

#### 3.4.1 CTC loss

The likelihood of the CTC is given in Eq.(8):

$$P_{\text{ctc}}(\boldsymbol{y} \mid \boldsymbol{H}_{\text{off}}) = \sum_{\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y})} P(\boldsymbol{z} \mid \boldsymbol{H}_{\text{off}}) = \sum_{\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y})} \left[ \prod_{t=1}^{T} P\left(\boldsymbol{z}_t \mid \boldsymbol{h}_t\right) \right], \quad (8)$$

where it serves as an auxiliary loss to refine the model parameters by minimizing the negative log-likelihood, as shown in the Eq.(9) below:

$$L_{\text{ctc}} = - \sum_{(\boldsymbol{H}_{\text{onl}} \rightarrow \boldsymbol{H}_{\text{off}}, \boldsymbol{y})} \log P_{\text{ctc}}(\boldsymbol{y} \mid \boldsymbol{H}_{\text{off}}), \quad (9)$$

#### 3.4.2 Auxiliary attention loss

The likelihood of an attention mechanism is described as in Eq.(10):

$$P_{\text{att}}(\boldsymbol{y} \mid \boldsymbol{H}_{\text{onl}}) = \prod_{s=1}^{S} P\left(y_s \mid \boldsymbol{y_{s-1}}, \boldsymbol{H}_{\text{off}}\right), \quad (10)$$

The auxiliary attention loss provides an alignment between the input and output sequences and brings model regularization by optimizing the corresponding negative log-likelihood, as shown below:

$$L_{\text{att}} = - \sum_{(\boldsymbol{H}_{\text{onl}} \rightarrow \boldsymbol{H}_{\text{off}}, \boldsymbol{y})} \log P_{\text{att}}(\boldsymbol{y} \mid \boldsymbol{H}_{\text{off}}). \quad (11)$$

#### 3.4.3 Auxiliary MLM loss

The MLM auxiliary loss in our framework estimates the token sequence using the full sequence given by $\boldsymbol{H}_{\text{off}}$ as shown in Eq.(3), similar to the attention mechanism. However, during the training phase, MLM distinguishes itself from attention by masking randomly selected tokens, denoted as $y_{mask}$, with a special token.

Subsequently, $y_{mask}$ is predicted based on the remaining unmasked tokens, $y_{obs}$, as $P_{\text{mlm}}(y_{\text{mask}}|y_{\text{obs}}, \boldsymbol{H}_{\text{off}})$. Here, the MLM refines the model parameters by minimizing the negative log-likelihood as outlined in the equation below:

$$\mathcal{L}_{\text{off}}^{\text{mlm}} = - \sum_{(\boldsymbol{H}_{\text{onl}} \rightarrow \boldsymbol{H}_{\text{off}}, \boldsymbol{y})} \log P_{\text{mlm}}(y_{\text{mask}}|y_{\text{obs}}, \boldsymbol{H}_{\text{off}}). \quad (12)$$

This approach allows the MLM to contribute to the refinement of the model parameters, enhancing the overall performance of the system.

Finally, the offline loss ($\mathcal{L}_{\text{off}}$) is computed using the weighted sum of individual loss objectives as defined in Eqs.(7), (9), (11) and (12):

$$\mathcal{L}_{\text{off}} = \lambda_{\text{ctc}}\mathcal{L}_{\text{off}}^{\text{ctc}} + \lambda_{\text{rnnt}}\mathcal{L}_{\text{off}}^{\text{rnnt}} + \lambda_{\text{att}}\mathcal{L}_{\text{off}}^{\text{att}} + \lambda_{\text{mlm}}\mathcal{L}_{\text{off}}^{\text{mlm}}, \quad (13)$$

where $\lambda_{\text{ctc}}, \lambda_{\text{rnnt}}, \lambda_{\text{att}}$ and $\lambda_{\text{mlm}}$ are tunable hyperparameters and are determined experimentally. However, for this work we used the hyperparameters as reported in [14] and obtained optimal results.

Finally, we define the total multi-task learning objective for end-to-end integration of online and offline encoders as:

$$\mathcal{L}_{\text{mtl}} = \lambda_{\text{onl}}\mathcal{L}_{\text{rnn-t}}^{\text{onl}} + \lambda_{\text{off}}\mathcal{L}_{\text{off}} \quad (14)$$

where $\lambda_{\text{onl}}$ and $\lambda_{\text{off}}$ are the weighting terms and $\mathcal{L}_{\text{onl}}$ is the online loss obtained from Eq.(5) and $\mathcal{L}_{\text{off}}$ is the offline loss obtained from Eq.(13).

## 4 Experiments

### 4.1 Dataset

In this study, we primarily focus on a specific subset, referred to as subset A, of the Corpus of Spontaneous Japanese (CSJ) [15]. This subset consists of academic lecture-based ASR tasks and comprises 236 hours of speech data.

For evaluation purposes, we have divided the dataset into three distinct tasks: eval 1, eval 2, and eval 3. These tasks contain 1.9 hours, 2.0 hours, and 1.3 hours of speech data respectively, providing a comprehensive and diverse set of data for our analysis.

### 4.2 Experimental setup

In our study, we utilized the ESPnet2 toolkit [16] as a foundation to build our baseline models and the proposed E2E framework. This framework is based on the concept of multi-task learning for ASR task. We integrated additional modules for encoder, decoder, and auxiliary losses into this framework, capitalizing on existing specialized architectures. Our online encoder is composed of twelve layers of 256-dimensional contextual block conformer, each layer having 1024 feed-forward dimensions and 4 attention heads. We apply a dropout rate of 0.1 to each layer. For block-processing [2], we configure the block size to 40 and maintain a look-ahead and hop size of 16 to optimize online streaming performance. Our offline encoder comprises of twelve layers of 256-dimensional

表 1: On Corpus of Spontaneous Japanese (CSJ) : Absolute (abs.) character error rate (CER) and relative (rel.) CERR numbers on CSJ data for i) single baseline models (B1, B3, B5, B6, & B7), ii) a baseline cascaded encoder with shared decoder models (B2 & B4) [12], iii) proposed end-to-end integration of online and offline encoders with separate RNN-T decoders (P1 & P2) and auxiliary losses (P3, P4 & P5). All the results are decoded with a beam size of 10.

| Mode | ID | Method | CSJ (SUBSET A) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | eval1 | | eval2 | | eval3 | |
| | | | abs.↓ | rel.(%)↑ | abs.↓ | rel.(%) ↑ | abs.↓ | rel.(%)↑ |
| **Online** | B1 | Context-Transducer (baseline) | 6.84 | | 4.95 | | 11.72 | |
| | B2 | Cascaded [12] (baseline) | 6.81 | (0.44) | 5.07 | (-2.42) | 11.89 | (-1.45) |
| | P1 | Online-Transducer **(ours)** | **6.67** | (2.49) | **4.86** | (1.81) | 11.79 | (-0.60) |
| **Offline** | B3 | Conformer-Transducer (baseline) | 5.78 | | 4.15 | | 9.94 | |
| | B4 | Cascaded [12] (baseline) | 5.60 | (3.11) | 4.04 | (2.65) | 9.68 | (2.61) |
| | P2 | Offline-Transducer **(ours)** | **5.48** | (5.19) | **3.89** | (6.27) | **9.50** | (4.42) |
| **Offline** | B5 | Conformer-CTC (baseline) | 5.50 | | 3.88 | | 9.76 | |
| | P3 | Conformer+$\mathcal{L}_{\text{off}}^{\text{ctc}}$ **(ours)** | 5.51 | (-0.18) | **3.72** | (4.12) | **9.62** | (1.43) |
| **Offline** | B6 | Conformer-Transformer (baseline) | 5.16 | | 3.87 | | 9.90 | |
| | P4 | Conformer+$\mathcal{L}_{\text{off}}^{\text{att}}$ **(ours)** | 5.29 | (-2.46) | **3.71** | (4.13) | **9.55** | (3.54) |
| **Offline** | B7 | Conformer-MLM (baseline) | 5.53 | | 4.00 | | 9.51 | |
| | P5 | Conformer+$\mathcal{L}_{\text{off}}^{\text{mlm}}$ **(ours)** | 5.58 | (-0.90) | **3.81** | (4.75) | 9.88 | (-3.74) |

full-context conformer [3], each layer having 1024 feed-forward dimensions and 4 attention heads. The output from the online contextual block conformer is channeled into the offline conformer block, resulting in a cascaded architecture. For the online encoder, we employ a separate RNN-T decoder with a 256-dimensional embedding prediction network and a 320-dimensional joint network. Conversely, for the offline encoder, we utilized a distinct RNN-T decoder with a 256-dimensional embedding prediction network and a 320-dimensional joint network. To augment the regularization of the offline encoder-decoder, we introduced auxiliary losses based on CTC, attention, and MLM mechanisms. We train this end-to-end architecture for 50 epochs with a learning rate of 0.0015 and warmup steps of 1500. We employ a training weight of 1 to the online encoder to maximize the performance capacity of the online ASR mode. However, for the offline mode and auxiliary losses, we adopt the training weights for $\lambda_{\text{rnnt}}, \lambda_{\text{ctc}}, \lambda_{\text{att}}, \lambda_{\text{mlm}}$ as proposed in [14], i.e., 0.10, 0.15, 0.30, and 0.45 respectively.

## 4.3 Main results

In this work, we evaluate the performance of our proposed E2E framework with several baseline models. These include the standalone online contextual block conformer transducer (Context-Transducer), the offline full-context conformer transducer (Conformer-Transducer), and a cascaded architecture with a shared RNN-T decoder, as proposed in a previous study [12]. To ensure a fair comparison, we maintained the same number of encoder layers (twelve) for both online and offline modes in all models, including the Context-T and Conformer-T.

For our first primary analysis, we perform an ablation study conducted on the Corpus of Spontaneous Japanese (CSJ) dataset, and is presented in Table 1. In this table, the standalone online and offline transducer baseline models are represented by Context-Transducer (`B1`) and Conformer-Transducer (`B3`), respectively. We also developed a baseline cascaded architecture with shared decoders (`B2` & `B4`) as proposed in the referenced study [12]. These models are compared against our proposed framework (`P1` & `P2`). The

character error rates (CER) listed in Table 1 are obtained using a beam width of 10 for both online and offline RNN-T modules. Our findings indicate that the proposed framework improved the performance of the online ASR path compared to the standalone online model. We observed a relative CER (CERR) improvement ranging from 1.81% to 2.49% across multiple evaluation sets. Moreover, our proposed framework also demonstrates substantial performance improvement for the offline ASR path, with a CERR between 4.4% and 6.3%. These results highlight the effectiveness of our proposed E2E framework in improving the performance of both online and offline ASR modules.

Next, we compare our auxiliary tasks $\mathcal{L}_{\text{off}}^{\text{ctc}}$ (P3), $\mathcal{L}_{\text{off}}^{\text{att}}$ (P4) and $\mathcal{L}_{\text{off}}^{\text{mlm}}$ (P5) against the separately trained Conformer-CTC (B5), Conformer-Transformer (B6), and Conformer-MLM (B7) models. Table 1 summarizes our experiments to understand how each task improved or degrades the performance on CSJ evaluation sets compared to the standalone models. Overall, most auxiliary tasks show improved performance on eval2 and eval3 test sets with an exception for test set eval1 which shows performance degradation. This degradation can be attributed to the fact that training weights in this study are optimized to improve the overall performance of the online and offline transducer modes. The performance for each auxiliary task can potentially be improved by assigning more optimal weights to the individual task. It will allow the model to focus more on optimizing the performance of each auxiliary task, rather than prioritizing the overall performance of the transducer modes.

# 5 Conclusion

In this study, we propose a novel approach to integrate online and offline ASR modules in an end-to-end manner, utilizing auxiliary losses. This framework is designed to optimize the combination of online and offline RNN-T decoders, leveraging the power of multi-task learning. The primary objective is to enhance the learning of contextual representations, thereby offering increased flexibility in the E2E-ASR framework. Our approach demonstrates a significant improvement in CERR for the CSJ corpus, compared to traditional cascaded architectures [12]. This improvement is particularly noticeable with the introduction of auxiliary losses, which provide additional regularization and refinement to the framework.

# 参考文献


[1] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," 2023.

[2] E. Tsunoo, Y. Kashiwagi, T. Kumakura, and S. Watanabe, "Transformer ASR with contextual block processing," in *Proc. ASRU*, 2019.

[3] A. Gulati, C. Chiu, J. Qin, J. Yu, N. Parmar, R. Pang, S. Wang, W. Han, Y. Wu, Y. Zhang, and Z. Zhang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020.

[4] B. L. et al., "A better and faster end-to-end model for streaming asr," in *Proc. ICASSP*, 2021.

[5] A. Graves, "Sequence transduction with recurrent neural networks," in *Proc. ICML*, 2012.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[7] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016.

[8] Y. Higuchi, S. Watanabe, N. Chen, T. Ogawa, and T. Kobayashi, "Mask ctc: Non-autoregressive end-to-end asr with ctc and mask predict," in *Proc. Interspeech*, 2020.

[9] F. Weninger, M. Gaudesi, M. A. Haidar, N. Ferri, J. Andr'es-Ferrer, and P. Zhan, "Conformer with dual-mode chunked attention for joint online and offline asr," in *Proc. Interspeech*, 2022.

[10] Y. Sudo, M. Shakeel, Y. Peng, and S. Watanabe, "Time-synchronous one-pass beam search for parallel online and offline transducers with dynamic block training," in *Proc. Interspeech*, 2023.

[11] J. Yu, W. Han, A. Gulati, C.-C. Chiu, B. Li, T. N. Sainath, Y. Wu, and R. Pang, "Dual-mode {asr}: Unify and improve streaming asr with full-context modeling," in *Proc. ICLR*, 2021.

[12] A. Narayanan, T. N. Sainath, R. Pang, J. Yu, C.-C. Chiu, R. Prabhavalkar, E. Variani, and T. Strohman, "Cascaded encoders for unifying streaming and non-streaming asr," in *Proc. ICASSP*, 2021.

[13] C. Liu, F. Zhang, D. Le, S. Kim, Y. Saraf, and G. Zweig, "Improving rnn transducer based asr with auxiliary tasks," in *Proc. SLT*, 2021.

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

[15] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese," in *Proc. LREC*, 2000.

[16] S. W. et al., "Espnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 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-063-04

# 曖昧な言語指示に対応可能な大規模言語モデルを用いた動作計画システムの開発

# Development of a Task Planning System Using a Large Language Model Capable of Handling Ambiguous Instructions

山尾晃世[1*]　金岡大樹[1]　磯本航世[1]　田向権[1,2]
Kosei Yamao[1],　Daiju Kanaoka[1],　Kosei Isomoto[1],　Hakaru Tamukoh[1,2]

[1] 九州工業大学大学院生命体工学研究科
[1] Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Japan
[2] 九州工業大学ニューロモルフィック AI ハードウェア研究センター
[2] Research Center for Neuromorphic AI Hardware, Kyushu Institute of Technology, Japan

**Abstract:** 人の言語指示から適切な動作を行う汎用的なサービスロボットの実現のためには，高精度な動作計画が必要となる．大規模言語モデルを活用した SayCan と呼ばれる動作計画システムは非常に高精度であるが，いくつかの問題を有している．我々は，SayCan の有する問題の中でも，具体的な対象を特定できない抽象名詞を含む曖昧な言語指示に対して，動作計画の精度が低下する問題と，システムが持つスキルの数に伴い推論時間が増加する問題に注目した．我々は，SayCan をベースとし，抽象名詞を聞き返す機能とルールベースでスキルを抽出する機能を導入した動作計画システムを提案する．提案システムは，言語指示の解釈を容易にし，適切な動作計画を可能とした．また，RoboCup@Home と呼ばれるホームサービスロボットの性能を評価する競技会において高い成績を獲得することにより，実環境下でも十分動作することを示した．

## 1　緒言

昨今，世界的な少子高齢化社会の到来を背景に，家庭内での生活を支援するホームサービスロボットの需要が高まっており，開発・研究が盛んに行われている [1–4]．人の言語指示から適切な動作を行う汎用的なサービスロボットの実現のためには，高精度な音声認識と適切な動作計画が必要となる．動作計画においては，世界の一般的な知識を活用することで，高精度な出力を行うことが可能である．そのため，一般的な知識を用いる動作計画に関する様々な研究がされている．

近年では，大規模言語モデル（Large Language Model: LLM）が注目を集めている [5]．LLM は数十億以上のパラメータを持ち，インターネット上に存在する大規模なデータを学習することで，様々な言語タスクにおいて高い性能を示し，世界の一般的な知識を有していると考えられている [6]．また，LLM は言語タスク以外でも高い性能を発揮しており，物体認識や動作計画システムにも応用されている．LLM を用いた動作計画として，Google は SayCan [7] を提案している．SayCan は，Say モジュールと Can モジュールの 2 つで構成されている．ロボットの各動作を表すスキルセットは予め与えれており，Say モジュールでは言語指示をもとに各スキルの尤度を推定し，Can モジュールでは現在のロボットの状態をもとに各スキルの尤度を推定する．それぞれのモジュールから出力される各スキルの尤度を基に次に実行するスキルを決定する．SayCan は高い性能を発揮しているが，その課題として，Can モジュールはロボットから得た大量の動作データを用いた強化学習で実装されているため，他のロボットへの応用が困難である．また，fruit や drink などの具体的な対象を特定できない抽象名詞を含むコマンドに対して精度が低下する問題と，システムが持つスキル数に伴い推論時間が増加する問題を有している．

以上より，本研究では，SayCan をベースとし，抽象名詞を聞き返す機能とルールベースでスキルを抽出する機能を導入したホームサービスロボットの動作計画システムを提案する．提案システムは，命令認識では，与えられたコマンドの中に抽象名詞があった場合，その具体的な名称を聞き返し返答結果と抽象名詞を置き

*連絡先：九州工業大学大学院生命体工学研究科人間知能システム工学専攻
〒 808-0135 福岡県北九州市若松区ひびきの 2-4
E-mail: yamao.kosei665@mail.kyutech.jp


図 1: HSR の外観と主なデバイス

換えたコマンドを認識結果とする．次に，ルールベースによる制約を用いて，予め用意したスキルセットから，スキル候補を抽出する．コマンドの認識結果から抽出した物体名などのキーワードとスキル候補を組み合わせてタスク候補を生成する．最後に，各タスク候補の尤度を LLM を用いて出力し，最も尤度が高いタスクを実行する．

本研究の貢献は以下である．

- 人間の話ことばを解釈を容易にし，より高精度な動作計画を可能にした
- ホームサービスロボットの性能を評価する競技会で高い性能を示し，実世界でも十分動作することを明らかにした

# 2 関連研究

## 2.1 RoboCup@Home

ホームサービスロボットの技術発展を目的に開催されている RoboCup@Home [8] と呼ばれる国際的な競技会がある．本競技会は，人間とロボットの協調を目標の一つに掲げており，音声認識や物体認識，ナビゲーション，マニピュレーションに関する競技が動的環境下で実施される．そのため，現実に近い家庭環境でロボットの性能を評価することができ，世界中で注目を集めている．RoboCup@Home は，使用するロボットの違いにより複数のリーグに分かれており，Domestic Standard Platform League (DSPL) では，トヨタ自動車株式会社が開発した Human Support Robot（HSR）[9] を標準機として採用し，ソフトウェアの性能のみを評価している．図 1 に，HSR の外観と搭載されている主なデバイスを示す．

RoboCup@Home では，GPSR とより難易度の高い EGPSR という競技がある．GPSR・EGPSR は，実際のロボットが自然言語による様々なコマンドを聞き，日常生活の環境において適切な行動を実行するというタスクである．ロボットへのコマンドは，RoboCup@Home が公開しているランダムコマンド生成器 [10] から出力されたものを用いる．ランダムコマンド生成器では，“Please find the fruits in the dishwasher” のように “fruits” などの対象を特定できない抽象名詞を含むコマンドが出力される可能性がある．GPSR と EGPSR の違いは，ランダムコマンド生成器から出力されるコマンドの難易度が異なる点である．EGPSR で用いるコマンドは，ドアの開閉などの難易度の高い動作が要求され，GPSR で用いるコマンドと比較して内容が複雑になっている．

## 2.2 SayCan

SayCan は，ロボットができるスキルを予め設定しており，与えられたコマンドからどのスキルを実行すればいいかを判断する．SayCan は Say モジュールと Can モジュールで構成されている．

Say モジュールは，LLM が持つ文章中の各単語の尤度を出力する機能を用いて，次に実行する可能性の高いスキルを予測する．与えられたコマンドとスキルの説明文を LLM に入力し，各単語の尤度を合算することで，最も尤度の高いスキルを実行する．論文中では，SayCan は，同じく Google が開発した PaLM [11] と呼ばれる LLM を用いて予測を行っている．

Can モジュールでは，カメラやセンサなどから外界の情報を取得し，どのスキルが実行できる可能性が高いかを予測する．例えば,「リンゴを持ってきて」というコマンドにおいては，カメラの画角内にリンゴがあれば,「リンゴを把持する」というスキルの尤度が高くなり，逆に画角内になければ「リンゴを把持する」の尤度は低くなり,「キッチンに行く」や「リンゴを探す」などの尤度が高くなる．

これら 2 つのモジュールの結果を統合して，次に実行するスキルを予測する．Can モジュールは強化学習を用いて学習を行っており，実際には大量のデータセットを準備する必要があり多くの時間を要するため，実装は非常に難しい．SayCan は，Can モジュールから出力される尤度を使わずに，Say モジュールから出力される尤度だけでも非常に高精度な動作計画を行うことが可能である．

また，“Bring me a fruit” などの具体的な対象を特定できない抽象名詞を含むコマンドに対して精度が低下することが，論文内の実験結果から確認できる．また，システムが持つスキル数に伴い推論時間が増加する問題を有している．

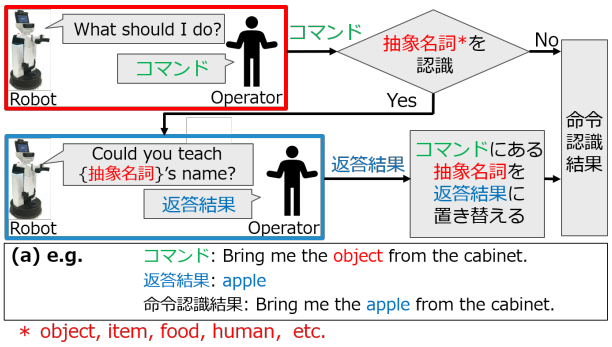



図 2: 命令認識の概略図

### 2.3 その他の動作計画システム

Obinata らが提案する動作計画システム [12] は，オープンボキャブラリーな物体認識を実現するため，Zero-shot での物体認識が可能なモデルを導入した．このモデルを導入することで，未知物体を認識するスキルをシステムに組み込み，より柔軟な動作計画を可能とした．このシステムは，RoboCup@Home JapanOpen2022 の GPSR タスクにおいて，高得点を獲得し，高い安定性を示している．

Shirasaka らの提案する動作計画システム [13] は，ロボットが動作を失敗するなどの障害が発生した場合に，対処するための自己回復機能を導入した．また，複数の基盤モデルと呼ばれる大量で多様なデータを用いて訓練され，様々なタスクに適応可能な大規模モデルを活用することで，動作計画の性能を向上させている．このシステムは，GPSR タスクにおいて，RoboCup@Home JapanOpen2023 で優勝し，RoboCup@Home2023 で 2 位を獲得することで非常に高い性能を示している．

## 3 提案システム

本章では，提案するシステムについて述べる．提案システムは命令認識と動作計画で構成されている．

### 3.1 命令認識

図 2 に，提案システムにおける命令認識の機能の概略図を示す．提案システムは，与えられたコマンドの中に，事前にデータベースに登録している object や item，food などの抽象名詞があった場合，その具体的な名称を聞き返す．“object” などの抽象名詞を含む曖昧なコマンドでは，ロボットが抽象名詞の対象を認識できない可能性があるため，認識可能な具体的な名称を取得するために行う．例として，図 2-(a) に示すように，抽象名詞である “object” の具体的な名称を聞き返し，返答結果に応じて，“object” を “apple” に置き換える．

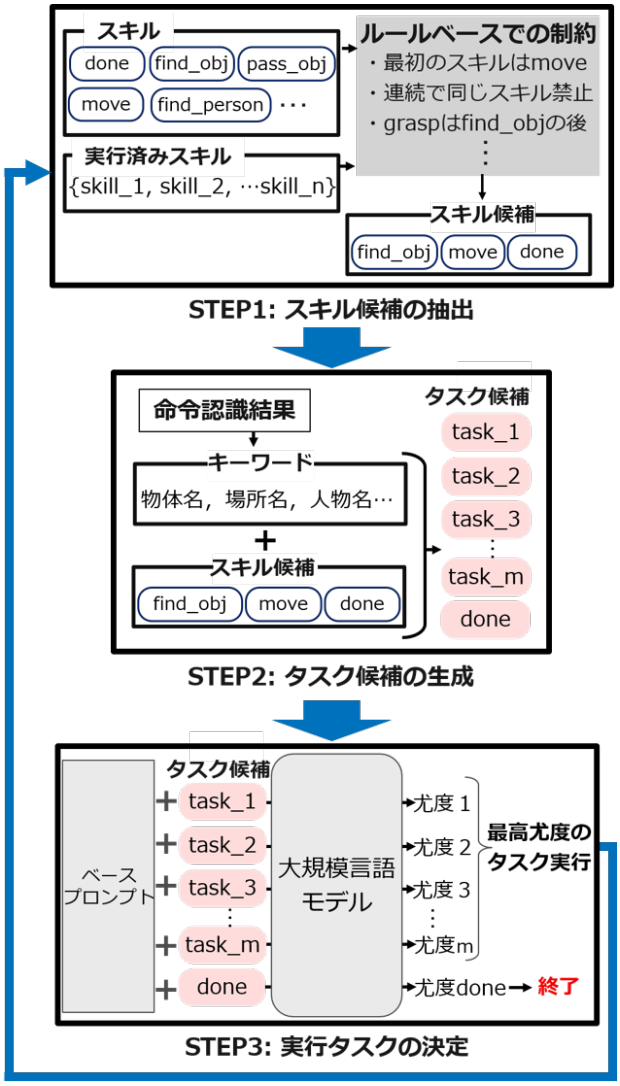



図 3: 動作計画の概略図

### 3.2 動作計画

図 3 に提案システムにおける動作計画の機能の概略図を示す．動作計画は大きく 3 つの処理に分けられる．

**STEP1：スキル候補の抽出** スキルセットから，ルールベースでの制約により，動作の流れに則したスキル候補を抽出する処理を行う．制約では，“最初に実行するスキルは `move`” や “連続で同じスキルは禁止”，“`grasp` は `find_obj` の後” などのスキルの順序による制約を行う．これにより，LLM に入力する回数を減らし，推論時間の短縮を行っている．また，極めて可能性の低いスキルを除外することで動作計画の精度向上にも繋がると考えられる．また，使用するスキルセットを表 1

表 1: 提案システムで用いたスキルセット

| スキル | タスク | 動作説明 |
|---|---|---|
| move | go to the {PLACE} | {PLACE} に移動 |
| follow | follow the target | 人を追跡 |
| find_obj | find the {OBJECT} on the {PLACE} | {PLACE} の {OBJECT} を探索 |
| find_person | find {PERSON} | {PERSON} を探索 |
| observe_obj | look at the {PLACE} to check objects | {PLACE} にある物体の名前などを取得 |
| observe_person | look at the {PLACE} to check people | {PLACE} にいる人の位置などを取得 |
| grasp_obj | grasp the {OBJECT} | {OBJECT} を把持 |
| put | take the {OBJECT} to the {PLACE} | {PLACE} に {OBJECT} を置く |
| pass_obj | pass the {OBJECT} | {OBJECT} を渡す |
| answer_question | answer a question | 人からの問いに答える |
| say | say {} | LLM から出力された文章を発話 |
| done | done | 動作計画を終了 |

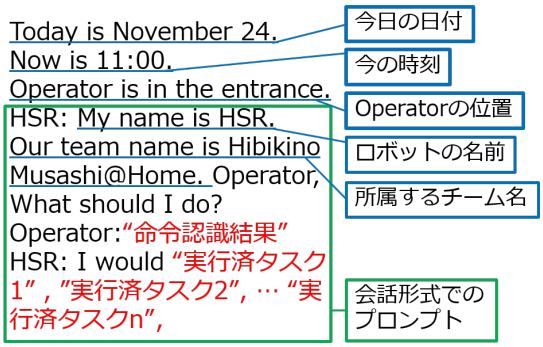



図 4: ベースプロンプトの内容

に示す．このスキルセットには，ホームサービスロボットが家庭内での様々なタスクを遂行するための基本的なスキルを有している．say スキルでは，LLM で発話する内容を生成し補完する．

**STEP2：タスク候補の生成** 命令認識結果から，事前にデータベースに登録している物体名や場所名，人物名をキーワードとして抽出する．その後，キーワードとスキル候補を組み合わせて，タスク候補を生成する．タスクは表 1 で示す形式で，生成される．

**STEP3: 実行タスクの決定** 図 4 に示すベースプロンプトと，タスク候補を組み合わせて LLM に入力することで，各タスク候補の尤度を求める．その後，最も高い尤度であるタスクを採用し実行する．ベースプロンプトでは，日付や時刻などの情報に加え，会話形式のプロンプトを通して，命令に従い実行する役割を LLM に与えている．

第 3.1 節で述べた命令認識の処理とこれら 3 つの処理を実施することで，抽象名詞を含む曖昧な言語指示に対応可能な動作計画が実現できる．

# 4 実験

## 4.1 概要

本研究では以下に示す実験を行い，提案システムの評価を行った．

1. コマンド生成器を用いて出力したコマンドによる評価実験
2. RoboCup@Home2023 で行われた実際の競技を用いた評価実験

実験 1 では，まず，RoboCup@Home の提供するコマンド生成器を用いて GPSR と EGPSR のコマンドをそれぞれランダムに 100 個ずつ生成した．本実験は，RoboCup@Home2023 を想定しており，生成されたコマンドに含まれる物体名や場所名は，RoboCup@Home2023 で使用されたものである．生成したコマンドを提案システムに入力し，その出力として得られた動作計画が，与えられたコマンドを実現可能であるかを評価した．具体的には，動作計画に含まれるスキルの順序や対象が，対応するコマンドを達成できるかどうかを判定した．

実験 2 では，2023 年 7 月にフランス・ボルドーで開催された RoboCup@Home2023・DSPL に Hibikino-Musashi@Home (HMA) として参加し，GPSR・EGPSR 競技において，提案システムを動作させ評価を行った．

また，実験条件として提案システム中の LLM は OpenAI 社の text-davinci-003 [14] を用いた．

## 4.2 実験 1：動作計画の評価結果

結果として，GPSR においては，61 個のコマンドにおける動作計画が実現可能であり，EGPSR において

表 2: 動作計画に失敗したコマンドの例と原因

| コマンド | 原因 |
|---|---|
| Tell me which are the three biggest objects on the desk | three biggest objects に対応できない |
| Deliver drinks to everyone in the kitchen | 命令の対象が複数では対応できない |
| Go to the desk, look for the mug, and place it on the sink | grasp が明示されていない |
| Could you please close the entrance door | ドアを開け閉めするスキルがない |
| Hand me some coke in a mug | 液体などを注ぐスキルがない |
| Place a mug on the desk and a knife on its left | 物の横に置くことができない |

は，27 個のコマンドにおける動作計画が実現可能であることを確認した．特に，コマンドにある抽象名詞を聞き返すことで対象となる具体的な名称を獲得し，動作計画を成功させることを確認した．例えば，“Please bring me the fruit on the desk” というコマンドでは，fruit という抽象名詞を聞き返し，apple という返答を得ることで，抽象名詞を置換させ，動作計画を成功させていることを確認した．動作計画が誤っていたコマンドの例と考えられる原因を表 2 に示す．

### 4.3 実験 2：RoboCup@Home2023 での競技結果

GPSR では，“bring me the object behind the lemon from the cabinet” というコマンドが出題された．このコマンドに対して，ロボットは抽象名詞である object の具体的な名称を聞き返した．Operator から tropical juice という返答がきたが，音声認識に失敗してしまい，抽象名詞の置換がされなかった．そのため，Operator に渡す物体を lemon であると判断し，動作計画に失敗した．これは，物体を Operator に持ってくるという点は成功しているため，部分的に点数を獲得した．

EGPSR では，“get acquainted with Morgan at the exit, then find him in the living room please” というコマンドに対して，指示通り，exit で Morgan を発見した．その後，Morgan に対して，“Hello, Morgan. I’m HSR from Hibikino Musashi@Home. Nice to meet you.” と発話した．最後に living room に移動し Morgan を発見した．このコマンドは成功したと判断され点数を獲得した．

表 3 に RoboCup@Home2023 での GPSR と EGPSR における上位 3 チームと点数を示す．結果的に HMA は，GPSR では 3 位，EGPSR では 1 位の成績を獲得し，実環境下でも十分動作することを示した．

表 3: RoboCup@Home2023 の結果

| | GPSR | | EGPSR | |
|---|---|---|---|---|
| | チーム名 | 点数 | チーム名 | 点数 |
| 1 位 | Tidyboy | 400 点 | **HMA** | **700 点** |
| 2 位 | TRAIL | 300 点 | TRAIL | 400 点 |
| 3 位 | **HMA** | **200 点** | Tidyboy | 300 点 |

## 5 考察

GPSR のコマンドにおいては，61 個の動作計画を成功した．Obinata らの動作計画システム [12] における同様の実験では，59 個のコマンドにおいて実行可能な正しい動作計画が出力された．僅かに我々の実験結果が上回り，提案システムの有効性を示した．EGPSR のコマンドにおいては，成功した数が 27 個と非常に少なかった．EGPSR のコマンドでは，ドアの開閉や液体を注ぐ動作，相対的な位置に対して物体を置くといった難易度の高いスキルを要求され，内容もより複雑であるため，動作計画の成功率が著しく低下したと考えられる．

また，表 2 に示すようなコマンドにおける動作計画の失敗が多かったことが確認できた．動作計画に失敗した原因は大きく 3 つに分けられる．1 つ目は，既存のスキルの不完全さである．“Tell me which are the three biggest objects on the desk” というコマンドにおいては，現在のスキルでは最も大きな object の情報を 1 つを取得し，伝えることは可能だが，複数を対象にすることは対応できていない．2 つ目は，必要な動作が明確に含まれていないコマンドでは，動作計画に失敗する場合がある点である．“Go to the desk, look for the mug, and place it on the sink” というコマンドにおいては，place という指示があるため，grasp_obj を行わなければいけない．しかし，コマンドに明示的に grasp を促す文章がないため grasp の尤度が高くならず，誤った動作計画を出力した．3 つ目は，必要なスキルの欠如である．“Could you please close the entrance door” というコマンドにおいては，ドアを開閉するスキルを作成していないため，動作計画に失敗した．

また，ロボットが有するスキルと対象となる物体名や場所名の組み合わせは非常に多く，全てを LLM に入力すると推論時間が大きくなる．ルールベースの制約を導入することで，LLM に入力するスキル数を減ら

すことができ，推論時間の短縮が可能であると考えられる．

## 6 結言

本研究では，SayCan をベースとし，抽象名詞を聞き返す機能とルールベースでスキルを抽出する機能を導入したロボットの動作計画システムを提案した．提案システムは，抽象名詞を含むコマンドに対し，対象となる具体的な名称を聞き返すことで，より高精度な動作計画を実現した．また，ルールベースの制約を設け，LLM に入力するタスク数を減らすことで，推論時間の短縮を行った．我々は，提案システムを用いた実験や競技会での評価を通して，システムの有効性を示した．また，提案システムの課題として，既存スキルの不完全さや，必要なスキルの欠如，必要な動作が明確に含まれていないコマンドでは，動作計画に失敗する場合がある点が挙げられた．より高精度な動作計画システム実現に向けて，ドアの開閉を行うスキルなどの開発や，LLM の有効性を最適化するためのプロンプトの工夫 [15]，カメラ画像などの外界の情報を取り入れた動作計画が必要であると考える．また，実際の家庭環境で動作する上では，LLM が有する一般的な知識だけでなく，各家庭が持つ家族の好みや習慣などの固有の知識も必要である．この固有の知識を獲得し，将来の動作に組み込むことを目指した研究も行われている [16,17]．今後の展望として，固有の知識を今回の提案システムに組み込むことで，抽象名詞をその都度聞き返すことなく，ロボットが以前に獲得した固有の知識を基に動作することを目指す．

## 謝辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務 (JPNP16007) の結果得られたものです．また，本研究は JSPS 科研費 23H03468，23K18495 の助成を受けたものです．

## 参考文献

[1] 株式会社富士経済：2023 年版 ワールドワイドロボット関連市場の現状と将来展望 サービスロボット編 (2023).

[2] Yano, Y., Isomoto, K., Ono, T. and Tamukoh, H.: Autonomous Waiter Robot System for Recognizing Customers, Taking Orders, and Serving Food, in *Proceedings of the 26th RoboCup International Symposium* (2023).

[3] Ono, T., Kanaoka, D., Shiba, T., Tokuno, S., Yano, Y., Mizutani, A., Matsumoto, I., Amano, H. and Tamukoh, H.: Solution of World Robot Challenge 2020 Partner Robot Challenge (Real Space), *Advanced Robotics* (2022).

[4] Isomoto, K., Yano, Y., Tanaka, Y. and Tamukoh, H.: Robust Trash Can Lid Opening System, in *Proceedings of the 2023 International Workshop on Smart Info-Media Systems in Asia (SISA)* (2023).

[5] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J. and Wen, rong J.: A Survey of Large Language Models, *arXiv preprint arXiv:2303.18223* (2023).

[6] Petroni, F., aschel, T. R., Riedel, S., Lewis, P. S. H., Bakhtin, A., Wu, Y. and Miller, A. H.: Language models as knowledge bases?, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473 (2019).

[7] Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R. J., Jeffrey, K., Jesmonth, S., Joshi, N. J., Julian, R., Kalashnikov, D., Kuang, Y., Lee, K.-H., Levine, S., Lu, Y., Luu, L., Parada, C., Pastor, P., Quiambao, J., Rao, K., Rettinghouse, J., Reyes, D., Sermanet, P., Sievers, N., Tan, C., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Xu, S., Yan, M. and Zeng, A.: Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, *arXiv preprint arXiv:2204.01691* (2022).

[8] RoboCup@Home, `https://www.robocup.org/domains/3`(Accessed on 29/08/2023).

[9] Yamamoto, T., Terada, K., Ochiai, A., Saito, F., Asahara, Y. and Murase, K.: Development of Human Support Robot as the research platform of a domestic mobile manipulator, *ROBOMECH Journal*, Vol. 6, No. 1, pp. 1–15 (2019).

[10] RoboCup@Home Command Generator, `https://github.com/kyordhel/GPSRCmdGen`(Accessed on 29/08/2023).

[11] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S. and Fiedel, N.: PaLM: Scaling Language Modeling with Pathways, *Journal of Machine Learning Research*, Vol. 24, No. 240, pp. 1–113 (2023).

[12] Obinata, Y., Kanazawa, N., Kawaharazuka, K., Yanokura, I., Kim, S., Okada, K. and Inaba, M.: Foundation Model based Open Vocabulary Task Planning and Executive System for General Purpose Service Robots, *arXiv preprint arXiv:2308.03357* (2023).

[13] Shirasaka, M., Matsushima, T., Tsunashima, S., Ikeda, Y., Horo, A., Ikoma, S., Tsuji, C., Wada, H., Omija, T., Komukai, D., Matsuo, Y. and Iwasawa, Y.: Self-Recovery Prompting: Promptable General Purpose Service Robot System with Foundation Models and Self-Recovery, in *Conference on Robot Learning 2023 Workshop on Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition* (2023).

[14] OpenAI GPT-3.5 API [text-davinci-003], `https://platform.openai.com/docs/models/gpt-3-5`(Accessed on 07/09/2023).

[15] Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V. and Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 24824–24837 (2022).

[16] Tanaka, Y., Tamukoh, H., Tateno, K., Katori, Y. and Morie, T.: A Brain-inspired Artificial Intelligence Model of Hippocampus, Amygdala, and Prefrontal Cortex on Home Service Robots, in *Proceedings of the 2020 International Symposium on Nonlinear Theory and Its Applications (NOLTA)*, pp. 138–141 (2020).

[17] 水谷 彰伸, 田中 悠一朗, 田向 権, 立野 勝巳, 野村 修, 森江 隆：大規模言語モデルと海馬モデルによるホームサービスロボット向け知識獲得システム, 電子情報通信学会スマートインフォメディアシステム研究会（SIS）, 第 123 巻, pp. 13–18 (2023).

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-05

# Back Translation in Sign Language Generation

Khan Nabeela Khanum*[1], Tan Sihan[1], Itoyama Katsutoshi[1,2], and Nakadai Kazuhiro[1]

[1] Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology
[2]Honda Research Institute Japan Co.Ltd.

**Abstract**—This survey paper provides a comprehensive overview of the evolving landscape of Sign Language Back Translation (SLBT) for Sign Language Generation (SLG). BT is utilized to convert spoken text into gloss, a written form of SL, serving as a crucial intermediary step in SLG. The paper explores various SLBT paradigms, highlighting key studies and encapsulating SLBT's trajectory. It addresses challenges and promising advancements in back translation, aiming for seamless communication between spoken and signed languages.

## 1. INTRODUCTION

The application of sign languages (SLs) within the deaf community has become an established mode of communication. Sign languages utilize multiple channels, incorporating both manual and non-manual features. to break the linguistic and cultural barrier between hearing and deaf communities, translation plays an important role However, according to [1], manual translations is too expensive and a rather difficult to practice in daily life. In this case, the most affordable means of communication is using Machine Translation (MT)[2]. Unlike spoken languages, sign languages lack standardized written scripts, rendering advanced Machine Translation (MT) models designed for text-based languages unsuitable for SLs. To address this challenge, various writing methods have been introduced for SLs, including Glosses [3], SignWriting[4], HamNoSys[5], Stokoe Notations[6], and Si5s[7]. Among these methods, glosses have emerged as the most popular choice. Glosses involve labeling signs with words from the corresponding spoken language, often including affixes and markers, providing a bridge between the visual nature of sign languages and the written form of spoken languages.

Glosses play a vital role in various sign language (SL) processing tasks like recognition, translation, and generation. Despite their limitations in capturing the full linguistic richness of SL [8], they are crucial in Sign Language Translation (SLT) applications, facilitating communication between hearing and deaf communities, especially in education and interpretation. Additionally, glosses provide valuable parallel data for MT training in SL processing research. Researchers often employ glosses as intermediaries, translating between SL and spoken language. For instance, during SL to spoken language or vice versa translation, gloss-to-text or text-to-gloss conversion acts as an intermediary step. When generating SL from spoken language, the initial process involves converting text to glosses, which are then used for SL generation. This paper specifically focuses on the initial stage for SL generation, i.e converting text to glosses.

Several research initiatives in MT have been undertaken in the past for the back translation (BT) of SL in order to generate SL, showcasing the innovative efforts in the realm of sign language accessibility. One notable project, ViSiCAST, funded by the European Union, aimed to enhance accessibility for deaf citizens using virtual animation or avatars presenting sign language [9]. Another significant European venture, the eSIGN Project, contributed to this mission by developing avatar-based technology for American Sign Language. DePaul University played a pivotal role in this initiative, creating an avatar named 'Paula' capable of conveying all linguistic parameters of ASL while translating English to ASL [10]. Additionally, the ProDeaf project emerged as a pioneering effort, converting Portuguese text and voice into Portuguese Sign Language (LIBRAS) to facilitate seamless communication between the deaf and hearing communities [11].

In this paper, our focus is on the exploration of sign language back translation from text to gloss. The structure of the paper is organized as follows. Section 2 elucidates the fundamental distinctions between sign language and spoken language and presents a review of related works in machine translation for sign language back translation (SLBT) . Section 3 outlines the evaluation metrics employed in various machine translation approaches. Section 4 encompasses a discussion on the current state of research, including its limitations, and outlines potential future directions for further investigation.

## 2. RELATED WORKS

### 2·1 Sign language v/s Spoken language

SL and spoken language exhibit fundamental differences in lexicons, grammar rules, and structure. Contrary to common misconceptions, SL is not a universal language; rather, each country possesses its distinct SL, such as American Sign language , German Sign Language, Japanese Sign Language, Indian Sign Language, Chinese Sign Language, and Greek Sign Language [12]. In the realm of communication, SL

relies on visual transmission, utilizing vision power instead of hearing power [13]. The mode of expression in SL involves intricate components, including single or both hands, facial expressions, and body movements. Notably, SL deviate from spoken languages in multiple aspects, ranging from grammatical variances to structural disparities and word order differences [14]. elucidate challenges in translating from Spanish to Spanish Sign Language, emphasizing issues like mapping semantic concepts to specific signs or generating multiple signs from one concept [15]. The translation process from Arabic text to Arabic Sign Language also faces hurdles due to grammatical rule discrepancies and differences in word order between the source and target languages [16]. Sequentiality is another significant distinction, where spoken languages follow a phonemic sequence, whereas SL incorporate non-sequential components concurrently, involving fingers, hands, and facial expressions [14]. This simultaneous nature is exemplified in Thai Sign Language, where the linguistic structure deviates from the linear organization of the Thai language [17]. Specific features unique to sign languages include non-manual components, the utilization of space as a lexical element, varied parts of speech represented by the same sign, the use of classifiers with morphological value, and diverse sentence structures [18]. These differences underscore the richness and complexity of sign languages, challenging preconceptions and highlighting the need for specialized linguistic understanding and translation methodologies.

### 2·2 Sign Language Back Translation (SLBT) methods

#### 2·2.1 Rule based Machine translation (RBMT)

The rule-based approach to text-to-gloss translation, as discussed in [19], is an early method in machine translation. Rule-Based Machine Translation (RBMT) relies on predefined linguistic rules for both source and target languages. It involves morphological, syntactic, and semantic analyses in both languages. The RBMT system includes components such as source language morphological analysis, parsing, translation, target language morphological generation, and final parsing. The Vauquois Pyramid Fig. 1 illustrates the complexity of rule-based approaches, representing various linguistic levels. While structured, RBMT systems face challenges with idiomatic expressions but form the foundation for machine translation techniques in sign languages.

#### 2·2.2 Corpus based Machine translation (CBMT)

Corpus-Based Machine Translation (CBMT) relies on bilingual text corpora for generating translations. While Rule-Based Machine Translation (RBMT) systems can produce accurate translations, they are

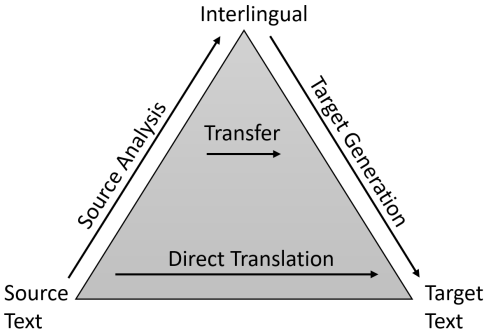


Fig. 1 Vauquois Pyramid

labor-intensive to develop, requiring manual crafting of linguistic resources and continuous rule additions, making the process time-consuming. In contrast, CBMT systems, also known as data-driven machine translations, leverage large bilingual datasets for translation.

**Example based Machine Translation (EBMT)**

Example-Based Machine Translation (EBMT), a concept first introduced by Makoto Nagao in 1984, relies on bilingual parallel corpora for its training, containing sentence pairs from both languages [20]. This approach has found pioneers in Sara Morrissey and Andy Way, who applied EBMT to Sign Language Machine Translation (SLMT) systems. Morrissey et al. employed the Marker Hypothesis to translate English to Dutch Sign Language, demonstrating its promise in segmenting English input text into chunks alignable with Sign Language annotations [21, 22]. Notably, ELAN annotation tool was utilized for sign language corpora, facilitating accurate alignment between English text and sign annotations. EBMT has also been successfully applied to languages with smaller corpora, such as Arabic Sign Language, where Almohimeed et al. utilized a 203-sentence corpus to translate Arabic text into Arabic Sign Language. The EBMT system operated on text chunks aligned with corresponding signs and, despite its limitations due to the quality of examples, yielded a word error rate (WER) of 46.7% and position-independent word error rate (PER) of 29.4% [23]. Additionally, researchers like Boulares et al. combined EBMT with genetic algorithms and fuzzy logic to translate English into American Sign Language (ASL). By integrating global and local alignment algorithms, they achieved effective proximity searches between words, demonstrating EBMT's potential for capturing complex linguistic structures [24]. For languages like Turkish Sign Language (TSL), where the grammar is poorly understood and datasets are limited, Selcuk-Simsek et al. proposed a bidirectional EBMT approach. Their system, incorporating a lexical supervision component (LSC) with morphological analyzers and disambiguation tools, achieved a BLEU score of 43% and a TER

score of 38% using k-fold cross-validation [25]. Although EBMT excels in limited datasets, it faces challenges in scalability due to the need for a substantial number of high-quality examples. As detailed in Table 5, this approach's suitability diminishes for larger datasets, prompting a shift to explore the implementation of Statistical Machine Translation for such scenarios.

**Statical Machine Translation (SMT)**

Statistical Machine Translation (SMT) emerges as a significant player within the Corpus-Based Machine Translation (CBMT) paradigm. SMT, rooted in probability distributions and Bayesian approaches, operates efficiently with large bilingual corpora. Early pioneers such as Koehn et al. laid foundational work, exploring word alignment and various phrase translation methods [26, 27]. However, challenges surfaced, especially in smaller-scale applications like translating German text into German Sign Language (DGS) [28]. Researchers like [29] addressed these hurdles, employing morpho-syntactic analysis to enhance translation quality, leading to a notable 9% improvement [**?**]. Data scarcity posed a significant obstacle, prompting innovative solutions such as introducing thematic roles to capture verb meanings, as demonstrated in Chinese to Taiwanese Sign Language translation [30]. Syntactic-semantic information integration further enhanced translation accuracy, as seen in Spanish to LSE translation, where modules like categorization and Factored Translation Modules (FTMs) boosted BLEU scores significantly [31]. However, challenges persisted, exemplified in translating Indian Sign Language glosses, where existing models faced limitations [32]. Additionally, Turkish Sign Language (TID) translation efforts showcased diverse approaches, including stemming and semantic tagging, but lacked manual evaluation [30]. These endeavors underscore the ongoing pursuit to refine SMT for sign languages, emphasizing the need for innovative strategies and hybrid approaches to address data limitations and enhance translation outcomes.

**Hybrid Machine Translation (HMT)**

The integration of multiple machine translation systems within a single framework, known as Hybrid Machine Translation (HMT) systems, has emerged as a critical advancement. Addressing the limitations of single machine translation systems, researchers have explored diverse hybrid approaches, combining methodologies like example-based, transfer-based, knowledge-based, and statistical translation. For instance, Hogan et al. combined various translation sub-systems, showcasing the potential of hybridization [33]. Wu et al. adopted a hybrid model by merging rule-based and statistical approaches for translating Chinese to Taiwanese Sign Language, demonstrating substantial progress but highlighting challenges related to corpora size and extensibility [34]. Morrissey et al. contributed significantly to the field, utilizing MaTrEx Machine Translation system and integrating Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT) methodologies. While achieving satisfactory results, challenges persisted in achieving natural sign language animations [35, 36]. San-Segundo et al. developed a comprehensive HMT approach incorporating rule-based, example-based, and statistical translators, enhancing the translation quality manifold. Their hierarchical structure and combination of techniques significantly improved results, outperforming individual techniques [37]. Additionally, researchers like Lopez-Ludena et al. refined hybrid systems, automating module generation and utilizing advanced translation strategies, producing remarkable improvements and paving the way for future advancements in SLMT [38, 39, 32]. These diverse hybridization efforts underscore the ongoing pursuit of enhancing SLMT, integrating advanced technologies, and addressing challenges, aiming for more accurate and natural sign language translations.

### 2·2.3 Neural Machine Translation (NMT)

Neural Machine Translation (NMT) stands as a transformative approach, utilizing artificial neural networks to predict word sequences. Manzano et al. employed NMT to translate English to American Sign Language (ASL) using the ASLG-PC dataset, yielding ASL glosses as output, albeit facing challenges due to a limited vocabulary size [40]. ATLASLang, translating Arabic to Arabic Sign Language, adopted NMT and surpassed previous systems with a BLEU score of 0.79 [41]. Text2Sign, an NMT system, utilized a Generative Adversarial Network and Motion Generation to produce sign videos from spoken language, showcasing robustness despite challenges related to avatar-based approaches [42]. Saunders et al. proposed an NMT approach focusing on automatic sign language production, resulting in enhanced Sign Language Production (SLP) performance [43]. Recognizing the importance of non-manual features, Saunders et al. extended their approach, encapsulating all sign articulators [44]. Ventura et al. advanced the field by generating realistic signing videos using the SIGN-GAN approach, outperforming baseline systems both quantitatively and in human perception evaluations [45]. These studies highlight the evolving landscape of NMT in SLMT, underscoring both advancements and areas for further exploration

## 2·3 Datasets

For SLBT, a diverse array of datasets plays a pivotal role in advancing the field. These datasets serve as the foundation for training and evaluating models capable of translating sign language expressions back into spoken language. Among the prominent datasets, RWTH-PHOENIX-Weather-2014T (PHOENIX14T)

[46] stands out as a widely employed resource, offering German sign language videos, gloss, and spoken language text. This dataset, segmented into parallel sentences, has become a benchmark for evaluating baseline models in sign language translation (SLT). CSL-Daily [47], a notable dataset for Chinese SLT, covers a spectrum of themes with sign language videos featuring normative and natural expressions. RWTH-PHOENIX-Weather 2014 (PHOENIX14) focuses on German sign language videos sourced from weather news programs, providing valuable content for SLT endeavors. ASLG-PC12 [48], despite lacking sign language videos, presents a massive repository of gloss–text pairs, particularly suited for Gloss-to-Text tasks. Lastly, Spreadthesign-Ten (SP-10), a multilingual sign language recognition dataset, contributes to the broader understanding of sign languages by encompassing videos and corresponding texts from various linguistic backgrounds. These datasets collectively form a rich resource landscape, enabling researchers to explore and enhance sign language back translation techniques, ultimately fostering more inclusive and effective communication between signers and non-signers.

Notably, the release of How2Sign [49] in 2021 marked a significant stride. This multimodal and multi-view continuous American Sign Language (ASL) dataset introduces new dimensions to the field. While previous studies predominantly relied on the RWTH-PHOENIX-Weather-2014T (PHOENIX14T), How2Sign brings a fresh perspective by providing a broader and more comprehensive dataset for ASL. However, due to the dataset's complexity in data processing, researchers have yet to explore ASL production using deep learning techniques. The How2Sign dataset boasts a duration and vocabulary approximately 7.53 and 5.74 times larger than the commonly used PHOENIX14T, respectively. This significant expansion in both temporal coverage and vocabulary size holds promise for advancing the capabilities of deep learning models in handling the intricacies of ASL production, further enriching the resources available for sign language back translation research.

### 2·4 Evaluation Metrics

Accurate assessment of Sign Language Back Translation (SLBT) systems is crucial for determining their effectiveness. This evaluation encompasses both manual and automatic methods, each offering unique insights into the system's performance.Manual evaluations, drawing on feedback from deaf individuals and Sign Language (SL) experts, provide qualitative perspectives. They assess factors such as usability, appeal, and user satisfaction, offering a holistic understanding of the generated sign language output.

Automatic metrics, tailored to different translation types, include Word Error Rate (WER), Sentence Error Rate (SER), Position Independent Word Error Rate (PER), Translation Error Rate (TER), BLEU, and NIST [50, 51, 52, 53]. BLEU, a precision-based metric, shows a strong correlation with human judgment in machine translation [54].

Advancements in Neural Machine Translation (NMT) systems like ATLASLang and SIGNGAN introduce additional evaluation metrics, such as SSIM, PSNR, and MSE, providing nuanced assessments of synthesized SL images or videos [43, 42]. Complex NMT approaches may incorporate metrics like METEOR and RIBES to assess word order and reordering events [55].

Balancing manual and automatic evaluations is crucial, especially in Rule-Based Machine Translation (RBMT) scenarios with limited corpora. Ongoing research aims to refine and expand performance metrics, ensuring a comprehensive understanding of SLMT system outputs.

## 3. Discussion

The discussion on implications for research and practice in sign language back translation outlines promising avenues for future exploration, categorized into rule-based machine translation, corpus-based machine translation, neural machine translation, and sign generation.

In the realm of rule-based machine translation, addressing the translation of complex sentences and the lack of formal sign language grammar analysis are identified as critical research areas. The discussion underscores the importance of improving accuracy and usability, particularly in handling complex linguistic structures.

Moving to corpus-based machine translation, the challenges of limited bilingual corpora and the need for data acquisition in multiple sign languages emerge as key areas for future investigation. The efficient functioning of corpus-based systems is contingent on extensive datasets, prompting researchers to focus on creating diverse bilingual corpora for a comprehensive range of sign languages. Multilingual efforts, exemplified by existing datasets, like DICTA-SIGN and MultiATIS++ corpus, lay the foundation for future developments, aiming to bridge communication gaps within different deaf communities.

The discussion extends to neural machine translation, where the integration of deep learning and artificial intelligence (AI) is seen as a promising frontier. Notable successes, such as Google Translator's development based on neural machine translation (GNMT), point towards the potential of incorporating AI and deep learning strategies into prevalent systems. This approach holds the promise of achieving similar breakthroughs in text-to-sign translation across multiple languages.

Lastly, the importance of the sign generation sys-

tem is emphasized, highlighting its crucial role in making more information and services available to the deaf community. Recognizing the significance of these avenues for future research, it becomes evident that addressing linguistic and cultural nuances, fostering interdisciplinary collaboration, and refining research methodologies are imperative. The discussion concludes by acknowledging the limitations of the review, notably the exclusion of articles in different languages and the study's focus solely on text-to-sign translation. These limitations underscore the ongoing need for inclusive and iterative approaches in advancing research in sign language back translation.

### 3·1 Limitations

Research on Sign Languages (SLs) reveals inherent challenges that pose obstacles to the development of robust linguistic models and technology. This section discusses two significant challenges: the Scarcity of SL Corpora and Ambiguity in Context.

#### 3·1.1 Scarcity of SL corpora

The under-resourced nature of SLs, categorized as low-density languages, results in a scarcity of technological tools and computerized linguistic resources, such as corpora or lexicons. Corpora play a vital role in linguistic research, providing a corpus of naturally occurring signed language data for analysis and model training. The limited availability of SL corpora impedes the advancement of natural language processing in SLs, hindering the development of computational tools tailored to these languages.

#### 3·1.2 Ambiguity in context

The absence of a standardized writing system for SLs and the reliance on video representations introduce challenges related to the ambiguity of context. Written languages rely on a standardized set of symbols, whereas SLs primarily use video for communication. This lack of a universally accepted writing system limits the creation and analysis of corpora, leading to challenges in disambiguating context in signed language utterances. In sign languages, the richness of facial expressions, body movements, and spatial components can introduce ambiguity in interpreting context. For instance, the same sign may have different meanings based on facial expressions or body language.

### 3·2 Future Direction

Efforts to surmount resource constraints in sign language back translation require a comprehensive strategy encompassing data augmentation, technological advancements, and collaborative endeavors. One key approach involves the expansion of sign language corpora through crowd-sourced initiatives, actively engaging the Deaf community. Collaborative platforms that facilitate the sharing of linguistic resources globally can foster a collective effort to alleviate resource scarcity, supporting the development of more inclusive machine translation models .

In addition to corpus expansion, researchers can explore transfer learning and pretraining models on larger, more general datasets to mitigate the impact of limited linguistic resources. Techniques such as domain adaptation and unsupervised learning can enhance model adaptation to the unique characteristics of sign language .

Complementing these strategies, the integration of data augmentation techniques proves instrumental in both overcoming resource constraints and enhancing translation accuracy. By synthetically expanding the training dataset through variations in signing speed, styles, facial expressions, and body movements, augmented datasets contribute to more robust models.

Moreover, future works should focus on developing context-aware models that account for the non-manual components, facial expressions, and body movements inherent in sign languages. This requires a nuanced understanding of the cultural and linguistic nuances embedded in sign language communication. So, a holistic approach to future research in SLBT should combine advancements in machine learning techniques, collaborative efforts, and an understanding of the unique linguistic aspects of sign languages. Integrating both general techniques for resource expansion and model adaptation, along with specific data augmentation methods, ensures a well-rounded strategy for addressing the challenges posed by resource constraints and enhancing the accuracy of sign language back translation.

## 4. Conclusion

In this paper, we survey the back translation in sign language for text to gloss conversion. We have discussed the importance of sign language back translation in sign language generation. we have also discussed the existing approaches and their limitations. We have explained the possible solutions of these limitations and some future directions with possible application of sign language back translation. In conclusion, this survey presents a comprehensive overview of sign language back translation, encapsulating key challenges and potential avenues for future research. The discussions span rule-based, corpus-based, and neural machine translation, each revealing distinct challenges and opportunities. The imperative to address the translation of complex sentences, formalize sign language grammar, and bridge data gaps in bilingual corpora for diverse sign languages underscores the interdisciplinary nature of this research. The potential breakthroughs promised by integrating deep learning and artificial intelligence into neural machine translation systems pose exciting prospects for advancing text-to-sign translation. The critical role of the sign generation system in fostering inclusivity and

improving services for the deaf community is highlighted. Acknowledging limitations, the study advocates for ongoing refinement, inclusivity, and collaboration in shaping the trajectory of sign language back translation research.

## References


[1] P. Isabelle and G. Foster, "Machine translation: overview," 2006.

[2] J. Porta, F. López-Colino, J. Tejedor, and J. Colás, "A rule-based translation from written spanish to spanish sign language glosses," *Computer Speech & Language*, vol. 28, no. 3, pp. 788–811, 2014.

[3] C. Valli and C. Lucas, *Linguistics of American sign language: An introduction*. Gallaudet University Press, 2000.

[4] A. C. da Rocha Costa and G. P. Dimuro, "Signwriting and swml: Paving the way to sign language processing," *Atelier Traitement Automatique des Langues des Signes, TALN*, vol. 2003, 2003.

[5] K. Kaur and P. Kumar, "Hamnosys to sigml conversion system for sign language automation," *Procedia Computer Science*, vol. 89, pp. 794–803, 2016.

[6] W. C. Stokoe Jr, "Sign language structure: An outline of the visual communication systems of the american deaf," *Journal of deaf studies and deaf education*, vol. 10, no. 1, pp. 3–37, 2005.

[7] R. A. Augustus, E. Ritchie, and S. Stecker, "The official american sign language writing textbook," *Los Angeles, CA: ASLized*, 2013.

[8] E. Pizzuto, P. Rossini, and T. Russo, "Representing signed languages in written form: questions that need to be posed," in *Proceedings of the Workshop on the Representation and Processing of Sign Languages, LREC*, vol. 2006, 2006.

[9] A. Othman and M. Jemni, "Designing high accuracy statistical machine translation for sign language using parallel corpus: case study english and american sign language," *Journal of Information Technology Research (JITR)*, vol. 12, no. 2, pp. 134–158, 2019.

[10] O. H. Al-Barahamtoshy and H. M. Al-Barhamtoshy, "Arabic text-to-sign (artts) model from automatic sr system," *Procedia Computer Science*, vol. 117, pp. 304–311, 2017.

[11] J. Rocha, J. Lensk, T. Ferreira, and M. Ferreira, "Towards a tool to translate brazilian sign language (libras) to brazilian portuguese and improve communication with deaf," in *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 1–4, IEEE, 2020.

[12] M. Alaghband, H. R. Maghroor, and I. Garibay, "A survey on sign language literature," *Machine Learning with Applications*, vol. 14, p. 100504, 2023.

[13] H.-Y. Su and C.-H. Wu, "Improving structural statistical machine translation for sign language with small corpus using thematic role templates as translation memory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1305–1315, 2009.

[14] V. Lopez Ludeña, R. San Segundo Hernández, C. González Morcillo, J. C. López López, and J. M. Pardo Muñoz, "Methodology for developing a speech into sign language translation system in a new semantic domain," 2012.

[15] R. San-Segundo, R. Barra, R. Córdoba, L. F. D'Haro, F. Fernández, J. Ferreiros, J. M. Lucas, J. Macías-Guarasa, J. M. Montero, and J. M. Pardo, "Speech to sign language translation system for spanish," *Speech Communication*, vol. 50, no. 11-12, pp. 1009–1020, 2008.

[16] H. Luqman and S. A. Mahmoud, "Automatic translation of arabic text-to-arabic sign language," *Universal Access in the Information Society*, vol. 18, pp. 939–951, 2019.

[17] S. Dangsaart, K. Naruedomkul, N. Cercone, and B. Sirinaovakul, "Intelligent thai text–thai sign translation for language learning," *Computers & Education*, vol. 51, no. 3, pp. 1125–1141, 2008.

[18] J. Porta, F. López-Colino, J. Tejedor, and J. Colás, "A rule-based translation from written spanish to spanish sign language glosses," *Computer Speech & Language*, vol. 28, no. 3, pp. 788–811, 2014.

[19] D. Kouremenos, K. Ntalianis, and S. Kollias, "A novel rule based machine translation scheme from greek to greek sign language: Production of different types of large corpora and language models evaluation," *Computer Speech & Language*, vol. 51, pp. 110–135, 2018.

[20] M. Nagao, "A framework of a mechanical translation between japanese and english by analogy principle," *Artificial and human intelligence*, pp. 351–354, 1984.

[21] N. Gough, *Example-based machine translation using the marker hypothesis*. PhD thesis, Dublin City University, 2005.

[22] S. Morrissey and A. Way, "An example-based approach to translating sign language," in *Workshop on example-based machine translation*, pp. 109–116, 2005.

[23] A. Almohimeed, M. Wald, and R. I. Damper, "Arabic text to arabic sign language translation system for the deaf and hearing-impaired community," in *Proceedings of the second workshop on speech and language processing for assistive technologies*, pp. 101–109, 2011.

[24] C. M. Bishop, "Mixture density networks," 1994.

[25] M. Selcuk-Simsek and I. Cicekli, "Bidirectional machine translation between turkish and turkish sign language: a data-driven approach," *Int. J. Nat. Lang. Comput*, vol. 6, no. 3, pp. 33–46, 2017.

[26] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pp. 177–180, Association for Computational Linguistics, 2007.

[27] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133, 2003.

[28] J. Bungeroth and H. Ney, "Statistical sign language translation," in *sign-lang@ LREC 2004*, pp. 105–108, European Language Resources Association (ELRA), 2004.

[29] D. Stein, C. Schmidt, and H. Ney, "Analysis, preparation, and optimization of statistical sign language machine translation," *Machine Translation*, vol. 26, pp. 325–357, 2012.

[30] D. Stein, J. Bungeroth, and H. Ney, "Morpho-syntax based statistical methods for automatic sign language translation," in *Proceedings of the 11th Annual conference of the European Association for Machine Translation*, 2006.

[31] V. Lopez Ludeña, R. San Segundo Hernández, R. d. Córdoba Herralde, J. Ferreiros López, J. M. Montero Martínez, and J. M. Pardo Muñoz, "Factored translation models for improving a speech into sign language translation system," 2011.

[32] V. López-Ludeña, C. González-Morcillo, J. C. López, R. Barra-Chicote, R. Córdoba, and R. San-Segundo, "Translating bus information into sign language for deaf people," *Engineering Applications of Artificial Intelligence*, vol. 32, pp. 258–269, 2014.

[33] C. Hogan and R. E. Frederking, "An evaluation of the multi-engine mt architecture," in *Conference of the Association for Machine Translation in the Americas*, pp. 113–123, Springer, 1998.

[34] C.-H. Wu, H.-Y. Su, Y.-H. Chiu, and C.-H. Lin, "Transfer-based statistical translation of taiwanese sign language using pcfg," *ACM transactions on Asian language information processing (TALIP)*, vol. 6, no. 1, pp. 1–es, 2007.

[35] S. Morrissey, "Assistive translation technology for deaf people: translating into and animating irish sign language," 2008.

[36] S. Morrissey and A. Way, "Joining hands: Developing a sign language machine translation system with and for the deaf community," 2007.

[37] R. San-Segundo, J. M. Montero, R. Cordoba, V. Sama, F. Fernández, L. D'haro, V. López-Ludeña, D. Sánchez, and A. García, "Design, development and field evaluation of a spanish into sign language translation system," *Pattern Analysis and Applications*, vol. 15, pp. 203–224, 2012.

[38] V. López-Ludeña, R. San-Segundo, C. G. Morcillo, J. C. López, and J. M. P. Muñoz, "Increasing adaptability of a speech into sign language translation system," *Expert Systems with Applications*, vol. 40, no. 4, pp. 1312–1322, 2013.

[39] V. Lopez Ludeña, R. San Segundo Hernández, C. González Morcillo, J. C. López López, and J. M. Pardo Muñoz, "Methodology for developing a speech into sign language translation system in a new semantic domain," 2012.

[40] D. Manzano, "English to asl translator for speech2signs," *Retrieved March*, vol. 16, p. 2021, 2018.

[41] M. Brour and A. Benabbou, "Atlaslang mts 1: Arabic text language into arabic sign language machine translation system," *Procedia computer science*, vol. 148, pp. 236–245, 2019.

[42] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2sign: towards sign language production using neural machine translation and generative adversarial networks," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 891–908, 2020.

[43] B. Saunders, N. C. Camgoz, and R. Bowden, "Progressive transformers for end-to-end sign language production," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 687–705, Springer, 2020.

[44] B. Saunders, N. C. Camgoz, and R. Bowden, "Adversarial training for multi-channel sign language production," *arXiv preprint arXiv:2008.12405*, 2020.

[45] B. Saunders, N. C. Camgoz, and R. Bowden, "Everybody sign now: Translating spoken language to photo realistic sign language video," *arXiv preprint arXiv:2011.09846*, 2020.

[46] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7784–7793, 2018.

[47] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1316–1325, 2021.

[48] A. Othman and M. Jemni, "English-asl gloss parallel corpus 2012: Aslg-pc12," in *sign-lang@ LREC 2012*, pp. 151–154, European Language Resources Association (ELRA), 2012.

[49] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto, "How2sign: a large-scale multimodal dataset for continuous american sign language," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2735–2744, 2021.

[50] A. Othman and M. Jemni, "English-asl gloss parallel corpus 2012: Aslg-pc12," in *sign-lang@ LREC 2012*, pp. 151–154, European Language Resources Association (ELRA), 2012.

[51] S. Morrissey and A. Way, "Lost in translation: the problems of using mainstream mt evaluation metrics for sign language translation," 2006.

[52] A. Kulesza and S. Shieber, "A learning approach to improving sentence-level mt evaluation," in *Proceedings of the 10th international conference on theoretical and methodological issues in machine translation*, European Association for Machine Translation, 2004.

[53] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231, 2006.

[54] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[55] H.-Y. Jung, J.-H. Lee, E. Min, and S.-H. Na, "Word reordering for translation into korean sign language using syntactically-guided classification," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 2, pp. 1–20, 2019.

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-06

# コミュニケーションから心へ：動物行動からの知見

岡ノ谷 一夫
Kazuo Okanoya

帝京大学
Teikyo University

## 概要

コミュニケーションとは，送信者が発信する信号により受信者が行動を変化させ，結果として送信者が利益を得るような動物個体間の相互作用である．動物の多様なコミュニケーション行動（社会的促進，情動伝染，援助行動，他者操作，交唱，メタ認知等）についてこの視点で分析し，心を仮定する必要があるのか，心を仮定することで何がわかるのかを議論する．同様な議論の方法で，人工知能に心を仮定する必要があるかどうかを考える．

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-07

# 生成モデルによる形質表現を活用した鳥類の鳴き声に関する進化モデルとフィールド実験の融合の試み

# An approach to integrating evolutionary models and field experiments on avian vocalization using trait representations based on generative models

鈴木麗璽[1*]　古山諒[1]　Zachary Harlow[2]　中臺一博[3]　有田隆也[1]
Reiji Suzuki[1]　Ryo Furuyama[1]　Zachary Harlow[2]　Kazuhiro Nakadai[3]　Takaya Arita[1]

[1] 名古屋大学，Nagoya University
[2] University of California, Berkeley
[3] 東京工業大学，Tokyo Institute of Technology

**Abstract:** 本研究は鳥類の鳴き声を題材とし，生成モデルによる形質表現を活用した進化モデルとフィールド実験の融合を目的とした予備的試みについて，オオルリ，ホシワキアカトウヒチョウを対象とする最近の事例を報告する．オス・メスの遺伝子を VAE の潜在空間上のベクトル，それから生成される鳥類音声のスペクトログラム画像を形質・選好性とみなした性選択モデルを構築した．実験の結果，生成鳴き声の中でも明瞭であり単純すぎないものが選択されがちであることが示唆された．さらに，生成モデルに基づく再生音が野生個体にどのような影響を及ぼすかについて，プレイバック実験による予備的検討を行った．野生のオオルリに対する実験では，進化実験でよく選択された鳴き声について対象個体の鳴き返しを抑制する傾向がありうることが示唆された．ホシワキアカトウヒチョウに対する実験では，人の耳にはノイズのように聞こえる生成音でも，なんらかの生態的な影響がありうることが示唆された．

## 1　はじめに

生物や社会集団の進化や相互作用を理解するための計算論的アプローチの一つに，エージェントベースモデルがある．これは，生物・社会集団における主体間の相互作用や進化のルールを記述し計算機上で動かすことで，集団全体に生じる複雑な構造や進化のダイナミクスを理解するものである．その中でも，生命・社会現象の定性的な理解を志向するエージェントベースモデルの多くは，単純な構成要素やルールを設定することが一般的であり，そこから生じる複雑な過程は抽象レベルの理解や知見をもたらし得る．しかし，同時に，これらのモデルは実世界や社会の独特な複雑さと比べ大きな差異があり，直接的に比較するのが難しい場合もある．

一方，近年目覚ましい発展を遂げている ChatGPT や Stable Diffusion といった生成モデルは，インターネット上などの大量データを利用した表現学習を通じて，言語や画像に含まれる構造を潜在空間として抽象化する．これにより，空間上の任意の特徴ベクトルから，元のデータには存在しない言わば架空のデータを生成することが可能となる．この技術によって，現実に即しつつも新奇で複雑な表現を計算機上で生み出せるようになった．これらの生成物のリアルさ，複雑性，新奇性は人々を驚かせ，特に大規模言語モデルの急速な普及は人間社会に大きな影響を与えている．

エージェントベース進化モデルに対して生成モデルがもたらすメリットは少なくとも 2 つあると考えられる [1]（図 1）．一つは，従来よりも現実的，新奇で複雑な形質表現をモデルに盛り込むことができる点である．本研究で注目する鳥類の鳴き声に関連する生物・生態音響学において，多数の動物音声のスペクトログラムの表現学習や次元圧縮手法の利活用が盛んに検討されている．Sainburg らは鳥類をはじめとする様々な動物の鳴き声の録音に対し，深層学習ネットワークによる表現学習や種々の次元圧縮法を用いることで，その潜在空間上の構造を分類や関係の分析に利用するための Python パッケージを公開している [2]．また，Thomas らは近年汎用的な次元圧縮手法として多方面で利用される UMAP[3] に基づく動物音声の潜在空間作成につ

*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:reiji@nagoya-u.jp

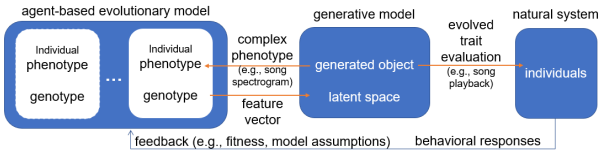



図 1: 生成モデル基づく形質表現によるエージェントベース進化モデルの拡張

いて詳細に検討している [4]．Best らは，変分オートエンコーダ（VAE）に基づく鳥類や海棲哺乳類の鳴き声のクラスタリングの精度について詳細に分析している [5]．これらは既存の収録音の自動識別や特徴の分析のために低次元空間での分布を利用する試みである．

同時に，生成モデルはその低次元空間の任意の位置から当初の高次元データを生成することが可能である．Sainburg らは，VAE などの表現学習に基づく潜在空間表現で複雑な非線形な音声特徴を操作しつつ生成できたり，特徴を線形的に表現できることを指摘している [6]．近年でも，マッコウクジラの音声を表現学習し，学習範囲外の生成音の特徴を分析してシグナルに潜在する役割を調べる手法も提案されている [7]．前述の UMAP も最近潜在空間からの高次元データの生成も可能であり，生成モデルとしての役割も期待される．このような潜在空間をエージェントベース進化モデルの遺伝子空間とし，各遺伝子から生成される音声を形質とすることで，実生態の形質と同等な複雑さを持ったリアルな音声の進化モデルを構築することが可能である．

もう一つのメリットは，進化・創発したリアルな形質を直接フィールド実験で利用したり，その影響を観測したりできることである．しかし，このような生成音の影響を実生態においてプレイバック実験等で調べる試みは知る限りまだわずかである．さらには，両メリットを組み合わせることで，進化モデルと野外生物との直接の相互作用を検討することが考えられる．このような取り組みは音声を介した動物とコンピュータやエージェント間の相互作用 [8] に関連するといえるが，生成モデルを介したエージェントベース進化モデルと実フィールドとの接続の観点に基づく試みは知る限りない．一方，我々は，ロボット聴覚技術を用いた鳥類生態観測システム HARKBird[9] を構築し，プレイバック実験に基づく野外鳥類の行動観測を試行する中で，HARK で分離した音源のクラスタリング等に表現学習や次元圧縮手法を以前から活用しており [10, 11, 12]，また，生成音のプレイバックによる影響についても一部調査を行ってきた [13]．

以上を踏まえ，本研究は鳥類の鳴き声を題材とし，生成モデルによる形質表現を活用した進化モデルとフィールド実験の融合を目的とした最近の予備的試みについて 2 つの事例を報告する．具体的には，起点となる進化モデルとして，Higashi らのオスの形質とメスのえり好みに関する数理モデル [14] に着想を得た性選択モデルを構築する．Higashi らのモデルでは，えり好みとその対象となるオスの形質がそれぞれ正負の整数値で表され，メスは自身のえり好みと評価対象のオスの形質を掛け合わせた値が大きいほどそのオスを選びやすい．ある好みを持つメスが好みのオスを選ぶことで個体内の遺伝子間の相関が高まることが繰り返されて生じるランナウェイ過程が正負の方向に同時に働き，同所的種分化が生じることが報告されている．本研究では，オス・メスの遺伝子を生成モデルの潜在空間上のベクトル，それから生成される鳥類音声のスペクトログラム画像を形質・選好性とみなし，メスは自身の選好性を表すスペクトログラム画像により近い形質を表すスペクトログラム画像を持つオス個体を繁殖相手としてより頻繁に選択する．本モデルを用いて，リアルで複雑な音声が発現しえり好みの対象となる性選択モデルではどのような進化ダイナミクスが創発し，どのような歌が選択されがちかを検討する．これをオオルリ，ホシワキアカトウヒチョウの二種に関する生成モデルを用いて行い，進化の一般的傾向について論ずる．

次に，生成モデルに基づく再生音が野生個体にどのような影響を及ぼすかについていくつかの実験の事例を紹介する．まず，オオルリの鳴き声に関する性選択モデルにおいて頻繁に選択された歌のいくつかを，同種オスの野生個体にプレイバックすることで，野外の鳥類に対してどのような影響を与えうるかを検討する．次に，ホシワキアカトウヒチョウを対象にして，原点周辺に存在する潜在空間上の実録音の写像位置から順に遠ざかるように音声を生成し，明瞭なものから構造が崩れるにつれ，どれほどシグナルとしての役割を保持しうるかを検討する．

## 2 生成モデルによる鳴き声の潜在空間の作成

本研究ではオオルリ（*Cyanoptila cyanomelana*）とホシワキアカトウヒチョウ（*Pipilo maculatus*）に関する 2 種の生成モデルを変分オートエンコーダ（VAE）を利用して作成した．オオルリについては，2018 年 5 月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おいて，16 チャネルマイクアレイ（DACHO; System in Frontier 社製）で録音された野外録音データを採用した．マイクは針葉樹が生い茂り近くに小川が流れる林道中に設置されており，数個体のオオルリが近隣になわばりをもっていたと考えられる．録音からオオルリのさえずりを Praat を用いて歌いだしから 3 秒間を切り出し，496 × 128 ピクセルの音声スペクトログラム画像に変換した．全 30

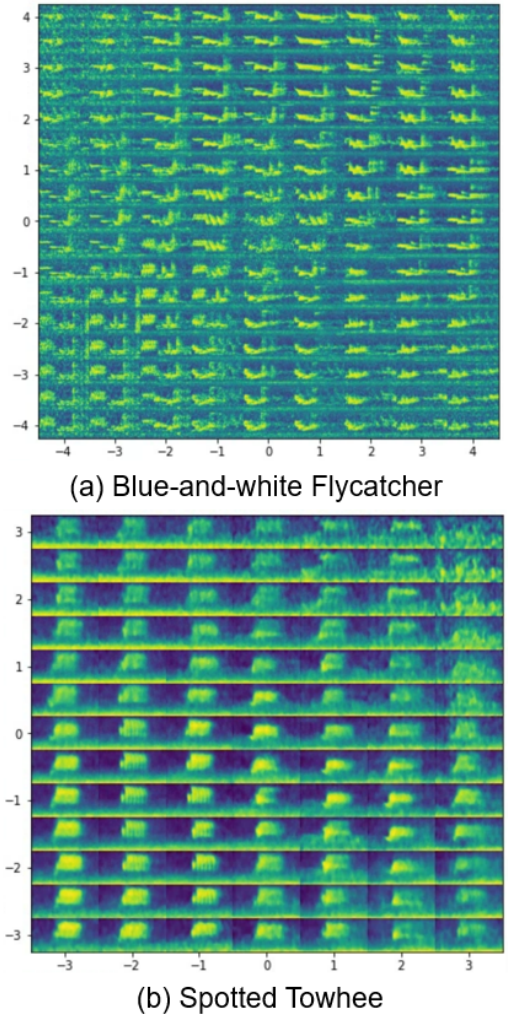



図 2: VAE で作成した 2 次元潜在空間上の鳴き声分布

種類からなる計 304 個のスペクトログラム画像をデータセットとして畳み込み変分オートエンコーダ VAE（8 畳み込み層と 3 完全接続層を持ち，2 次元にまで圧縮するエンコーダと，エンコーダと対称なデコーダ）の表現学習を行った．ネットワークの構成は Sainburg らの VAE[2] を参考にした．

図 2(a) は，生成モデルの 2 次元潜在空間において，各座標位置をデコーダに入力して生成されたスペクトログラムを規則的に並べたものである．同図から，座標の原点付近を中心に異なる種類の歌が集まって分布しており，外側に広がるに従いノイズが大きくなっていることがわかる．明瞭に生成された鳴き声には，オオルリの歌の特徴であるいくつかの音素に続く高周波の「ジジッ」まで再現されているものもある．実録音に対応する生成音を比較すると，生成音は対応する音声をおおむね再現しているが，細かな周波数や強弱の振動等はぼける傾向があった．

図 2(b) は，ホシワキアカトウヒチョウの鳴き声分布である．2023 年 5 月 14 日，米国カリフォルニア州の UC Berkeley の自然保護区である Blue Oak Ranch Reserve の林道沿いに生息する，複数のホシワキアカトウヒチョウのなわばり周辺において，8 チャネルマイクアレイ（TAMAGO-03; System in Frontier 社製）で録音を行った．野外鳥類音源定位分離ソフトウェア HARKBird[9] を用いて音源定位・分離し，同種の歌（2 秒）を約 1000 個取り出した．これらを上記と同様の方法で VAE で学習した．同種は個体ごとに短いさえずりをいくつか持っており，同図からそれらが少しずつ音響特徴を変えながら中心付近に分布していることがわかる．また，中心付近から離れるに従い，音響構造が明瞭でなくノイズが多くのった状況であることもわかる．

以上から，HARK の音源分離による鳴き声スペクトログラムを潜在空間上にマップすることができることが分かった．一方，各音源の種類ごとに間隔をあけてクラスタが形成されるまでには至らず，これはサンプルサイズが影響している可能性がある．

## 3　生成スペクトログラムを用いた鳴き声と選好性の進化

### 3.1　モデル

前述の Higashi らの性選択による同所的種分化の数理モデルをもとに，前節で作成した 2 種の鳴き声生成モデルそれぞれについて，2 次元の潜在空間上の座標を遺伝子とし，その座標から生成されるスペクトログラム画像をオスの鳴き声，メスの選好性とした性選択モデルを構築した（図 3）．各 $N$ 個体からなるオス集団とメス集団を考える．各個体は潜在空間上の位置（$(x, y)$ 座標のペア）を 2 つ実数値の遺伝子型として持ち，それぞれ自身がオスの際に発現する歌，メスの際に発現する歌に対する選好性に関する遺伝子とする．オス個体は，自身の歌の遺伝子型を入力したデコーダネットワークから生成されるスペクトログラム画像が表す歌をさえずるとする．また，各メス個体は，オスがさえずる歌に対する好みとして，好みの遺伝子型から生成されるスペクトログラムを生得的な選好性の鋳型として配偶者選択に用いる．

各メスは 1 個体のオスを選択して繁殖する．各メスはすべてのオスを $exp^{-\beta \times x}$ で評価する．ここで，$x$ は注目するオスの歌スペクトログラムと自身の選好性スペクトログラムの各画素値の差の平均，$\beta$ は係数である．すべてのオス個体から，評価値に比例した確率で 1 個体のオスを選択する．同式は，歌のスペクトログラムと選好性が近いオスほど選ばれやすいことを示して

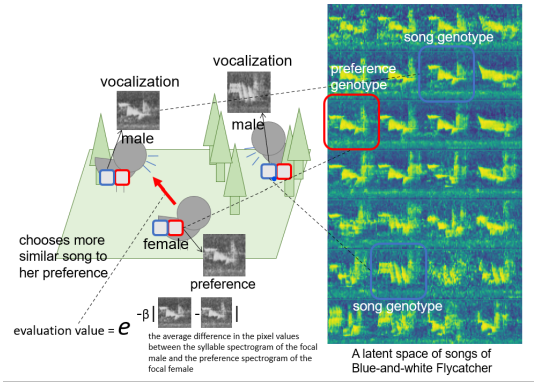



図 3: 生成スペクトログラムを用いた形質表現に基づく性選択モデル

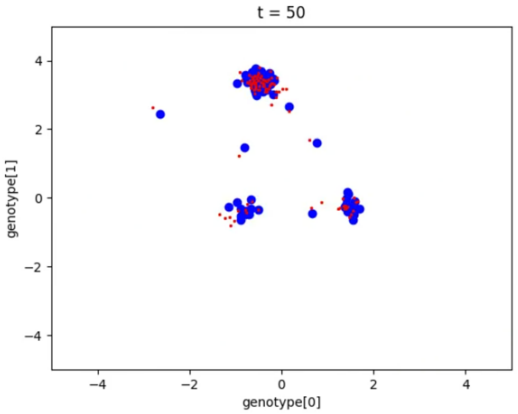



図 4: 進化実験におけるオスの遺伝子分布の例．青：鳴き声遺伝子，赤：選好性遺伝子．50 世代目のスナップショット（最終世代では上と右のクラスタのみ生存）．

いる．自身の遺伝子と選択した個体の遺伝子で 2 体の子個体のペアがつくられる．この際，確率 $p_c$ で BLX-$\alpha$ 交叉 [20] が生じ，加えて確率 $p_m$ で各遺伝子に突然変異（平均 0，標準偏差 $\sigma$ の正規乱数の値を加算）が生じる．ペアのうち一方をオス，もう一方をメスとランダムに決定する．1 試行は初期集団の各遺伝子の値を $[-W, W]$ の一様乱数から決定するものとして，$T$ 世代にわたって行った．

## 3.2 進化実験

本研究では，$N$=100，$W$=5.0，$\beta$=0.3，$\alpha$=1.1，$T$=100，$p_c$=0.5，$p_m$=0.15，$\sigma$=0.2 を採用した．両種の生成モデルを用いた実験に共通することとして次のことが分かった．数試行の分析から，図 4 に示されるように，各試行では初期集団から鳴き声遺伝子と選好性遺伝子が相関し合いつついくつかの集団に分化する傾向が見られた．これは，鳴き声と選好性のスペクトログラムの代わりに遺伝子の座標そのものを採用し，選好性は座標間の距離の逆数で計算した場合において，原点付近に急速に収束する結果と異なった．つまり，スペクトログラムの複雑な特徴が，Higashi らが示したような複数同時に生じるランナウェイ過程（オスのある特徴を好む遺伝的特性を持つメスが何らかの要因（遺伝的浮動やその他の淘汰圧，外界の影響など）で増加すると，メスはその特徴をより持ったオスを配偶者として選んで子個体をつくるため，次世代でその鳴き声遺伝子を持つ個体は同時にそれを好む遺伝子を持つ傾向が強まる．これは，メスがその特徴を持ったオスを選ぶことが自身の好み遺伝子も間接的に選ぶことにつながる．これが繰り返されてよりその特徴を持った鳴き声と好みが同時に集団に広まっていく過程）のような形で同所的なすみわけをもたらしたと考えられる．

一方，分化した結果収束する遺伝子は試行毎に大きく異なった．そこで，多数実験を行い全体的な傾向を抽出することにした．図 5(i) は，2 種それぞれ (a, b) に関して，全 (a)500 試行または (b)2000 試行における最終世代のオスの鳴き声遺伝子の頻度分布を KDE（kernel density estimation）分布で示したものであり，色が青・緑・黄・赤となるに従い頻度が高いことを示している．(2) は比較のための図 2 の再掲，(3) は潜在空間上のスペクトログラムに対して音響複雑性指標（Acoustic Complexity Index, ACI）[15] を計測したものであり，色が青・緑・黄・赤となるに従い ACI が高いことを示している．

まず，オオルリに関して，鳴き声遺伝子分布 (i) から，選択された鳴き声は全体として原点の周囲の広い範囲に分布することがわかる．これは，スペクトログラム分布 (ii) と比較すると，ノイズが少なく鳴き声が比較的明瞭に生成されている範囲に分布しがちであることがわかる．例えば，最も頻繁に選択された鳴き声遺伝子分布 (i) 中の 1 の位置の鳴き声は，典型的なオオルリのさえずりが明瞭に再生されるものであった．ACI 分布 (iii) と比較すると，ACI が低めの黄緑からやや明るい青色付近に分布していることがわかる．ACI は環境ノイズを取り除いた録音内に存在する生物音声を定量的に抽出する指標であり，各周波数ビンごとのパワーの時間変化が大きいほど大きい．同図では，ノイズが大きいスペクトログラムに対してそのまま適用しているため，ノイズの大きい領域では赤く極端に高い値になっているが，進化した鳴き声遺伝子はそこを避けて分布しており，ノイズが少なく明瞭な鳴き声が選択されがちであることがわかる．また，最も値の低い領域，例えばオオルリでは図中上付近も避けられているように見え，同時に単純すぎても選択されにくいこともわ

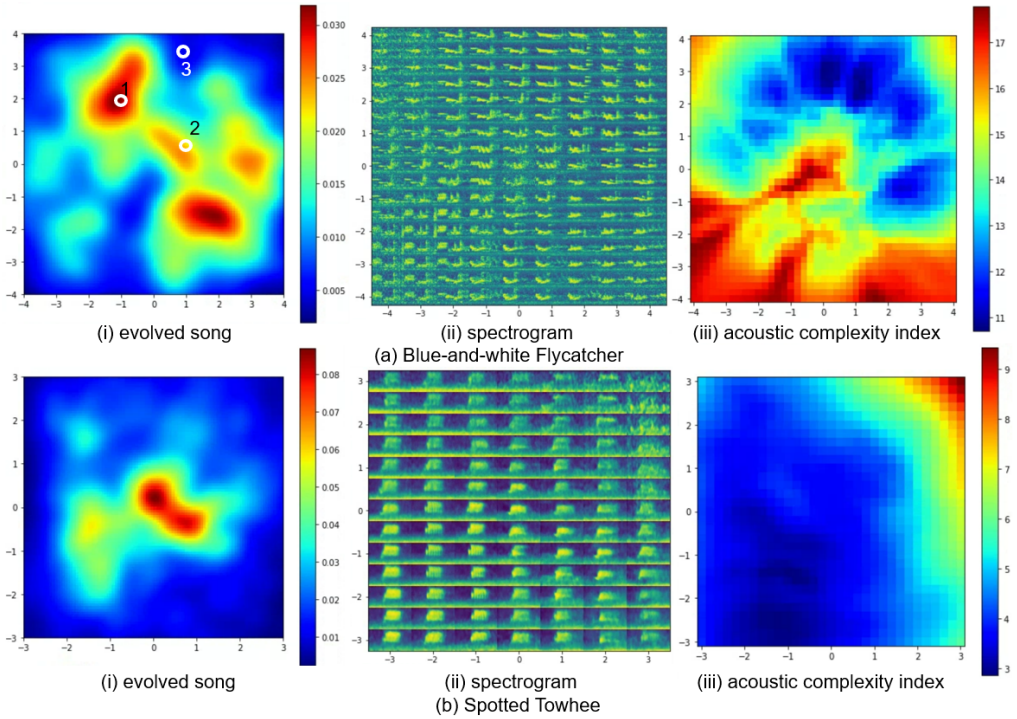



図 5: 進化実験結果．(a) オオルリ，(b) ホシワキアカトウヒチョウ．(i) オスの鳴き声遺伝子の分布，(ii) スペクトログラムの分布，(iii) 音響複雑性（ACI）指標の分布．

かる．ホシワキアカトウヒチョウについても，中心付近だがやや上方と右下にずれた位置のピークと，左下に中頻度で選択されたピークがあり，明確ではないものの上記の傾向を反映しているといえる．これらから，本進化モデルでは明瞭であり単純すぎない鳴き声が選択されがちであることが示唆された．

## 4　生成音を用いたプレイバック実験

前節での進化モデルで選択された歌や生成音が実際の生態でどのような特徴を持つか検討するため，オオルリ，ホシワキアカトウヒチョウそれぞれについて生成音を用いたプレイバック実験を試行した．状況設定が大きく異なるため，両種についてそれぞれ説明する．

### 4.1　オオルリに対する実験

進化モデルで選択された歌が実際の個体に与える影響の予備的検討のため，2023 年 6 月 27 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，オオルリのオス 1 羽に対して前節の生成音を用いたプレイバック実験を行った．実験場所は，周囲を針葉樹で囲まれた林道で，Raspberry Pi に接続され長時間録音が可能な USB マイクロホンアレイ（TAMAGO-03, システムインフロンティア社製）を複数台設置し，各マイクで同時に録音した．現地を縄張りとしているとみられるオス個体に対して，図 5(i) 中の 1〜3 の位置に対応する生成音と，同フィールドの別の場所で録音されたオスの鳴き声，“CD 鳴き声ガイド日本の野鳥”[16] に収録されたメスの鳴き声を加えた計 5 種である．実際には，より多くの種類の鳴き声も用いて 26 日にも実験を行ったが，エゾハルゼミのや多種の鳴き声が大きく鳥類の行動に影響する可能性があったり，音源定位が難しくなるため，27 日の状況の良い録音のみを分析に用いた．それぞれについて 6.5 秒の間隔を置いて 10 回再生するファイルを作成し，3 回繰り返し再生した．

実験間に 30 分以上の間隔を挟み，実験場所付近にオスの個体が接近した際に実験を開始し，プレイバック前，プレイバック中，プレイバック後の約 5 分ずつの鳴き声を調査対象とした．HARKBird を用いて 2 つのマイクで到来方向に関する音源定位を行い，同時刻に両マイクから定位されたオオルリの鳴き声の方向を用いて 3 点測量の方法で音源の位置を計算した．HARKBird の設定はオオルリの鳴き声をできるだけ抽出するように設定したが，音源との距離やその他の音の影響で抽出できない場合があり，すべての音源の位置情報を取り出すことはできなかった．また，基本的には同じ場所でさえずりを繰り返すため，60m を超える移動は外

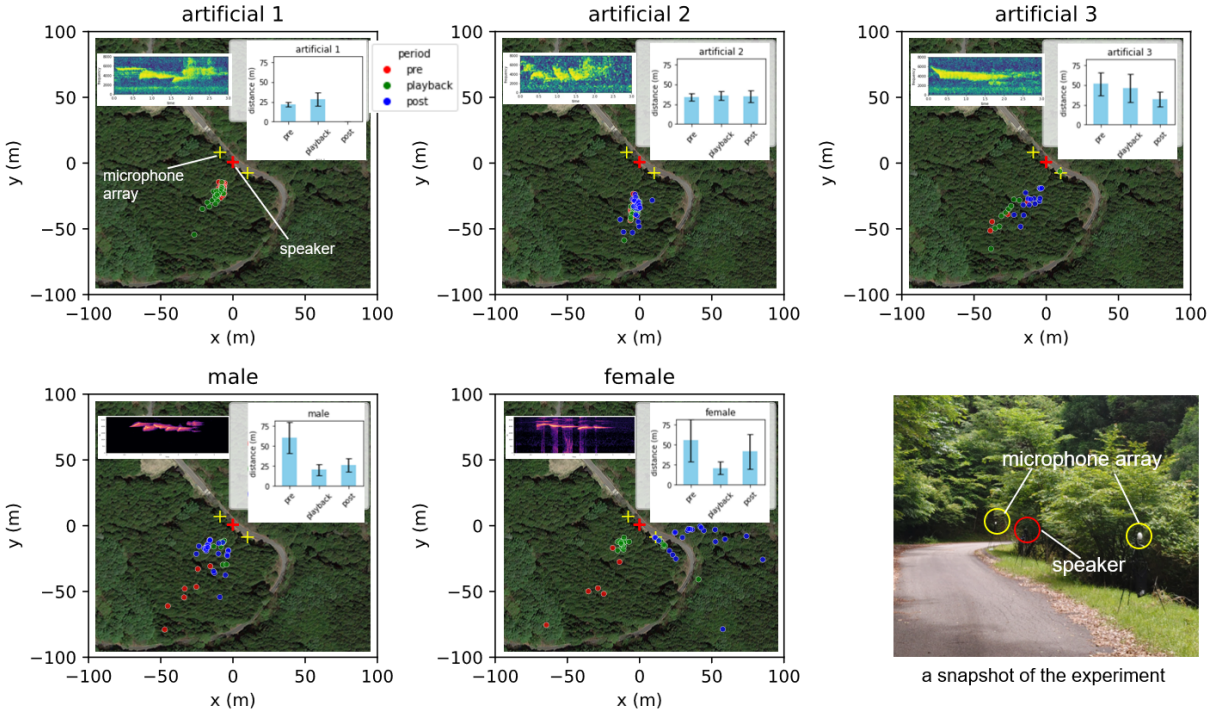



図 6: オオルリに対するプレイバック実験結果.

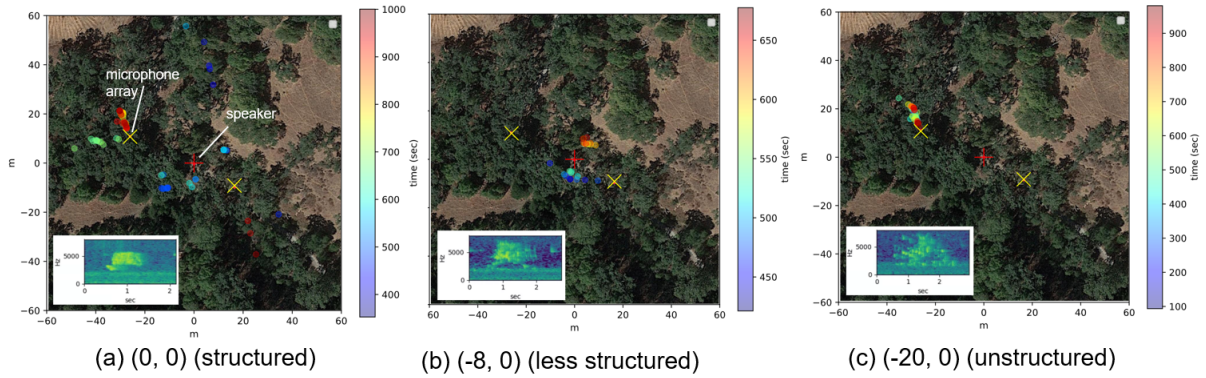
れ値として除外した．以上のように予備的な分析であるが，大まかな分布を知るには有益であると考えた．

図 6 は各再生音ごとの，プレイバック前・中・後で定位された音源の二次元分布と，スピーカからの距離（平均，標準偏差）を示したものである．対象の個体は，林道沿いに設置したスピーカ・マイクから見て南側（図中下側）の林道に囲まれた森の針葉樹の上でさえずっていたことがわかる．図中下段のオス・メスの鳴き声の場合，やや遠い位置でのさえずりからプレイバック中に 20m を下回る極めて近い位置にまで接近してさえずっており，プレイバック音に対する積極的な鳴き返しであると推測される．なお，オオルリはメスもさえずる場合がある北半球において稀な種として知られており，かつて実施した実験においてもメスの鳴き声のプレイバック音に対して反応があったことを報告している [17]．森の中でさえずる様子はわかっても，その中の詳細な位置関係を現地で観測するのは容易ではないため，このようなデータは繊細な生成音の影響の分析に役立つことが示唆される．

次に，上段の 3 つの生成音を用いた実験に注目する．これらの生成音 1〜3 は，それぞれ，進化実験において，高頻度・中頻度・低頻度（もしくはほぼ選択されず）で選択された鳴き声であり，これらがもたらす影響を検討する目的で実施した．まず，生成音 2，3 の中頻度・低頻度で選択された鳴き声の場合，実験期間を通して明確なさえずり位置やスピーカからの距離の変化は見られなかったが，同様な位置でのさえずりが維持された．一方，高頻度で選択された生成音 1 の場合は，当初から近い位置でさえずっていたのが，プレイバック中にやや離れたのちにその場から去る結果になった．これが，進化実験の結果よく選択された鳴き声が明瞭で特徴的であり強力な生成音であるため鳴き返しをあきらめたものか，単にプレイバック音に対する興味を失ったものか判断はつかないが，同図の実験以外の設定において個体が去ってしまう状況はこれ以外に観測されておらず，特徴的な行動であったことが考えられる．また，実験を通して，対象個体の当初のさえずり頻度が低い場合はプレイバック期間に頻度が増加する一方，当初の頻度が高い場合は逆にプレイバック期間に頻度が下がる傾向もやや見られ，影響の可能性が示唆されるが，検討が必要である．また，実験時期が繁殖期の後期であり，より活発な時期における調査では異なる知見が得られる可能性がある．

## 4.2 ホシワキアカトウヒチョウに対する実験

VAE で生成した 2 次元潜在空間上では，原点の周囲に鳴き声の特徴がよく表現された音が生成され，そこからさらに広がるにつれて構造が崩れノイズの多い音声が生成された．この傾向において，原点から遠ざかるに従ってもなおどれほど実際の個体に対してシグナ




(a) (0, 0) (structured)　(b) (-8, 0) (less structured)　(c) (-20, 0) (unstructured)

図 7: ホシワキアカトウヒチョウに対するプレイバック実験結果.

ルとしての影響がありうるかを検討するため，ホシワキアカトウヒチョウの鳴き声潜在空間において原点からの距離の異なる 3 つの生成音を用いてプレイバック実験を実施した．

2023 年 5 月 17 日，米国カリフォルニア州 UC Berkeley の自然保護区である Blue Oak Ranch Reserve の林道において実験を実施した．同フィールドの林道沿いには複数のホシワキアカトウヒチョウの縄張りが点々と存在しており，複数個体の反応が得られる可能性の高い場所にスピーカ，および，2 台のマイクアレイ（TAMAGO-03, システムインフロンティア社製）を 40m ほど離して設置し，各実験 30 分以上の間隔をあけて実験を行った．録音は午前中に行われ，20 分毎のファイルに保存された．プレイバック期間を含む各 20 分の録音におけるプレイバック対象個体の歌（地鳴きは含まず）を HARKBird で音源定位し，三点測量の方法で 2 次元定位した．この時，到来方向は各音源が定位された時間を通した値の平均値を用いた．なお，いずれの場合もプレイバックをきっかけに単一，複数の個体がスピーカ付近に飛来したことを目視で確認している．

図 7(a) は，3 種のうち原点 (0, 0) の位置から生成された音声を再生した場合の結果である．この場合，図左下にあるようにホシワキアカトウヒチョウの典型的な鳴き声が明瞭に生成されている．同図から，時間変化に伴ってスピーカ周辺で複数の鳴き声が定位されていることがわかる．また，類似の時刻を表す同系色の音源が複数個所に分布しており，2，3 個体が飛来し鳴き返していることを示している．これは，生成音が周辺の複数個体による強い反応を引き起こし，相互に影響し合ったことを示していると考えられる．

図 7(b) は，原点から離れて音響構造が崩れた生成音 (-8, 0) を再生した場合であり，1 個体が飛来し，スピーカ周辺にかなり接近し林道をまたぎつつ鳴き返す様子が観測された．明確に積極的な威嚇の行動であり，構造が明確でなく人工的でノイズを含んだ音声であっても反応を引き起こしたのは興味深いといえる．

図 7(c) は，さらに原点から離れた (-20, 0) の位置からの生成音を再生した場合である．この時，西側（図中左側）になわばりをもつ個体がその境界と思われる林道の端まで来て，スピーカからは遠い位置を維持しつつ鳴き返し，プレイバック終了後まもなく去っていった．実際の鳴き声ほどではないにせよ，何らかの興味を惹いたと考えられる．金属的なノイズのように聞こえる再生音でもなお，個体の行動に影響したことは興味深いと同時に，生成音には明示的でない形で生態に影響する特徴が含まれうることも示唆していると考えられる．

総じて，VAE の潜在空間において，よく鳴き声の特徴が反映される原点周辺から遠ざかるに従い，鳴き返し等の反応が弱くなるが，相当離れたほぼノイズのような鳴き声でも反応が得られる傾向が観察された．なお，同フィールドの別の場所で同様の実験を行った際には，(0, 0) を再生した場合にのみ個体の鳴き返しの反応が観測された．その際の近隣は当初の場所よりも静かで，全体的に活発でない様子であったが，部分的に上記の傾向を支持する結果であるといえる．

## 5　おわりに

本研究は鳥類の鳴き声を題材とし，生成モデルによる形質表現を活用した進化モデルとフィールド実験の融合を目的とした最近の予備的試みについて，オオルリ，ホシワキアカトウヒチョウを対象とする事例を報告した．オス・メスの遺伝子を VAE の潜在空間上のベクトル，それから生成される鳥類音声のスペクトログラム画像を形質・選好性とみなした性選択モデルを構築した．実験の結果，生成鳴き声の中でも明瞭であり単純すぎないものが選択されがちであることが示唆された．

次に，生成モデルに基づく再生音が野生個体にどのような影響を及ぼすかについて，プレイバック実験による予備的検討を行ったところ，野生のオオルリに対する実験では，対象個体のさえずりが維持される場合や，進化実験でよく選択された鳴き声が対象個体のさえずりの抑制や忌避をもたらす場合がありうることが示唆された．また，ホシワキアカトウヒチョウに対する実験では，明瞭な生成音は野生個体の積極的な反応を招き，人の耳にはノイズのように聞こえる生成音であっても生態的な影響がありうることが示唆された．

ChatGPT や Stable Diffusion のような生成 AI が社会に浸透し，人を驚かせたり，仕事を効率化したり，楽しませたりするのと同時に，社会が生成されたコンテンツであふれたり知らないうちに人の行動にバイアスがかかったりするなどの課題がある．上記の知見は類似の課題が自然と AI 社会との接点においても生じうることを示唆していると考えられる．一方，これらの影響をうまく利用すれば，生成されたり進化する音を利用した環境保全など，エージェント進化モデル・人工システムと自然・生態との新しい接点やあり方を考えることができるといえる．今後はこのような可能性も考えつつ，ロボット聴覚技術を活用した進化モデルとフィールド実験の融合の試みを進めていきたいと考えている．

## 謝辞

本論文の研究の一部は JSPS 科研費 JP17H06383（#4903），JP19KK0260，JP21K12058，JP20H00475 の助成を受けた．また，カリフォルニアの録音のアノテーションでは Hao Zhao 氏の協力を得た．ここに謝意を表する．

## 参考文献

[1] Reiji Suzuki, Shinji Sumitani, Chihiro Ikeda, and Takaya Arita. A modeling and experimental framework for understanding evolutionary and ecological roles of acoustic behavior using a generative model. In *Proceedings of ALIFE 2022*, isal_a_00542, 2022.

[2] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Latent space visualization, characterization, and generation of diverse vocal communication signals. *bioRxiv*, 870311, 2020.

[3] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv e-prints*, arXiv:1802.03426, 2018.

[4] Mara Thomas, Frants H. Jensen, Baptiste Averly, Vlad Demartsev, Marta B. Manser, Tim Sainburg, Marie A. Roch, Ariana Strandburg‐Peshkin. A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations. *Journal of Animal Ecology*, 91(8), 1567–1581, 2022.

[5] Paul Best, Sébastien Paris, Hervé Glotin and Ricard Marxer. Deep audio embeddings for vocalisation clustering. *PLoS ONE*, 18(7), e0283396, 2023.

[6] Tim Sainburg and Timothy Q. Genter. Toward a computational neuroethology of vocal communication: From bioacoustics to neurophysiology, emerging tools and future directions. *Frontiers in Behavioral Neuroscience*, 15, 811737, 2021.

[7] Gašper Beguš, Andrej Leban and Shane Gero. Approaching an unknown communication system by latent space exploration and causal inference. *arXiv e-prints*, 10.3389/fnbeh.2021.811737, 2023.

[8] Roger K. Moore, Ricard Marxer, and Serge Thill. Vocal interactivity in-and-between humans, animals, and robots. *Frontiers in Robotics and AI*, 3, 61, 2016.

[9] Reiji Suzuki, Shiho Matsubayashi, Richard W. Hedley, Kazuhiro Nakadai, and Hiroshi G. Okuno. HARKBird: Exploring acoustic interactions in bird communities using a microphone array. *Journal of Robotics and Mechatronics*, 27, 213–223, 2017.

[10] Shinji Sumitani, Reiji Suzuki, Shiho Matsubayashi, Takaya Arita, Kazuhiro Nakadai, and Hiroshi 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.

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

[12] 炭谷晋司, 鈴木麗璽, 有田隆也, 和多和宏, 松林志保, 中臺一博, 奥乃博. 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. 第 58 回人工知能学会 AI チャレンジ研究会資料, pp. 12–20, 2021.

[13] 炭谷晋司, 松林志保, 鈴木麗璽, 有田隆也, 中臺一博, 奥乃博. 生成モデルに基づく鳴き声を用いた鳥類に対するプレイバック実験の試行. 第 55 回 AI チャレンジ研究会資料, pp. 6–11, 2019.

[14] Masahiko Higashi, Gaku Takimoto, and Norio Yamamura. Sympatric speciation by sexual selection. *Nature*, 402 (6761), 523–526, 1999.

[15] N. Pieretti, A. Farina, and D. Morri. A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI). *Ecological Indicators*, 11, 868–873, 2011.

[16] 松田道生. CD 鳴 き声ガイド日本の野鳥. 日本野鳥の会, 東京, 2016.

[17] 古山諒, 鈴木麗璽, 炭谷晋司, 有田隆也. 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. 第 58 回人工知能学会 AI チャレンジ研究会資料, pp. 6–11, 2021.

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-08

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

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

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

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

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

## 1　はじめに

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

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

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

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

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

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

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

## 2 関連研究

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

### 2.1 伝達関数のオンライン適用

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

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

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

### 2.2 補間を用いた伝達関数サイズの削減

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

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

# 3 提案手法

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

## 3.1 離散伝達関数モデル

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

## 3.2 フーリエ伝達関数モデル

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

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

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

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

フーリエ係数 $C_n$ は離散フーリエ変換を使って次のように計算できる．

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

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

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

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

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

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

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

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

また，フーリエ係数は以下のように表される．

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

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

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

## 3.3 フーリエ伝達関数モデルのオンライン適応

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

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

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

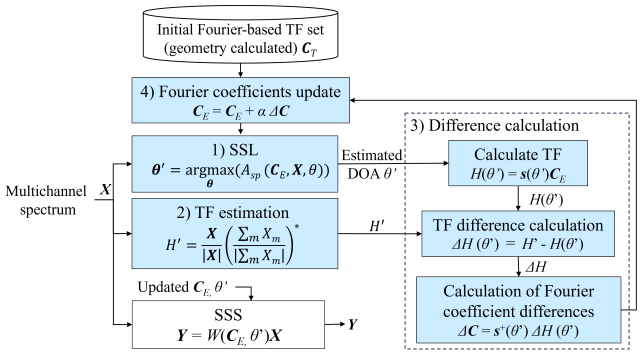



図 1: フーリエ伝達関数モデルにおけるオンライン適応

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

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

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

ここで $m$ と $*$ はマイクインデックスと共役演算子を表す．

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

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

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

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

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

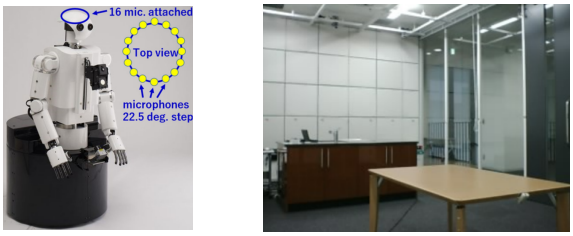
$$\Delta \boldsymbol{C} = \boldsymbol{s}^+(\theta')\Delta H(\theta'), \tag{14}$$

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

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

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

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




(a) HEARBO (15ch)　　(b) 実験室

図 2: 実験環境

# 4 実験

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

## 4.1 実験 1: 適用後の伝達関数比較

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

## 4.2 実験 2: 音源定位評価

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

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

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

### 4.3 実験 3: 音源分離評価

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

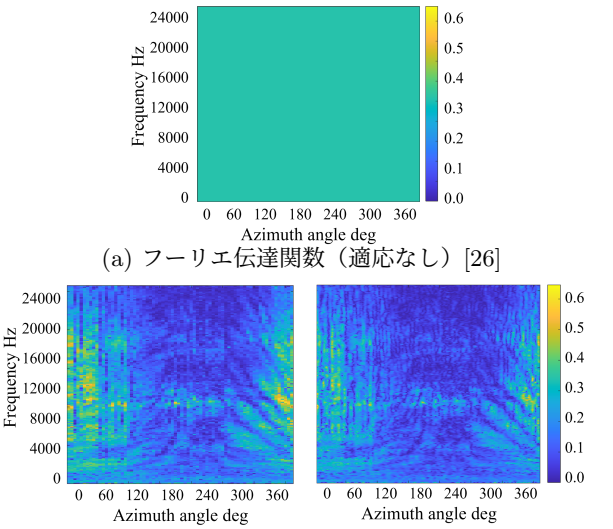
$$SDR(y) = 10\log10(||y_{\mathrm{t}}||^2/||e_{\mathrm{r}}||^2), \qquad (17)$$

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

## 5 実験結果

### 5.1 実験 1: 適応後の伝達関数比較

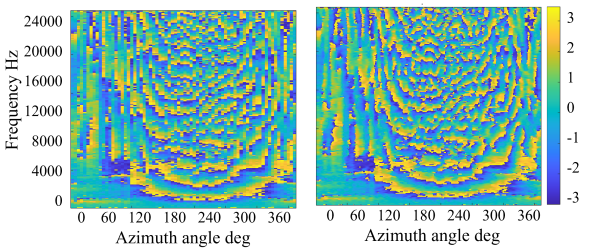
図 3 は，(a) 幾何計算によって算出したフーリエ伝達関数（適応なし）[26]，(b) オンライン適応手法を用いた離散伝達関数 [35]，(c) 提案したオンライン適応手法を用いたフーリエ伝達関数の振幅スペクトルを示す．自由音場を想定した幾何計算により算出された伝達関数は，音源到来方向や周波数成分によらず一定の振幅特性を持つ（図 3a）のに対し，オンライン適応手法を用いることで，実験環境の音響特性を反映するように伝達関数が更新された（図 3b, 3c）．提案手法と従来手法はほぼ同等の振幅特性を示していることから，提案手法が従来手法と同様に実際の環境に適応することができたことがわかる．また，適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



(a) フーリエ伝達関数（適応なし）[26]

(b) 離散伝達関数 [35]　(c) フーリエ伝達関数 (ours)

図 3: 伝達関数の振幅スペクトル



(a) 離散伝達関数 [35]　(b) フーリエ伝達関数 (ours)

図 4: 伝達関数の位相スペクトル

### 5.2 実験 2: 音源定位評価

#### 5.2.1 音源定位性能

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

#### 5.2.2 適応率の影響

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

#### 5.2.3 フーリエ次数の影響

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

表 1: 音源定位の標準誤差

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

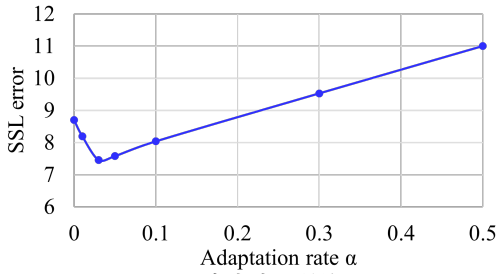


図 5: 適応率の影響

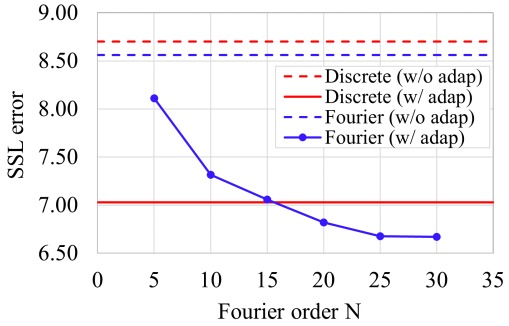


図 6: フーリエ次数の影響

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

### 5.3 実験 3: 音源分離評価

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

表 2: 音源分離タスクにおける SDR 改善 (dB)

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

## 6 議論

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

## 7 結論

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

# 参考文献


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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-09

# オープンソースAIツールの手術動画解析への適用

## Application of open source AI tools to surgical video analysis

山下 一郎[1*]　山田 敏哉[1]　宮崎 淳[1,2]
西田 健次[3]　村上 貴志[4]

[1]（株）オレンジテクラボ
[1] Orange Tech Lab Inc.
[2] 東京国際工科専門職大学
[2] International Professional University of Technology in Tokyo
[3] 東京工業大学 工学院 システム制御系
[3] Dept. of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology
[4] 東京歯科大学市川総合病院 心臓血管外科
[4] Tokyo Dental College Ichikawa General Hospital

**Abstract:**
模擬手術環境における看護師の技量評価を自動化するためのAIツールの適用可能性の検討，および，評価を行った．具体的には，器械出し看護師から医師への手術器材の手渡しの自動検出を試みた．器材の手渡しは，手術中の看護師と医師の協調作業の重要な側面を捉えるものである．本研究では，看護師と医師の手首間距離の分析や発声認識を通じて，手渡しタイミングの適切さを判断する新しい手法を開発した．現行の技術では誤検出や録音の質の問題があったが，これらは将来の改善のための重要な指針を提供する．また，視線検出に代えて顔の向き推定を行うことで，手術室における看護師の行動パターンを把握するための基礎的なデータを得ることができた．本研究の成果は，医療現場におけるAIの活用を進め，手術室での看護師の作業を効率化し，全体的な医療の質を向上させる重要な一歩となる．

## 1　はじめに

器械出し看護師は手術において重要な役割を果たしており，その技量を向上させるために種々の訓練手法が提案されてきた．

Skovらは器械出し看護師に対してシミュレーションに基づく教育を行うことで，ストレスを軽減し手術パフォーマンスが向上することを示しており [1]，Siahらはバーチャルリアリティシミュレーションを用いて，看護学生の有効性，態度，信頼度を評価した [2]．しかし，訓練の結果，看護師の技量が向上したか否かの評価に関しては，人間の判断に頼る部分が多いと言える．原らは初心者看護師に対するシミュレーションによる教育の効果を検証したが，技量の評価に関してはアンケートなどに基づいて行っており，客観的な評価指標を得てはいない [3]．また，Nasiriらは初心者看護師の器械出し技量を評価するためのチェックリストを開発し，その信頼性を評価した [4]．評価項目としての有効性は示しているが，アンケートに基づく評価指標となっており，人間の判断が基盤となっていると考えられる．Bracqらは熟練看護師と初心者看護師のバーチャル手術室におけるエラー認識の差を評価しているが，これは必ずしも看護師の技量を評価したものではなかった [5]．

一方，近年では深層学習などのAI技術を用いて，手術室での動作解析を行う試みがなされてきており，それを活用することにより，看護師の技量評価を行える可能性もある．Funkeらは三次元CNNを用いて時空間特徴を学習することにより，手術における動作認識を行うことを目指した [6]．また，岸らはCNNを用いて時間的姿勢特徴を学習することにより，手術手技認識を行うことを目指し [7]．模擬手術に対する動作認識において，一部の動作に対しては90%程度の認識精度が得られたが，平均的な精度は75%程度であり，動作ごとの認識精度の差が大きいことが示されていた．Goldbraikhらは時間方向の畳み込みネットワーク (temporal Convolutional network）の一種であるMS-TCN++による手術動作認識を目指した [8]．Khalidらは深層学習による手術動作

*ichiro.yamashita@orange-tech-lab.com

の認識とパフォーマンス測定手法を評価した [9]．北口らは深層学習を用いた実時間での手術の段階認識手法を提案し [10]，腹腔鏡下 S 字結腸切除術において，高い認識率を示した．Menegozzo らは時間遅れニューラルネットによる運動データに基づく手術動作認識を提案した [11]．しかし，これらの研究は，主に，術者の動作を認識することを目的としており，看護師の技量を評価するためのものとはなっていない．

また，器材の検出，追跡を行うことで，器材の手渡しのタイミングを検出することができるが，Bajraktari らの CNN を用いた機器検出による手術補助システムでは，ベンチマークセットに対しては良好な識別率を示しているが，実際に用いられる手術器具の種類をカバーしきれてはいない [12]．

看護師，および，術者の厳密な動作認識，器材の追跡などを行えば，器材の手渡しタイミングなどの検出が可能になり，看護師動作の適否の評価が可能になる．しかし，現実には，動作認識のための学習サンプルが不足しており，汎用性のある認識評価システムを構築することは困難であると考えられる．そこで，看護師の技量評価を行うための学習サンプルが不足していることを前提に OpenPose[13]，YoloV5[14]，Whisper[15] などの AI ツールを活用して，簡便に得られる特徴量を元に，看護師の技量評価が可能か否かを検証することとした [16]．

## 2　器械出し看護師の技量評価自動化への特徴量抽出実験

器械出し看護師の技量は，看護師から術者に対して適切なタイミングで適切な器材が渡されているか否かで評価できると考えられる．器材の手渡しの検出，追跡を行えば，手渡しのタイミングを検出することは可能となるが，現状では高い精度での手渡しタイミングを検出することは難しく，また，タイミングの適切さの評価基準は定まってはいない．そこで，現状の AI ツールによって，看護師の技量評価に資する特徴量が抽出できるか否かを確認する実験を行った．模擬手術の動画に対して OpenPose による看護師，および術者の姿勢解析，YoloV5 による器材の検出，Matlab の Speech detection[17] による音声区間切り出し，Whisper による音声認識を適用し，どの程度の認識が可能で，実際に機材の手渡しが行われたタイミング検出への有効性の検証を行った．本節では，模擬手術動画の撮影環境，および，動画に対する AI ツールの適用結果について述べる．

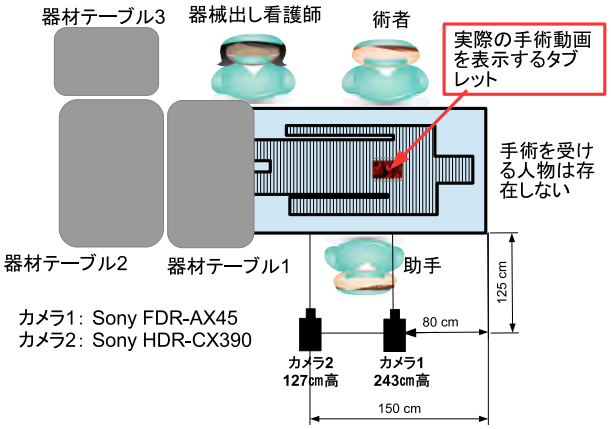



図 1: 実験環境

図 2: 取得データ例

### 2.1　実験環境

図 1 に，模擬手術動画撮影の環境を示す．模擬手術の内容は，心臓手術における開胸から人工心肺の開始，上行大動脈遮断までである．実際の手術で撮影された術野の動画を手術台上に置かれたタブレットで再生しながら，手術手順を追った．カメラ 1 は器材テーブルおよび看護師，医師の手元を上から写すために，高さ 243 cmに設置し，カメラ 2 は看護師の顔を写すために高さ 127 cmに設置した．音声の録音は，ビデオカメラ付属のマイクを利用した．また，看護師，術者，助手の三者とも，今回は熟練者である．取得された動画に OpenPose を適用した例を図 2 に示す．

### 2.2　器材の検出

YoloV5 による器材の検出の状況を図 3 に示す．器材の学習には Kaggle の Labeled Surgical Tools and Images[18] を用いたが，今回使用した機材に対応したデータセットではなかったため，良好な性能を得ることはできなかった．

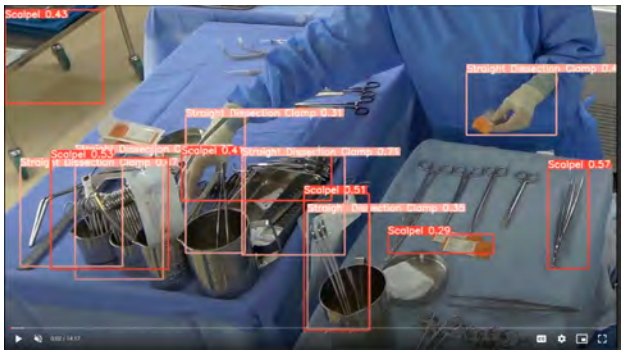



図 3: YoLoV5 による器材の検出

表 1: 手首間の距離による手渡し検出精度

| 看護師 | 術者 | 手渡し回数 | 検出回数 | 精度 |
|---|---|---|---|---|
| 右手 | 右手 | 22 | 35 | 62.9% |
| 右手 | 左手 | 1 | 1 | 100% |
| 左手 | 右手 | 7 | 64 | 11.0% |
| 左手 | 左手 | 1 | 3 | 33.3% |
| トータル | | 31 | 103 | 30.1% |

## 2.3 手渡しタイミングの検出

OpenPose による姿勢検出を元に，器材の手渡しタイミングの検出を試みた．看護師と術者の手首の位置を推定し，その距離が近づいた時を「機材手渡し」の瞬間とした場合の検出精度を検証した．実際に機材が手渡された瞬間を目視で確認し，その時刻をゼロとして，看護師と術者の手首の位置を図 4 に示す．術者が手を伸ばした後に看護師の手が近づいていくことが確認できる．

看護師と術者の手首間の距離が 500 画素よりも小さくなったと検出された回数は 103 であり，そのうち実際に機材が手渡しされていたのは 31 回であった．その内訳は，看護師の右手から術者の右手への手渡しは距離による検出 35 回に対して実際の手渡しは 22 回，看護師の右手から術者の左手は検出 1 回に対して手渡し 1 回，看護師の左手から術者の右手は検出 64 回に対して手渡し 7 回，看護師の左手から術者の左手は検出 3 回に対して手渡し 1 回となっている．看護師の左手と術者の右手の距離による誤検出が多いのは，両者の位置関係により，手渡しを行わない場合でも距離が近づいてしまうためであろうと考えられる (表 1)．

## 2.4 手首間距離最小と手渡しの時間差

手首間距離最小による手渡しの検出精度は高くないものの，手首間距離最小の瞬間と実際に手渡しが行われた瞬間の時間差によって，看護師の技量を推定できる可能性がある．そこで，実際に手渡しが確認された事例において，手首間距離最小と手渡しの時間差を検証した (図 4)．図 5 に手首間距離最小と手渡しの瞬間の時間差のヒストグラムを示した．時間差が負の値をとるのは，手渡しが行われるよりも前に手首間の距離が最小になっていることを示す．本実験では，手首間距離最小のほとんどが手渡しの前後 0.2 秒の間に入っていることが示された．

## 2.5 術者の発声と手渡しの時間差

術者の発声から手渡しまでの時間についても評価した．音声認識ツール Whisper では発生時刻が特定できないため，MATLAB の speech detection を適用し音声活動検出 (Voice Activity Detection: VAD) を行い，その結果を Whisper で解析してみたが満足のいく認識結果は得られなかった．しかし，器材の種類にかかわらず手渡しを検出するのであれば，VAD によって検出された発声時刻を手掛かりとすることはできる．図 6 に術者の発声と手渡しの時間差を示した．多くの手渡しは，術者の発声から 2 秒以内に行われており，4 秒以上遅れているものは少ないことが示された．

## 2.6 看護師の顔の向き

手術中に看護師が何処を（何を）見ているのかは，看護師の技量を推定するのに重要な手掛かりとなると考えられている．しかし，手術中は保護メガネを着用しているため，視線の検出は困難である．そこで，顔の向きを検出することにより，看護師の視点推定に代えることとした．顔の向き検出に関しても，マスクにより口などは隠されており，手掛かりとなるのは目（メガネ），鼻，首であった．図 7 に，顔の向きの検出結果を示し，図 8 に，顔が向きの傾向を示す．図中，-150 から 0 は画面向かって左のテーブルに向いており，0 から 100 は前の機材テーブル，100 から 200 は術野，あるいは，術者に向いていると考えられる．この結果，看護師は器材テーブル全体と術野（あるいは術者）を，ほぼ半々に見ていることを示している．

## 2.7 手渡し時刻の判読が困難な例

今回の実験の録画からは手渡し時刻の判読が困難な例を幾つか見つけられた (図 9)．その原因は，以下に示す 5 つであった．

1. 術者が直前に別の動作を行う
2. 長い器材を渡す

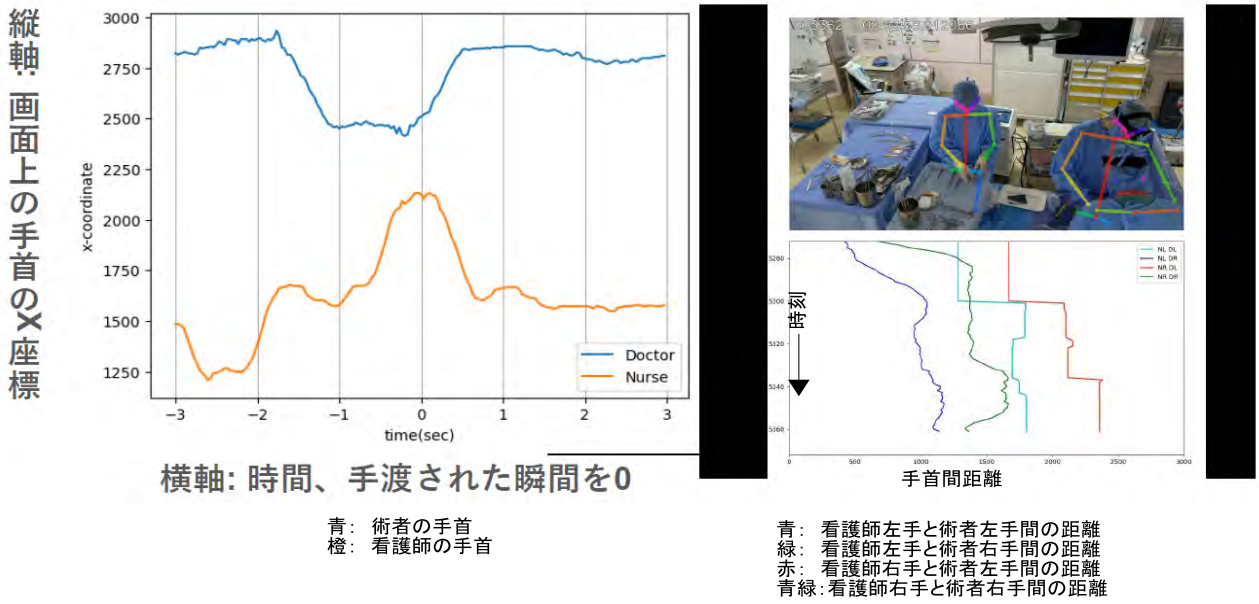



図 4: 術者が手を出してから渡されるまでの時間

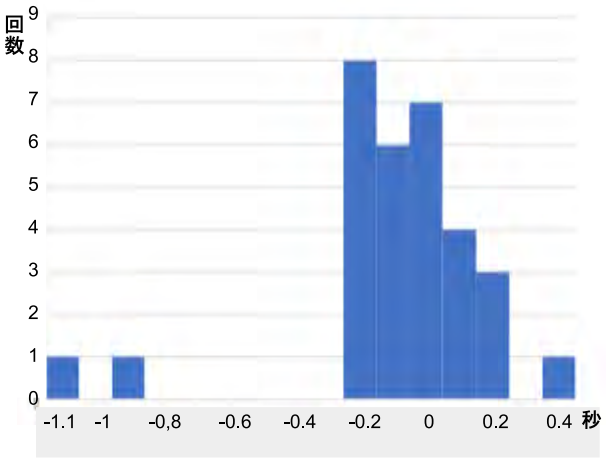



図 5: 手首間の距離最小と手渡しの時間差

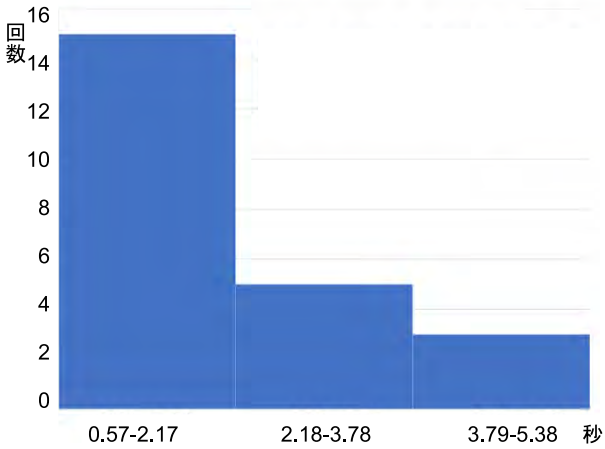


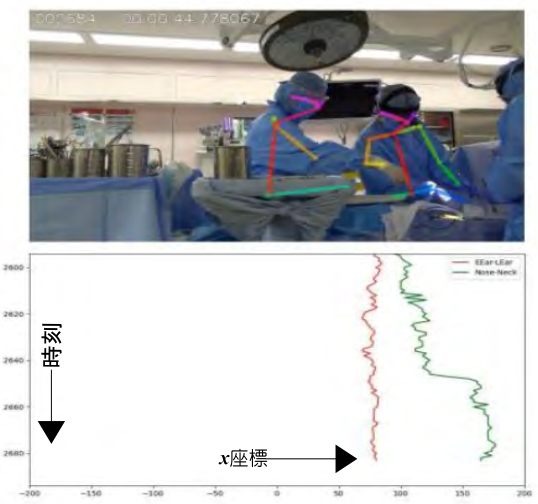

図 6: 術者の発声と手渡しの時間差

3. 看護師が先に手を伸ばす
4. 術者が手を伸ばさない
5. 術者が左手で受け取る

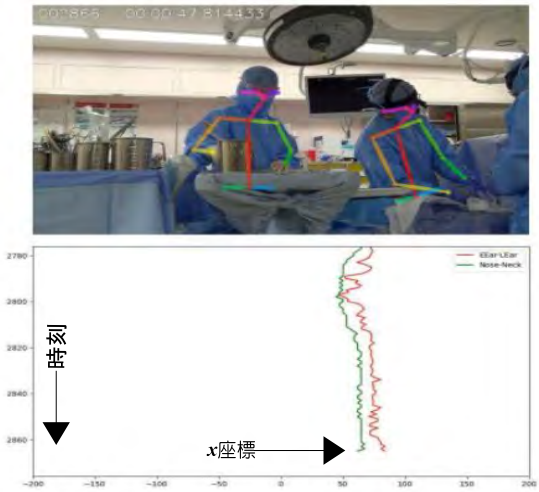
## 3 考察

看護師の技量評価を自動化するために AI ツールを用いて手術器具の手渡しを検出する試みを行った．現行の AI ツールの適用により，一部の課題は確認されたが，これらの課題は将来の研究の有望な方向性を示している．具体的には，看護師と医師の手首間の距離に基づく検出法は，単一カメラの限界を露呈したが，これは複数カメラの統合によって克服できる可能性がある．また，誤検出はあったものの，検出漏れはなく，潜在的な手渡しの候補を絞り込むための有効なツールとして機能した．一方で，現状では，手術器材の学習サンプルが不足しているため，器材の検出追跡に関して十分な性能が得られないことが確認された．

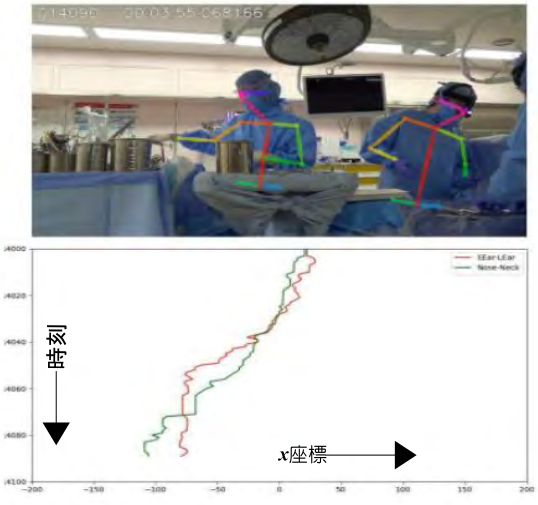
術者の発声に基づく検出では，録音状態の問題から，Whisper による器材名の識別は出来なかった．これは，ビデオカメラ付属のマイクによる録音のため，マスクによる声のこもり，周囲の雑音の影響が大きかったと考えられる．術者にヘッドセットを装備してもらうことなどにより改善が期待される．また，器材名の識別




(a) 術野方向




(b) 器材テーブル1方向




(c) 器材テーブル2, 3方向

図 7: 看護師の顔の向き検出

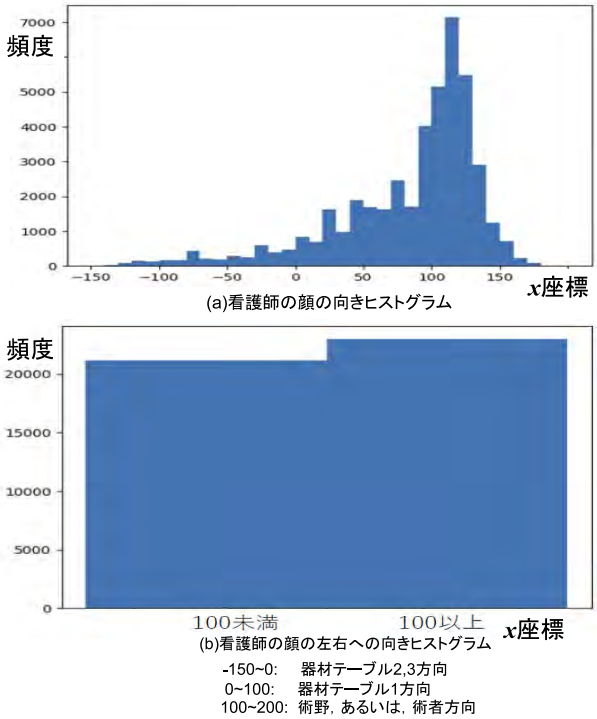



図 8: 看護師の顔の向きの傾向

は出来なくとも，術者の発声を捉えることで，手渡し検出，および，手渡しまでの時間推定が可能である．これを利用することで，手渡し検出精度の向上が期待されるとともに，看護師の技量評価の一助となると考えられる．

手術中の看護師が何処を見ているかも技量評価の大きなポイントとなるが，術中は保護メガネを着用しているため視線検出は困難である．そこで，視線検出に代え顔の向き推定を行った．目（保護メガネ）と鼻が検出できたため，大まかな顔の向き推定が可能であることが示された．これにより，看護師が何処を見ているかの推定が可能であることが示された．

## 4 結論

本研究では，模擬手術動画に AI ツールを適用し，看護師の技量評価の自動化の可能性を検証した．現在の技術では誤検出の問題が存在するが，これは今後の改善の余地を残している．特に，術者の要求と手渡しのタイミングの差を推定することで，看護師の技量評価の基準を設定する新しいアプローチが見出された．さらに，術者の発声を検出トリガとして利用することで，精度の向上が期待できる．今後の研究では，器材検出のための学習モデルの構築，複数カメラの視点統合，録音状態の改善など，さらなる精度向上に向けた取り組みが予定されている．これらの進展は，手術室での看

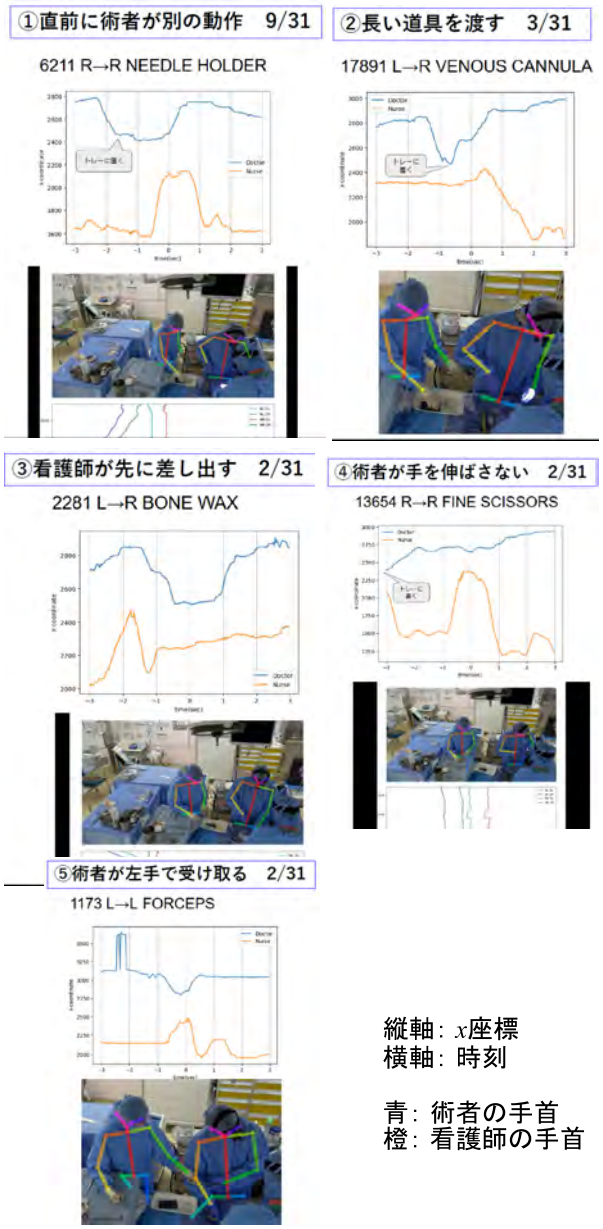



図 9: 手渡し時刻の判読が困難な例

護師の役割をより効果的にサポートし，全体的な医療の質を向上させる大きな一歩となると考えられる．

今後は，器材検出のための学習モデルの構築，複数カメラ視点の統合，録音状態の改善など，精度向上のための検討を行っていきたい．

## 参考文献

[1] Rebecca Andrea Conradsen Skov, Jonathan Lawaetz, Lars Konge, Lise Westerlin, Eske Kvanner Aasvang, Christian Sylvest Meyhoff, Katja Vogt, Tomas Ohrlander, Timothy Andrew Resch, and Jonas Peter Eiberg. Simulation-based education of endovascular scrub nurses reduces stress and improves team performance. *Journal of Surgical Research*, Vol. 280, pp. 209–217, 2022.

[2] Rosalind CJ Siah, Ping Xu, Cheang L Teh, and Alfred WC Kow. Evaluation of nursing students' efficacy, attitude, and confidence level in a perioperative setting using virtual-reality simulation. In *Nursing Forum*, Vol. 57, pp. 1249–1257. Wiley Online Library, 2022.

[3] Kentaro Hara, Tamotsu Kuroki, Masashi Fukuda, Toru Onita, Hiromi Kuroda, Emi Matsuura, and Terumitsu Sawai. Effects of simulation-based scrub nurse education for novice nurses in the operating room: A longitudinal study. *Clinical Simulation in Nursing*, Vol. 62, pp. 12–19, 2022.

[4] Morteza Nasiri, Shahrzad Yektatalab, Marzieh Momennasab, and Fatemeh Vizeshfar. Development and assessment of validity and reliability of a checklist to evaluate the circulating and scrub skills of operating room novices (cssorn checklist). *Journal of Education and Health Promotion*, Vol. 12, , 2023.

[5] Marie-Stéphanie Bracq, Estelle Michinov, Marie Le Duff, Bruno Arnaldi, Valérie Gouranton, and Pierre Jannin. Training situational awareness for scrub nurses: Error recognition in a virtual operating room. *Nurse education in practice*, Vol. 53, p. 103056, 2021.

[6] Isabel Funke, Sebastian Bodenstedt, Florian Oehme, Felix von Bechtolsheim, Jürgen Weitz, and Stefanie Speidel. Using 3d convolutional neural networks to learn spatiotemporal features for

automatic surgical gesture recognition in video. In *International conference on medical image computing and computer-assisted intervention*, pp. 467–475. Springer, 2019.

[7] Shota Kishi, Nozomu Suzuki, Shota Tsuyuki, Takio Kurita, Fujio Miyawaki, and Akinori Hidaka. Convolutional neural network based on temporal pose features for surgical procedure recognition. In *Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications*, Vol. 2021, pp. 60–64. The ISCIE Symposium on Stochastic Systems Theory and Its Applications, 2021.

[8] Adam Goldbraikh, Netanell Avisdris, Carla M Pugh, and Shlomi Laufer. Bounded future ms-tcn++ for surgical gesture recognition. In *European Conference on Computer Vision*, pp. 406–421. Springer, 2022.

[9] Shuja Khalid, Mitchell Goldenberg, Teodor Grantcharov, Babak Taati, and Frank Rudzicz. Evaluation of deep learning models for identifying surgical actions and measuring performance. *JAMA network open*, Vol. 3, No. 3, pp. e201664–e201664, 2020.

[10] Daichi Kitaguchi, Nobuyoshi Takeshita, Hiroki Matsuzaki, Hiroaki Takano, Yohei Owada, Tsuyoshi Enomoto, Tatsuya Oda, Hirohisa Miura, Takahiro Yamanashi, Masahiko Watanabe, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surgical endoscopy*, Vol. 34, pp. 4924–4931, 2020.

[11] Giovanni Menegozzo, Diego Dall' Alba, Chiara Zandona, and Paolo Fiorini. Surgical gesture recognition with time delay neural network based on kinematic data. In *2019 International symposium on medical robotics (ISMR)*, pp. 1–7. IEEE, 2019.

[12] Flakë Bajraktari, Kathrin Fleissner, and Peter P Pott. A comparison of two cnn-based instrument detection approaches for automated surgical assistance systems. In *Current Directions in Biomedical Engineering*, Vol. 9, pp. 599–602. De Gruyter, 2023.

[13] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] Glenn Jocher et al. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, Oct. 2020.

[15] Vadisetti G.P. Niranjan A. Saranu K. Sarma R. Shaik M.A.B. Gudepu, P.R. and P Paramasivam. Whisper Augmented End-to-End/hybrid Speech Recognition System - CycleGAN pproach. In *InterSpeech 2020*, pp. 2302–2302, 2020.

[16] 村上貴志. 音声トリガーも利用した AI による手術自動分析システムの提案と初期評価報告. `https://www.surgicaleducation.jp/images/ses2023/10th_program.pdf`, 2023.

[17] 深層学習を使用したノイズに含まれる音声区間の検出. `https://jp.mathworks.com/help/deeplearning/ug/voice-activity-detection-\in-noise-using-deep-learning.html`.

[18] Labeled Surgical Tools and Images. `https://www.kaggle.com/datasets/dilavado/labeled-surgical-tools`.

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-10

# ドローン聴覚におけるヒストグラム情報を用いた音源定位手法の提案 -周波数抽出とスケーリングの導入による性能向上-

# Proposal of sound source localization method for drone audition using histogram -Improvement of performance by introducing frequency extraction and scaling-

小松崎和泉[1*]　干場功太郎[1]　岩附信行[1]
Izumi Komatsuzaki[1]　Kotaro Hoshiba[1]　Nobuyuki Iwatsuki[1]

[1] 東京工業大学
[1] Tokyo Institute of Technology

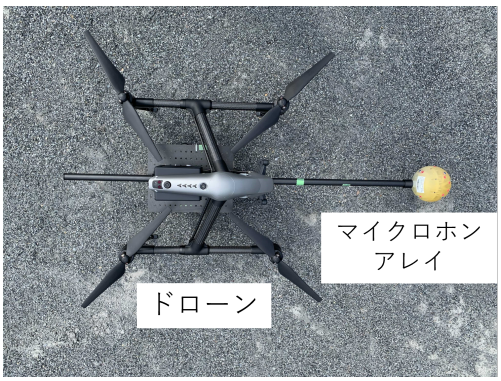
**Abstract:** ドローンを用いた被災者捜索のための音源探査技術において，これまで，著しい時刻変化を伴うドローン自身のエゴノイズに対する耐性，広い捜索範囲，低い計算コスト，汎用性をすべて満たす音源定位手法の開発を目的に，ヒストグラム情報と周波数情報を用いて空間スペクトルにおけるエゴノイズの除去を動的に行う音源定位手法の提案を行った．本稿では，これまでの提案手法におけるリアルタイム性および定位精度をより向上するために，目標音成分の存在する周波数を抽出し，周波数毎に適した基準値を設けてエゴノイズを除去することにより，エゴノイズ近傍を含む広い範囲に存在する目標音に対して，より正確な定位が可能となるよう提案手法の改良を行った．実環境での屋外実験とシミュレーションにより，提案手法の性能を評価した結果，本稿で紹介する改良した提案手法を用いることで，高いリアルタイム性を持ちながら，より高いノイズ耐性と広い捜索可能範囲を獲得することができ，本手法の有用性が確認された．

## 1　はじめに

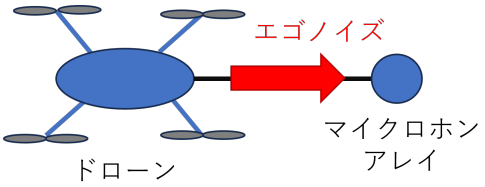
近年，災害地において，人が侵入できない場所にも容易に侵入できること，迅速な活動ができることから，ドローンを用いた要救助者の捜索手法が注目されている．ドローンを用いた捜索では，カメラによる方法が一般的であるが [1]，暗い時間帯の捜索活動，および瓦礫等に埋もれた被災者といったカメラに映らない対象の捜索は困難である．そこで，音情報による捜索手法の確立を目的に，ドローン搭載マイクロホンアレイを用いた音源探査技術の開発が進められている．このような技術はドローン聴覚と呼ばれ，様々な研究が行われている [2][3]．ドローンを用いた音源探査の実用化にあたり，課題の一つがドローンのエゴノイズである．風や飛行状態の影響により著しい時刻変化を伴うエゴノイズに対する耐性，広い捜索範囲，ドローン搭載の小型コンピュータを用いて実時間で捜索を行うためのリアルタイム性，どのような機体・状況でも捜索可能な高い汎用性を持った音源探査手法が求められる．

これまで，音源探査手法として，その分解能の高さから MUSIC (MUltiple SIgnal Classification) 法 [4] に基づく音源定位手法が提案されてきている．一般的な MUSIC 法である SEVD-MUSIC (MUSIC based on Standard Eigen Value Decomposition) 法は，計算コストが低くリアルタイム性が高い反面，ノイズ耐性が低い．そこで，事前収録したノイズの相関行列を用いてノイズ成分を除去する GEVD-MUSIC (MUSIC based on Generalized Eigen Value Decomposition)[5] や GSVD-MUSIC (MUSIC based on Generalized Singular Value Decomposition)[6] が提案された．これらは，SNR (Signal-to-Noise Ratio) の低い状況でも高い音源定位性能を持つが，計算コストが高く，事前に収録したノイズ情報を用いているため時刻変化するノイズへの耐性および汎用性がない．時刻変化するノイズへの耐性の強化および汎用性を補うことを目的に，直前の時刻の収録音をノイズと仮定してノイズ除去を行う iGEVD-MUSIC (incremental GEVD-MUSIC)[7] や iGSVD-MUSIC (incremental GSVD-MUSIC)[8] が提案されている．しかし，これらは，計算コストが高いことに加え，著しく時刻変化するノイズに対する耐性は不十分である．AFRF-MUSIC (MUSIC using Active Frequency ange Filter)[9] は，直前の時刻の収録音を用いて解析周波数を制限することで，小さい計算コストでありながら，エゴノイズを抑制することに成功し

*連絡先：東京工業大学　工学院　機械系
〒 152-8552　東京都目黒区大岡山 2-12-1 I1-27
E-mail: komatsuzaki.i.aa@m.titech.ac.jp

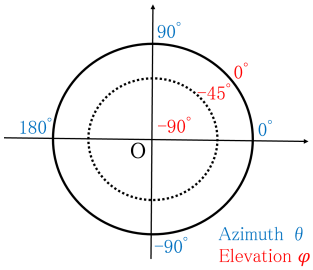



**(a)** マイクロホンアレイ搭載ドローンの一例.

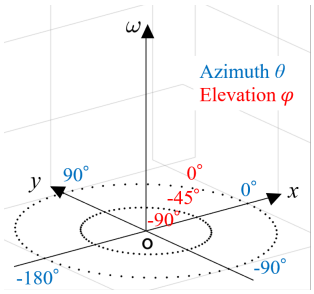



**(b)** マイクロホンアレイに対するエゴノイズの到来方向.

**図 1:** 一定の方向からエゴノイズが到来する配置のマイクロホンアレイ搭載ドローン.




**(a)** 二次元空間スペクトルの座標設定.




**(b)** 三次元空間スペクトルの座標設定.

**図 2:** 座標設定.

た．しかし，こちらも著しく時刻変化するノイズへの耐性は不十分である．また，角度制限 SEVD-MUSIC[10] は，一定の方向からエゴノイズが到来するようなマイクロホンアレイを設計し，SEVD-MUSIC にて得られた空間スペクトルに対し，目標音の捜索範囲からエゴノイズの到来範囲を事前に除外することでノイズ耐性を向上させる．しかし，除外範囲をあらかじめ設定する必要があり，狭い範囲を除外すると時刻変化を伴うノイズに対する十分な耐性が見込めず，広い範囲を除外すると捜索範囲が狭まる．

これらの問題を解決するため，これまでに，過去の情報を用いず，得られた現在時刻の空間スペクトルから，ヒストグラム情報と周波数情報に基づきエゴノイズ成分の判定を行い，動的に捜索範囲の制限による目標音成分の抽出を行う手法である HIST-MUSIC (MUSIC with HISTogram information)[13] および HIST-MUSIC-3D (three-dimentional HIST-MUSIC)[14] を提案した．これにより，高いノイズ耐性，広い捜索範囲，リアルタイム性，汎用性をすべて満たすことができた．しかし，被災者捜索にて活躍するためには，リアルタイム性およびノイズと近傍の方向にある目標音に対する目標音成分の抽出精度が不十分であり，更なる向上が必要と考えた．

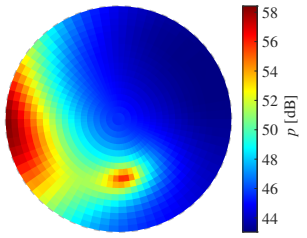
本稿では，著しい時刻変化を伴うノイズに対する耐性，広い捜索範囲，高いリアルタイム性，高い汎用性をすべて満たす音源定位手法の開発を目的に，HIST-MUSIC-3D の改良を行い，より精度の高い空間スペクトルからの目標音成分抽出手法を提案する．本手法では，得られた現在時刻の空間スペクトルから，ヒストグラム情報と周波数情報に基づき，目標音成分の存在する周波数の抽出と，各周波数で最適なノイズ判定の基準値の設定を行うことで，目標音成分の抽出を行う．周波数抽出により，計算コストの大きいノイズ判定の処理が行われる周波数が制限されること，また，周波数毎のノイズ判定の基準値の設定により，各周波数に存在する目標音成分を適切に抽出できることから，リアルタイム性を向上させながら，目標音成分の抽出精度も向上させることができると期待される．本稿では，提案手法のノイズ耐性，捜索範囲，リアルタイム性および汎用性について，シミュレーションおよび屋外実験により評価する．

## 2 HIST-MUSIC-3D

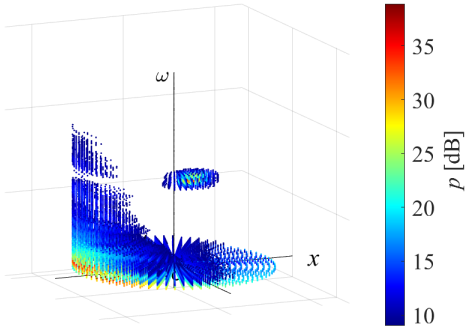
これまでに，現在時刻の空間スペクトルのヒストグラム情報と周波数情報のみからノイズの判定を行い，動的に探索範囲の制限を行う HIST-MUSIC-3D を提案した [14]．空間スペクトルは，計算コストが低い SEVD-MUSIC[4] により得られる MUSIC スペクトルとする．また，これまでに図 1 のようなマイクロホンアレイ搭載ドローンの構造が開発されており，本論文では，一定の方向からエゴノイズが到来するマイクロホンアレイの使用を前提に考える．以下に HIST-MUSIC-3D のアルゴリズムを示す．

SEVD-MUSIC 法では一般的に，周波数毎に算出した空間スペクトル $P(\psi,\omega)$ を周波数方向に加算した二次元空間スペクトル $\overline{P}(\psi)$ を解析に用いる．

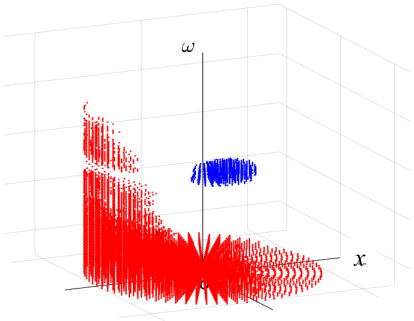
$$\overline{P}(\psi) = \sum_{\omega} P(\psi,\omega) \tag{1}$$



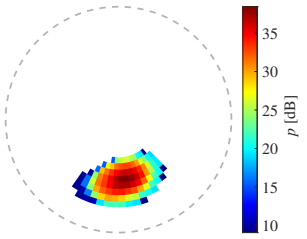

**(a)** $\overline{P}(\psi)$.




**(b)** *Pcut*.




**(c)** *Pcut* におけるエゴノイズ成分（赤色），および目標音成分（青色）の識別.




**(d)** $P'$.

**図 3:** HIST-MUSIC-3D における目標音成分抽出の過程.

$\omega$ とは周波数ビンを表し，$\psi$ はマイクロホンアレイに対する方位角 $\theta$，仰角 $\phi$ から，$\psi = (\theta, \phi)$ と定義する．二次元空間スペクトルの一例を図 3a に示す．ここでは，図 2a の設定軸に従ってプロットされており，各方向から到来した音のパワーをカラーマップで示している．図 3a の目標音方向は，方位角 $-90°$，仰角 $-45°$ である．本手法では，周波数方向に加算する前の周波数情報を含む三次元空間スペクトル $P(\psi, \omega)$ に着目する．

一般的に空間スペクトルのピーク検出により音源定位を行うが，目標音成分の最大パワーより大きいエゴノイズ成分がある場合，正確な目標音源の定位ができない．そこで，目標音成分の最大パワーより小さくかつ近傍な基準値を設定し，基準値より大きいエゴノイズ成分をピーク検出範囲から除外することで，捜索範囲が最大かつ正確な定位が可能になる [13]．基準値として，空間スペクトル $P(\psi)$ の全要素のヒストグラム $H$ を求める．

$$H(p) = \text{histogram}(P(\psi)) \tag{2}$$

ここで，$p$ は空間スペクトルのパワーに対する階級である．得られた $H$ のピーク以降の変曲点を基準値 $p_t$ とする [13]．

$$p_t = p|_{H''(p)=0} \tag{3}$$

各周波数ビンの空間スペクトル $P(\psi, \omega)$ に対して，$p_t$ 以上のパワーを持つ範囲 $\Psi cut(\omega)$ を抽出する．

$$\Psi cut(\omega) = \{\psi | P(\psi, \omega)| > p_t\} \tag{4}$$

$\Psi cut(\omega)$ に含まれる三次元の空間スペクトルの一例を図 3b に示す．ここでは，図 2b の設定軸に従ってプロットされている．エゴノイズは一定方向より到来することから，基準方向 $\psi_0 = (\theta_0, \phi_0)$ を設定し，$\Psi cut(\omega)$ に対して，$\psi_0$ を含まない連続している部分を目標音成分 $\Psi target(\omega)$ として抽出する．

$$\Psi target(\omega) = \Psi cut(\omega) \not\ni \psi_0 \tag{5}$$

それぞれの成分を図示したものが図 3c であり，赤色がエゴノイズ成分，青色が目標音成分を表す．そして，得られた $\Psi target$ に対応する空間スペクトルを周波数方向に足し合わせ，二次元の空間スペクトル $\overline{P'}(\psi)$ を得る (3d)．

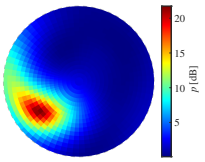
$$\overline{P'}(\psi) = \sum_{\omega} P(\Psi target, \omega) \tag{6}$$

$\overline{P'}(\psi)$ に対して，最大値をとる方向を目標音方向 $\psi target$ として検出する．

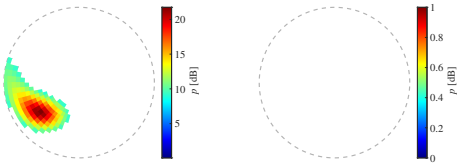
$$\psi_{target} = \text{argmax}_{\psi} \overline{P'}(\psi) \tag{7}$$

## 3 提案手法の改良

HIST-MUSIC-3D を用いた場合，すべての周波数ビンに対してノイズ除外操作を行っていたため，目標音成分が存在していない周波数に対してもノイズ除外操作を行うといった余分な計算コストがかかっている問題点があった．また，すべての周波数ビンに対して等しい基準値でノイズ判定を行っていたため，各周波数ごとの空間スペクトルの値に合ったノイズ判定の基準値でノイズの判定が行えていない問題点があった．その一例を図 4c に示す．図 4a はある 1 つの周波数ビンの空間スペクトルである．図 4b は，図 4a に対して，すべての周波数で等しく設定された基準値以上の成分をプロットしたものであるが，定位に影響のあるノイ

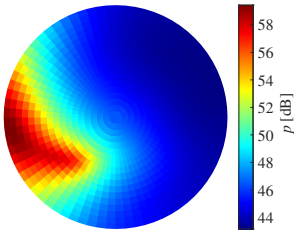



(a) 空間スペクトル.

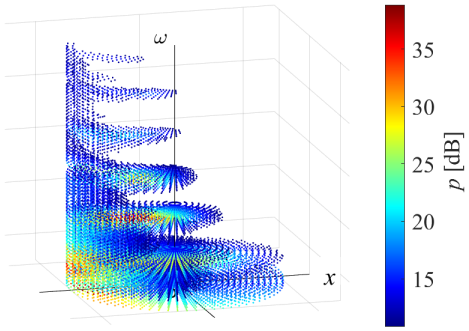



(b) 基準値以上の成分.　　(c) ノイズ除去後.

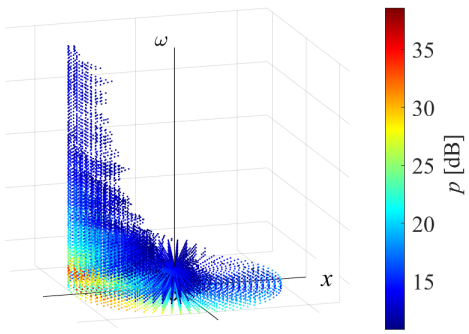
**図 4:** HIST-MUSIC-3D の失敗例.




(a) 二次元空間スペクトル.




(b) 三次元空間スペクトル (目標音とノイズの混合音の処理結果).




(c) 三次元空間スペクトル (ノイズのみの信号の処理結果).

**図 5:** 空間スペクトルの一例.

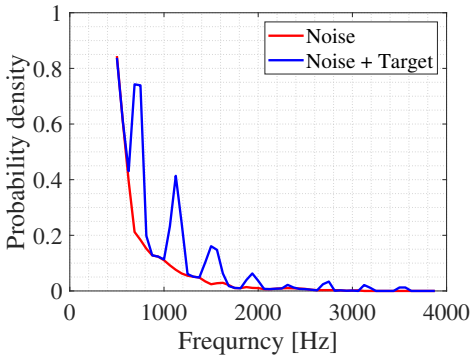



**図 6:** 全方向に対する $\Psi cut(\omega)$ が占める確立密度.

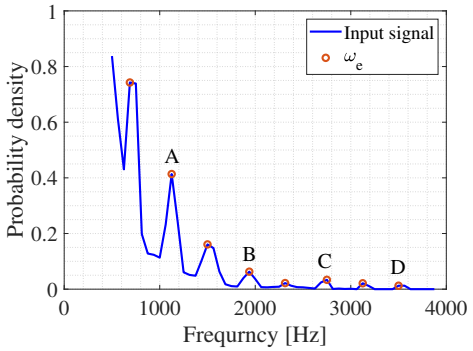
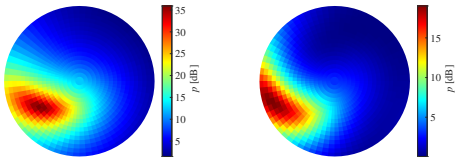
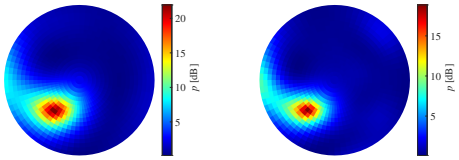



**図 7:** $N$ と $\omega_e$ の一例.




(a) 図 7 中の周波数 A における空間スペクトル.　(b) 図 7 中の周波数 B における空間スペクトル.




(c) 図 7 中の周波数 C における空間スペクトル.　(d) 図 7 中の周波数 D における空間スペクトル.

**図 8:** 抽出した周波数における空間スペクトルの一例.

ズが除外しきれておらず，図 4c のように，目標音成分の抽出に失敗している．これらの問題を解決するため，目標音の存在する周波数の抽出およびヒストグラムのスケーリングを用いた各周波数で異なる基準値の設定を行う．以下にそのアルゴリズムを示す.

## 3.1　周波数抽出

三次元空間スペクトルにおいて，基準値 $p_t$ 以上の要素は，ノイズのみの場合は図 5c，目標音とノイズを含

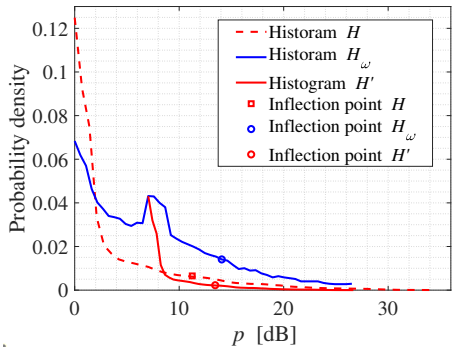
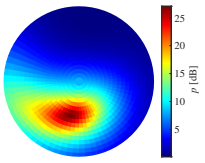
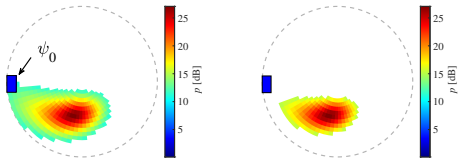



**図 9:** ヒストグラムのスケーリング.




**(a)** 空間スペクトルの一例.




**(b)** 基準値以上の成分 (基準値 $p_t$).

**(c)** 基準値以上の成分 (基準値 $p_t\omega$).

**図 10:** 基準値の違いによる空間スペクトルの除去される範囲の変化.

む場合は図 5b のようになる．このとき，それぞれのプロットされている周波数毎の基準値 $p_t$ 以上の要素数，つまり全要素数に対する基準値 $p_t$ 以上の要素数の確立密度を求める (図 6)．ノイズのみの場合は，周波数が高くなるにつれ確立密度が滑らかに減少するものの，目標音とノイズを含む場合は，目標音成分の存在する周波数で極大値をとる．よって，全要素数に対する基準値 $p_t$ 以上の要素数の確立密度が極大値となる周波数を抽出することで，目標音成分の存在する周波数を抽出することが可能となる．これより，計算コストが大きいノイズ除外操作を目標音成分の存在する周波数対してのみ行うことができるようになり，計算コストが削減されると期待される．

三次元空間スペクトルの全方向に対する $\Psi cut(\omega)$ が占める確立密度を $N$ としたとき，抽出する周波数 $\omega_e$ は次のアルゴリズムで求められる．

$$N(\omega) = Psicut(\omega)/\Psi \tag{8}$$

$$\omega_e = \text{arglocalmaximum}_\omega(N(\omega)) \tag{9}$$

図 7 は $N$ と $\omega_e$ の一例である．抽出した周波数のうち，A〜D の周波数における空間スペクトル $P(\psi,\omega)$ を図 8 に示す．このとき，目標音方向は，方位角 $-135°$，仰角 $-45°$ である．周波数が低い A，B では，ノイズのパワーも大きいため，目標音成分がノイズに引っ張られてしまっているが，全ての抽出した周波数で目標音成分の存在が確認できる．特に，高周波数である C，D では，目標音成分が正確な方向にはっきりと存在していることがわかる．

## 3.2 周波数毎のノイズ判定基準値の設定

HIST-MUSIC-3D では，すべての周波数ビンに対して等しい基準値でノイズ判定を行っていたため，各周波数ごとの空間スペクトルの値に合ったノイズ判定の基準値を周波数毎に設定することで，目標音抽出の精度が向上できると期待される．

HIST-MUSIC-3D にて，ノイズ判定の基準値 $p_t$ である変曲点を求める際，空間スペクトルのヒストグラムを滑らかな関数にするために，近似関数を適応している．しかし，近似計算は計算コストが大きいため，すべての周波数に対してヒストグラムに近似関数を適応し変曲点を求めることは，リアルタイム性を考慮すると現実的ではない．そこで，小さい計算コストで周波数毎のノイズ判定基準値の設定するため，空間スペクトルの全要素のヒストグラムを，空間スペクトルの各周波数の要素のヒストグラムにスケーリングすることで，簡易的に各周波数のヒストグラムにおける変曲点を求める．アルゴリズムは以下の通りである．

スケーリングを行う空間スペクトルの全要素のヒストグラム $H$ に対して，$H$ の極大値のうち階級 $p$ の大きい点における極大値 $pd_{lm}$ と階級 $p_{lm}$，階級の最大値 $p_{max}$ を求めておく．

$$pd^M_{lm} = \text{localmax}(H(p)) \tag{10}$$

$$p^M_{lm} = p|_{H(p)=pd^M_{lm}} \tag{11}$$

$$p_{lm} = \max(p^M_{lm}) \tag{12}$$

$$pd_{lm} = H(p_{lm}) \tag{13}$$

$$p_{max} = \text{argmax}(H(p)) \tag{14}$$

各周波数の空間スペクトルの要素 $P(\psi,\omega)$ のヒストグラム $H_\omega$ を求める．

$$H_\omega = \text{histogram}(P(\psi,\omega)) \tag{15}$$

式 (10)-(14) と同様に，$H_\omega$ の極大値のうち階級の大きい点における極大値 $pd^\omega_{lm}$ と階級 $p^\omega_{lm}$，階級の最大値 $p^\omega_{max}$ を求める．そして，$pd_{lm}$ を $pd^\omega_{lm}$ に，$p_{lm}$ を $p^\omega_{lm}$ に，$p_{max}$を$p^\omega_{max}$ に合わせるよう，$H$ をスケーリングしたヒストグラム $H'(\omega)$ を求める．

$$H'(p) = aH(bp+c)|_{pd_{lm}=pd^\omega_{lm},p_{lm}=p^\omega_{lm},p_{max}=p^\omega_{max}} \tag{16}$$

$H'$ において，$H$ の変曲点が対応する点 $p_t\omega$ を求める．

$$p_t\omega = p_t|_{H'} \tag{17}$$

この値を周波数毎に算出し，それぞれの周波数におけるノイズ判定の基準値として用いる．

スケーリングを行ったヒストグラムの一例を図 9 に示す．赤色の点線で表された $H$ を青の実線で表された

$h$ に合わせてスケーリングすることで，赤色の実線で表された $H'$ を得た．ヒストグラムのスケーリングに伴い，赤四角の位置にある $H$ の変曲点は，赤丸の位置に移動する．これは，近似を用いて算出した，青丸の位置にある $h$ の変曲点と近い階級値を取る．よって，ヒストグラムのスケーリングを行うことで，近似を用いずに簡易的に各周波数のヒストグラムにおける変曲点を求めることができることがわかった．また，図 10 に改良前と改良後のノイズ判定の基準値を用いた際の，目標音抽出の精度を示す．図 10a に対して，改良前の基準値 $p_t$ を用いたとき，多くのノイズが除外しきれず，青四角で示されたドローンの方向である基準方向 $\psi_0$ を含む連続した部分に目標音成分も含まれてしまっている (図 10b)．一方で，改良後の基準値 $p_t\omega$ を用いたときは，目標音成分が $\psi_0$ と離れて抽出できており，目標音がノイズと誤判定されずに抽出できていることがわかる (図 10c)．

## 4 評価実験

提案手法のノイズ耐性，捜索範囲，計算コストおよび汎用性を検証するため，評価実験を行った．屋外実環境にて収録したエゴノイズと，シミュレーションにより作成した任意の方向から到来した目標音を加算することにより評価用信号を作成し，解析を行った．

エゴノイズは，DJI 社製 Inspire 2 (図 11a) と ACSL 社製 MS-06LA (図 11b) の 2 種類のドローンに搭載したマイクロホンアレイにて収録した．プロペラには normal と high の異なる 2 種類のプロペラを用いた [11]．normal は一般的なプロペラであり，低周波数に強いエゴノイズが発生する．high altitude はエゴノイズの低周波数成分を抑制する代わりに高周波数のパワーが強い特徴を持つ．DJI 社製 Inspire 2 を用いた際のマイクロホンアレイには，下半球に 12ch，上半球に 4ch の MEMS マイクロホンが設置されている 16ch 球形マイクロホンアレイ (図 11c)[12] を用いた．また，ACSL 社製 MS-06LA を用いた際のマイクロホンアレイには，上半球の 4ch は使用せず，下半球のみの 12ch 球形マイクロホンアレイを用いた．本マイクロホンアレイでは，サンプリング周波数 16 kHz，量子化ビット数 24 bit で音響信号が収録される．マイクロホンアレイはドローンの中心から 600 mm の位置に設置した (図 c )．高度 10 m でホバリング中，および速度 1 m/s，2 m/s，3 m/s で飛行時のエゴノイズを収録した．目標音のサンプルには声およびホイッスルを用い，方位角が $-180° \leq \theta < 180°$，仰角が $-90° \leq \phi \leq 0°$ の範囲で 5° 刻みで到来方向を設定し，幾何計算により得た伝達関数から各方向から到来した目標音を作成した．そして，収録したノイズと作成した目標音を，SNR が 4 dB 刻みで $-20 \sim 0$ dB となるよう加算し，評価用信号を作成した．

まず，評価用信号を SEVD-MUSIC により解析し，MUSIC スペクトルを求めた．SEVD-MUSIC のパラメータは，目標音源数を 2，相関行列に用いる平均化フレーム数を 50，最小解析周波数を 500 Hz，最大解析周波数を 4000 Hz とした．得られた MUSIC スペクトルに対し，HIST-MUSIC-3D[14]，本稿で改良した提案手法において，周波数抽出のみ導入した手法 (以後，HIST-MUSIC-3D+FE (Frequency Ectraction) と呼ぶ)，周波数毎のノイズ判定の基準値のみ導入した手法 (以後，HIST-MUSIC-3D+SC (SCaling) と呼ぶ)，周波数抽出を行った後，抽出された周波数のみに対して周波数毎のノイズ判定の基準値の設定と以降のノイズ除外操作を行う手法 (以後，HIST-MUSIC-3D+FE,SC と呼ぶ) により処理を行い比較した．表 1 に比較条件，図 12 にそれぞれの比較条件の流れを示す．基準方向は，$\psi_0 = (-180°, 0°)$ とした．

ノイズ耐性および捜索範囲は，normal のプロペラを取り付けた DJI 社製 Inspire 2 のエゴノイズを用いた評価用信号に対して処理を行った．

リアルタイム性は，仮想マシンを用い，評価用信号に対して処理を行い評価した．仮想マシンは Jetson Nano を想定し，メモリ数 4GB，CPU4 コアとした．評価指標にはリアルタイム性能を示す指標である RTF (Real Time Factor) を用いた．RTF は，(演算時間)/(処理信号の時間) のように算出する．つまり，RTF $< 1$ の場合，処理信号の時間以内に演算が終わるため，遅延時間が蓄積せずにリアルタイムに処理が行えることとなる．0.5 秒の信号を用いて 200 回試行を行い，平均値を評価値とした．

汎用性は，high のプロペラを取り付けた DJI 社製 Inspire 2 と ACSL 社製 MS-06LA の 2 種類のエゴノイズを用いた評価用信号に対し処理を行い，音源定位性能を比較した．

**表 1:** 比較に用いた手法

| |
|---|
| 1. SEVD-MUSIC |
| 2. HIST-MUSIC-3D |
| 3. HIST-MUSIC-3D+FE |
| 4. HIST-MUSIC-3D+SC |
| 5. HIST-MUSIC-3D+FE,SC |

## 5 結果

SNR と目標音源，ドローンの機種，プロペラ，飛行状況の異なる様々な条件で作成した評価用信号を，表 1 で示した各手法で処理した結果を図 13 に示す．

空間スペクトルは図 2a の設定軸に従ってプロットされており，各方向から到来した音のパワーをカラーマップで示している．SNR はすべての結果で −12dB である．エゴノイズは，a，b はともにエゴノイズに normal のプロペラを取り付けた DJI 社製 Inspire 2，c は high のプロペラを取り付けた DJI 社製 Inspire 2，d は ACSL 社製 MS-06LA のエゴノイズを用いた．目標音サンプルは，a は声，b〜d はホイッスルとした．目標音方向は，a,c,d が方位角 $\theta = -135°$，仰角 $\phi = -45°$，b が方位角 $\theta = -160°$，仰角 $\phi = -13°$ の結果である．い


(a) マイクロホンアレイ搭載ドローン (DJI 社製 Inspire 2).


(b) マイクロホンアレイ搭載ドローン (ACSL 社製 MS-06LA).


(c) 12ch 球形マイクロホンアレイ.

**図 11:** 実験にて使用したドローンおよびマイクロホンアレイ.

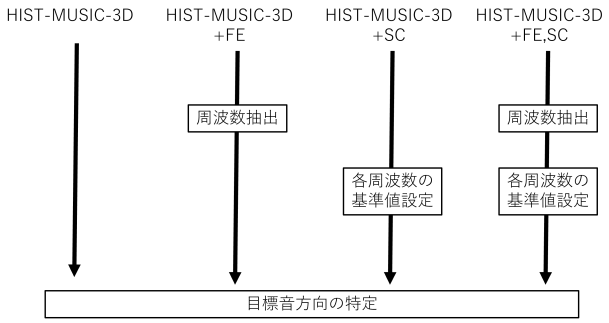



**図 12:** 比較条件の流れ.

ずれの結果も，エゴノイズは空間スペクトルの左側に，目標音は左下側，ノイズに対して右下側にノイズと重なる位置に現れる．

## 5.1 ノイズ耐性と捜索範囲の評価結果

ノイズ耐性と捜索範囲の評価結果を図 13a,b に示す．

SEVD-MUSIC の結果は，エゴノイズの右下に目標音が飛び出すように現れているが，エゴノイズの方がパワーが大きい．HIST-MUSIC-3D は，a の結果は目標音の抽出が行えているが，より目標音がエゴノイズと近い方向にある b の結果では目標音がノイズと誤認識され抽出ができていない．HIST-MUSIC-3D+FE も，同様の理由で，b の結果では目標音がノイズと誤認識され抽出ができていない．しかし，HIST-MUSIC-3D の結果 b にて左上側に僅かであるが残っているパワーの大きいエゴノイズが，HIST-MUSIC-3D+FE の b の結果では除外できている．これより，目標音の存在する周波数の抽出することは，リアルタイム性を向上させるだけでなく，目標音の定位に影響のあるパワーの大きいエゴノイズのみ存在する周波数を除外する効果もあるがわかった．HIST-MUSIC-3D+SC は，b の結果で目標音方向とエゴノイズ側にずれた位置に最大ピークが表れており，定位に影響のあるパワーの大きいエゴノイズが除外しきれていないため，正確な定位が行えていない．しかし，HIST-MUSIC-3D では除外されてしまった目標音成分の存在が確認できるため，周波数毎にノイズ判定の基準値を設定することで，目標音成分抽出の精度が向上していることがわかる．一方で，HIST-MUSIC-3D+FE,SC は，どちらの結果も目標音の抽出が行えており，目標音が最大ピークとなった．これは，目標音成分の存在する周波数の抽出を行うことで，パワーの大きいエゴノイズのみ存在する周波数では除去しきれなかったノイズが周波数抽出の時点で除去できるため，HIST-MUSIC-3D+SC は，b の結果では定位が失敗していたものの，HIST-MUSIC-3D+FE,SC では定位か可能になったと考えられる．よって，目標音の存在する周波数の抽出を行い，周波数毎に最適なノイズ判定の基準値を設けることで，エゴノイズと近い方向にある目標音を含む広い捜索範囲において，定位が可能であると期待される．

さらに，評価用音響信号 50 フレーム (25 秒分)，飛行状況 4 種類，目標音 2 種類，音源方向 1,297 通り，SNR6 通りの全 3,112,800 回の試行を行い，各手法における，捜索可能率 (図 14b) を求めた．

音源定位の成功率は，全試行のうち，定位が成功した試行数の割合を示す．つまり，ノイズ耐性と捜索可能範囲を統合した結果である．ここでは，表 1 で示した各手法に加えて，従来手法として，GEVD-MUSIC，iGEVD-MUSIC，AFRF-MUSIC，角度制限 SEVD-MUSIC についても評価を行った．GEVD-MUSIC で使用するノイズ相関行列には、別日の実験にて収録した DJI 社製 Inspire 2 のエゴノイズを用いた．iGEVD-MUSIC と AFRF-MUSIC のパラメータは，ノイズと定位音源のフレーム差を 50，ノイズの相関行列のフレーム数は iGEVD-MUSIC は 100，AFRF-MUSIC は 50 とした．また，AFRF-MUSIC の解析周波数帯域は 500〜4000 Hz の範囲内の 500Hz とした．角度制限 SEVD-MUSIC は，−90〜90°（角度制限 SEVD(1)）と −135〜135°（角度制限 SEVD(2)）の 2 つの捜索範囲を設定した．

HIST-MUSIC-3D は，ほとんどの SNR で従来手法に比べて高い成功率を獲得したが，SNR が −4 0dB のとき，AFRF-MUSIC よりも成功率が下回った．一方で，HIST-MUSIC-3D+FE，HIST-MUSIC-3D+SC，HIST-MUSIC-3D+FE,SC は，すべての SNR で従来手法に比べて高い成功率を獲得した．しかし，HIST-MUSIC-3D+FE は，SNR が −20dB のとき，HIST-MUSIC-3D より低い成功率となった．これは，HIST-MUSIC-3D は，エゴノイズと近い方向にある目標音に

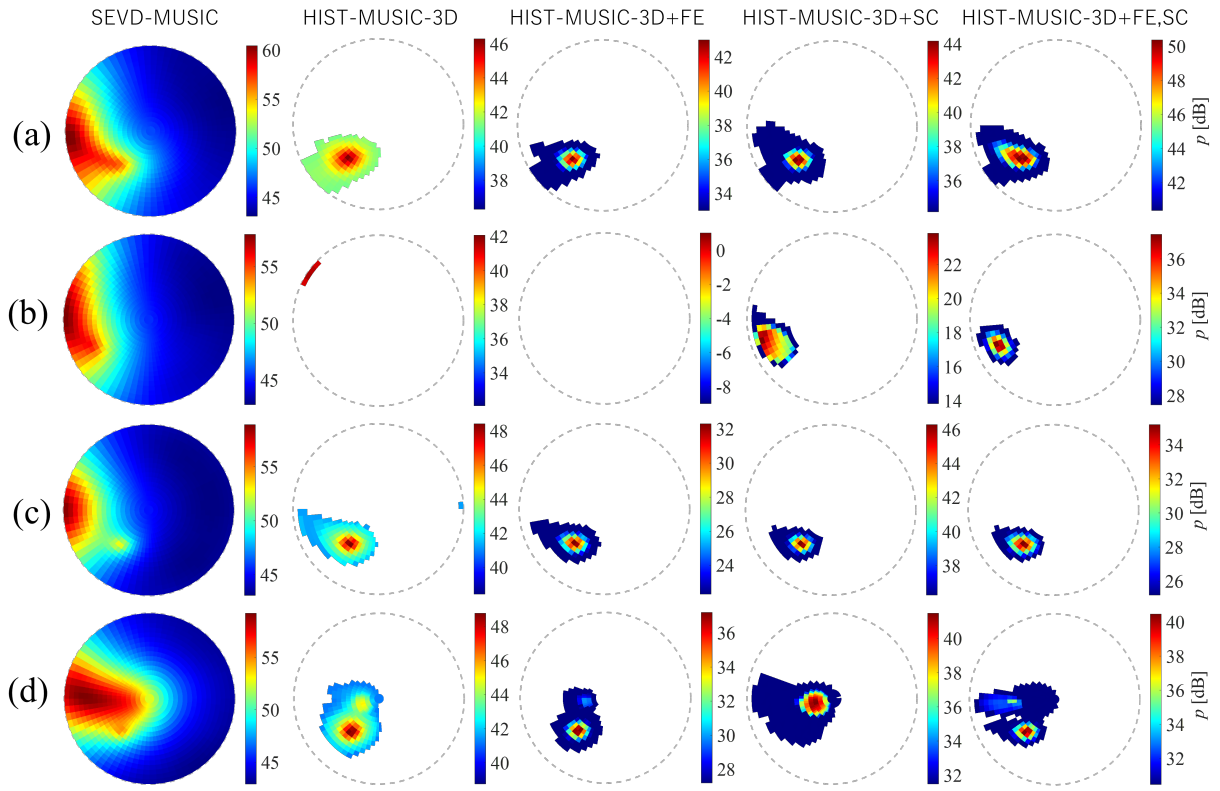



**図 13:** 各手法を用いて得られた空間スペクトル．SNR = −12dB.
(a) エゴノイズ：Inspire 2 (normal)，目標音源：声，$\psi_{target} = (-135°, -45°)$，(b) エゴノイズ：Inspire 2 (normal)，目標音源：ホイッスル，$\psi_{target} = (-160°, -25°)$，(c) エゴノイズ：Inspire 2 (high altitude)，目標音源：ホイッスル，$\psi_{target} = (-135°, -45°)$．(d) エゴノイズ：MS-06LA，目標音源：ホイッスル，$\psi_{target} = (-135°, -45°)$．

**表 2:** 各手法での RTF の結果．

| 処理条件 | RTF |
|---|---|
| SEVD-MUSIC | 0.92 |
| HIST-MUSIC-3D | 0.97 |
| HIST-MUSIC-3D+FE | 0.95 |
| HIST-MUSIC-3D+SC | 1.09 |
| HIST-MUSIC-3D+FE,SC | 0.96 |

対して，目標音が除外されているにも関わらず，目標音付近 (目標音方向と誤差 $\mp 5°$ 以内) のエゴノイズが除外しきれず最大ピークとなり定位成功となる場合があること，また，目標音の存在する周波数の抽出に失敗し，定位が失敗していることも原因として考えられる．HIST-MUSIC-3D+SC はすべての SNR で最も高い捜索可能率を獲得した．これは，HIST-MUSIC-3D+FE,SC では，目標音の存在する周波数の抽出ができず定位が失敗している場合があることが原因として挙げられる．一方で，SNR が −20dB と低いとき，HIST-MUSIC-3D+FE,SC の成功率は HIST-MUSIC-3D+SC とほとんど等しく，高性能であることがわかった．

## 5.2 リアルタイム性の評価結果

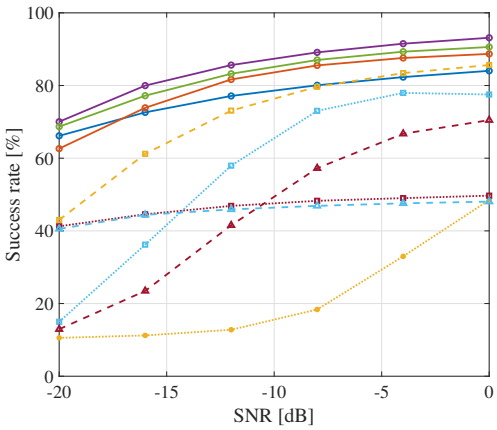
リアルタイム性の評価結果を表 2 に示す．

SEVD-MUSIC は各方向の MUSIC スペクトルを固有値展開により算出しているため，計算コストが小さい．HIST-MUSIC-3D は SEVD-MUSIC に単純な計算を加える処理を行うため，SEVD-MUSIC より僅かに大きくなったもののリアルタイム性を獲得できている．HIST-MUSIC-3D+FE は，最も RTF が小さくなった．これは，抽出された周波数のみに対してノイズ除外操作を行うため，計算コストが比較的大きいノイズ除外操作の回数が減ることで，周波数抽出の計算を加えても処理全体としては計算コストが小さくなったためである．HIST-MUSIC-3D+SC は，RTF $> 1$ となり，リアルタイム性が獲得できなかった．これは，すべての周波数に対してヒストグラム算出とノイズ除外操作を行うため，計算コストが高くなったことが原因である．HIST-MUSIC-3D+FE,SC は，RTF $> 1$ となり，また，これまでの提案手法に比べても小さい RTF となった．よって，周波数の抽出を行うことで，ヒストグラムの算出とノイズ除外操作を行う周波数を制限することで，リアルタイム性を獲得できることがわかった．

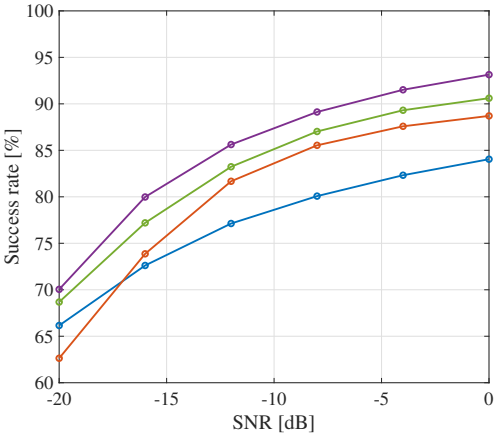
## 5.3 汎用性

汎用性の評価結果を図 13c,d に示す．

SEVD-MUSIC の結果は，左側にエゴノイズ，その右下側に目標音が確認できるが，エゴノイズの方がパワーが大きい．HIST-MUSIC-3D は，どちらの結果も




**(a)** 各手法の成功率の比較.




**(b)** (a) の拡大図（提案手法のみ）.

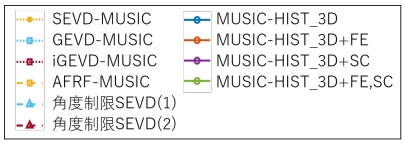



**図 14:** 音源定位の成功率.

目標音の定位が行えている．HIST-MUSIC-3D+FE も，両方の結果で目標音が最大ピークとなったが，d の結果において，これまでの提案手法の結果にも表れているノイズが，目標音の上側に残っている．よって，SEVD-MUSIC の d の結果のようにエゴノイズが仰角方向に大きく広がっているとき，HIST-MUSIC-3D+FE では，エゴノイズの先端部分，つまり仰角 $-90°$ 付近のエゴノイズ成分が目標音成分のようにふるまうことが原因で，ノイズのみの周波数をすべて除外できないことがわかった．また，この除外しきれていないノイズは，HIST-MUSIC-3D+SC で大きく現れており，最大ピークとなってしまっている．こちらも，エゴノイズの先端部分の成分が目標音成分のようにふるまうことで，エゴノイズの先端成分を目標音として誤って抽出するようなノイズ判定の基準値が設定されてしまっていることがわかる．HIST-MUSIC-3D+FE,SC についても，同様に目標音の上側にエゴノイズが除外しきれず残っているが，目標音が最大ピークとなり，周波数の抽出と周波数毎のノイズ判定の基準値の設定をどちらも取り入れる手法は，汎用性を獲得できることがわかった．

### 5.4 結果のまとめ

ノイズ耐性，捜索範囲の結果から，HIST-MUSIC-3D+SC が最も捜索可能率が高く，少し性能が下回るものの，HIST-MUSIC-3D+FE,SC も特に低い SNR において高い捜索可能率を獲得していることがわかった．一方で，HIST-MUSIC-3D+FE は，捜索可能率がこれまでの提案手法と同程度であり，他の改良した提案手法に比べて劣っていた．計算コストの結果から，HIST-MUSIC-3D+FE は最も計算コストが小さく，また HIST-MUSIC-3D+FE,SC もリアルタイム性を獲得していることがわかった．一方で，HIST-MUSIC-3D+SC は，リアルタイム性を獲得できなかった．汎用性の評価結果から，HIST-MUSIC-3D+FE と HIST-MUSIC-3D+FE,SC は汎用性を得られたが，HIST-MUSIC-3D+SC は汎用性を得られなかった．以上のことを踏まえて，HIST-MUSIC-3D+FE,SC が最もすべての求められる音源定位手法の性能を満たしており，最も有用性があると期待できる．

## 6 結論

本稿では，著しい時刻変化を伴うノイズに対する頑健性，広い捜索範囲，低い計算コスト，汎用性をすべて満たす音源定位手法の開発を目的に，過去の情報を用いず，得られた現在の空間スペクトルから，ヒストグラム情報と周波数情報に基づき，目標音成分の存在する周波数の抽出と各周波数で最適なノイズ判定の基準値を設定することで，ノイズの判定と目標音成分の抽出を行う手法を提案した．これまでの提案手法では，パワーの大きいエゴノイズが除外しきれない場合があることや，エゴノイズと近傍の方向に存在する目標音がエゴノイズと誤判定され定位が不可能であるといった問題点があったが，ノイズ除外操作を行う周波数を目標音の存在する周波数のみ制限すること，各周波数に存在するエゴノイズ，目標音成分のパワーに合ったノイズ判定の基準値を設けるよう提案手法を改良することにより，これまでの提案手法における問題点を解決することができた．評価実験の結果，高いノイズ耐性，広い捜索範囲，リアルタイム性，汎用性をすべて考えたとき，周波数の抽出と周波数毎のノイズ判定の基準値を設定する処理をどちらも導入する手法が最も有用性があると確認できた．今後は，屋外実環境にて評価を行い，改良した提案手法の実用性の評価を行っていく．

## 謝辞

本研究は，JSPS 科研費 22K14218 の助成を受けた．

## 参考文献


[1] 加藤，寺島，高見: 要救助者の複数ドローンによる協調探索のためのエッジサーバ集約型自動スケジューリング手法とシミュレーション評価マルチメディア，分散協調とモバイルシンポジウム *2019* 論文集, pp.291–296 (2019)

[2] L. wang, A. Cavallaro: Deep-Learning-Assisted Sound Source Localization From a Flying Drone *IEEE Sensors Journal* VOL. 22, NO. 21, pp. 20828-20838 (2022)

[3] L. wang, A. Cavallaro: Drone Ego-Noise Cancellation for Improved Speech Capture using Deep Convolution Autoencoder Assisted Multistage Beamforming *2022 25th International Conference on Information Fusion (FUSION)* (2022)

[4] R. O. Schmidt: Multiple Emitter Location and Signal Parameter Estimation *IEEE Trans. Antennas and Propagation*,VOL. 34, NO. 3, pp. 276-280 (1986)

[5] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, H. Tsujino: Intelligent Sound Source Localization for Dynamic Environment, *IEEE/RSJ International Conference on Intelligent Robots and Systemd* (2009)

[6] K. Nakamura, K. Nakadai, G. Ince: Real-time Super-resolution Sound Source Localization for Robots *Proc. of IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)* pp. 694-699 (2012)

[7] K. Okutani, T. Yoshida, K. Nakamura, K. Nakadai: Intelligent Sound Source Localization for Dynamic Environment, *Outdoor Auditory Scene Analysis Using a Moving Microphone Array Embedded in a Quadrocopter* (2012)

[8] T. Ohata, K. Nakamura, A. Nagamine, T. Mizumoto, T. Ishizaki, R. Kojima, O. Sugiyama, K. Nakadai: Outdoor Sound Source Detection Using a Quadcopter with Micrphone Array *J. of Robotics and Mechatronics* VOL. 29, NO. 1, pp. 177-187, (2017)

[9] K. Hoshiba, K. Nakadai, M. Kumon, H. G. Okuno: Assessment of MUSIC-Based Noise-Robust Sound Source Localization with Active Frequency Range Filtering, *Journal of Robotics andMechatronics*, VOL.30, NO. 3, pp. 426-435, (2018)

[10] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, H. G. Okuno: Design of UAV-Embedded Microphone Array System for Sound Sorce Localization in Outdoor Environments *Sensors* VOL. 17, NO. 11, pp. 1-16, (2017)

[11] R. Noda, T. Nakata, K. Senda, H. Liu : Multi-scale morphological effect on noise level and frequency characteristics of drone propellers *QUIET DRONS International e-Symposium on UAV/UAS Noise* pp. 77–142, (2020)

[12] K. Nonami, K. Hoshiba, K. Nakadai, M. Kumon, H.G. Okuno, Y. Tanabe, K. Yonezawa, H. Tokutake, S. Suzuki, K. Yamaguchi, S. Sunada, T. Takaki, T. Nakata, R. Noda, H. Liu, S. Tadokoro: Recent R&D Technologies and Future Prospective of Flying Robot in Tough Robotics Challenge *Disaster Robotics - Results from the ImPACT Tough Robotics Challenge, Satoshi Tadokoro Ed., Springer International Publishing* pp. 77–142, (2019)

[13] 小松崎，干場，武田，菅原: ”ヒストグラム情報を用いた時刻変化の著しい雑音に対する体制の高い音源定位手法の提案”，第 *40* 回日本ロボット学会学術講演会, RSJ2022AC4J3-06, (2022)

[14] 小松崎，干場，岩附:”ドローン聴覚におけるヒストグラム情報と周波数情報を用いた音源定位性能向上の検討”，第 *61* 回人工知能学会 *AI* チャレンジ研究会, pp. 26-32, (2022)

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-11

# 高分解能な音源定位のための展開可能なマイクロホンアレイの設計

# Design of deployable microphone array for high-resolution sound source localization

LEE DONG WOO[1]　干場功太郎[1]*　岩附信行[1]
Lee Dong Woo[1]　Kotaro Hoshiba[1]　Nobuyuki Iwatsuki[1]

[1] 東京工業大学
[1] Tokyo Institute of Technology

**Abstract:** 災害による被災地において，迅速な捜索・救助活動は重要な課題である．近年，マイクロホンアレイをドローンに搭載し，被害者の声といった音を頼りに被災者を捜索する音源定位の研究が行われている．これまで，ドローンの周囲にフレームを設置し，フレーム上にマイクロホンを配置した大型のマイクロホンアレイや，ドローンのアームの先に設置する小型の球形マイクロホンアレイなどが開発されているが，前者は定位の分解能は高いが，フレームの設置が難しい，飛行時の安定性に欠けるという問題点，後者は設置が容易であり，飛行時の安定性も高いが，分解能が低いといった問題点があった．そこで本稿では，飛行時にはコンパクトに収納され，被災者捜索時には大きく展開可能なマイクロホンアレイを提案する．航空宇宙工学分野で提案された展開構造物を参考に，少ない駆動源で展開可能な過拘束リンク機構からなる展開構造物を設計し，展開型マイクロホンアレイを試作した．屋内実験により，提案した展開型マイクロホンアレイの定位の分解能を評価し，その有用性を確認した．

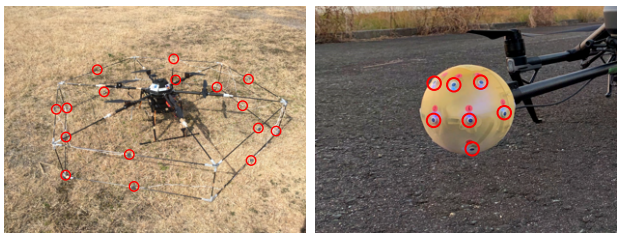
## 1　はじめに

災害が発生した際，最も重要なのは，被害者の捜索救助活動である．The first 72 hour Response として知られるように，災害発生から 72 時間が過ぎると被害者の生存率が減少すると言われており，迅速な捜索救助活動が求められる [1]．近年，人による被害者捜索が困難である場所において，迅速な捜索を行うためにドローンを用いた捜索活動が注目されている [2]．しかし，現在開発されているドローンを利用した捜索活動は，カメラ画像を用いることが多いため [3, 4, 5]，人が瓦礫等に埋もれている場合や鮮明な画像が取得できない夜間などの暗い時間帯には捜索活動を行うことが難しいという問題がある [6]．そこで，声などといった被災者由来の音響信号を，複数のマイクロホンにより構成されるマイクロホンアレイを用いて取得し，音源位置を探索することで被災者の場所を特定する音源探査手法が研究されている [7]．また，ドローンから計測信号を照射し，地表からの反射波をマイクロホンアレイにより観測し，その伝搬時間と到来方向により地形のセンシングを行う研究も行われている [8]．このような技術はドローン聴覚と呼ばれ，さまざまな研究が行われている．

一般的に，マイクロホンアレイを用いて音が到来する方向を観測する場合，構成する各マイクロホンの間隔が大きいほど解析結果の分解能が高いとされている [9]．図 1 に，これまでに音源定位に用いられてきたマイクロホンアレイの代表例を示す．1 つ目は図 1a のように，ドローンの周囲に直径 1.8 m のフレームを設置し，フレーム上に 16 個のマイクロホンを貼り付けたものである [10]．こちらの各マイクロホンの間隔は大きく，分解能は高いが，フレームの設置が難しく，またフレームが大きいため，飛行の安定性に欠ける．2 つ目は図 1b のように，直径 100 mm 程度の球体の筐体に 16 個のマイクロホンアレイを埋め込んだものである [11]．設置が容易になり，飛行時の安定性も高いが，各マイクロホンの間隔が小さいため，分解能が低い．そこで，飛行時には小さく収納され，音源探査時には大きく展開できる構造を持ったマイクロホンアレイを開発することができれば，高い音源定位の分解能，飛行時の高い安定性，設置の容易さを満たすことができる可能性がある．

展開可能な構造物は特に航空宇宙工学分野でよく研究されている．中でも，人工衛星のアンテナと太陽光パネルの分野で最も研究が進んでいる．人工衛星のアンテナは大きければ大きいほど地球との通信能力が高い．同様に，太陽光パネルも面積が大きいほど発電能力が高い．しかし，大きなアンテナや太陽光パネルをそのまま人工衛星に付けて宇宙空間に輸送するのは，輸送費用的に非効率であり，また飛行時に受ける力に耐えられる構造にする必要がある．そのため，人工衛星を発射する時にはコンパクトに収納され，宇宙空間で

*連絡先：東京工業大学　工学院　機械系
〒 152-8552　東京都目黒区大岡山 2-12-1 I1-27
E-mail: hoshiba@rmsv.mech.e.titech.ac.jp



**(a)** Large microphone array. **(b)** Small microphone array.

**図 1:** Example of drone-embedded microphone arrays. Microphones are marked as red circles.

大きく展開できるような展開構造物に対する研究が注目されている．Shah らは折り紙パターンを分析し，軽量，低費用で大きく展開できる折り紙パターンアンテナを提案した [12]．Wang らは花びらの折りたたみ・展開過程をモチーフとし，新しい平面展開機構を提案し，太陽光パネルを製作した [13]．しかし，これらの構造は質量が大きい，展開時に必要な駆動源が多いなどといった問題点があるため，これらをドローン搭載マイクロホンアレイに採用するには適切ではない．そこで，展開可能なリングトラス型アンテナを作るために，平面展開メカニズムから基本ユニットを提案し，基本ユニットの合成により展開可能なアンテナ構造物を構築した Han らの研究に着目した [14]．

本稿では，音源定位の分解能，飛行時の安定性の向上を目的に，ドローン搭載可能な展開型マイクロホンアレイの開発を行う．少ない駆動源で展開でき，かつ少ないリンクで構成される構造物である，Han らが提案した展開構造物の基本ユニットから最適なユニットを選定し，それらを反復配置することで，正八角形の展開構造物を設計した．設計した展開構造物にマイクロホンアレイを設置し，室内実験により音源定位の性能を評価した．

## 2 展開構造物の設計およびプロトタイプ製作

本章では，展開可能なマイクロホンアレイのための展開構造物の設計・試作を行う．

### 2.1 基本ユニットの選定

ドローンのペイロードは限られているため，駆動源を含めた展開構造物は軽量である必要がある．よって，少ない駆動源で展開でき，かつ少ないリンクで構成される軽量な構造物である，Han らが提案した展開構造 [14] を採用する．本展開構造では，図 2 のような単純な 5 つの基本ユニットが提案されている．本稿では，提案されている 5 つの基本ユニットから，適当な基本ユニットを選択する．地表の音源を高精度に定位するためには，マイクロホンアレイは雑音源，つまりプロペラより下に設置されるのが好ましい．しかし，ドローン下部には着陸用の脚があるため，自由に展開することが難しい．そこで，あらかじめドローンに付属している脚の代わりに，展開型マイクロホンアレイを着陸用の脚として使うことを考える．その場合，展開前後の高さを維持させるために，長方形の高さを保ちながら幅を変化させる必要があるため，図 2(a)，(e) は適当ではない．また，飛行中に空中で展開することを考えると，直進対偶 (P ジョイント) と比べ，回転対偶 (R ジョイント) はより柔軟性の高い動きが可能なため [12]，図 2(c)，(d) も適当ではない．以上の理由から，図 2(b) の 6R 基本ユニットを選定した．選定した 6R 基本ユニットを 8 個反復配置することにより，正八角柱形となる構造物を構築する．構築した構造を図 3 に示す．正八角形の一辺の中点から頂点までの長さを $l$，正八角形の中心点から頂点までの長さ (半径) を $r$ とすると，正八角形の中心角は 0° であるため，式 のように $r$ と $l$ の関係を表すことができる．

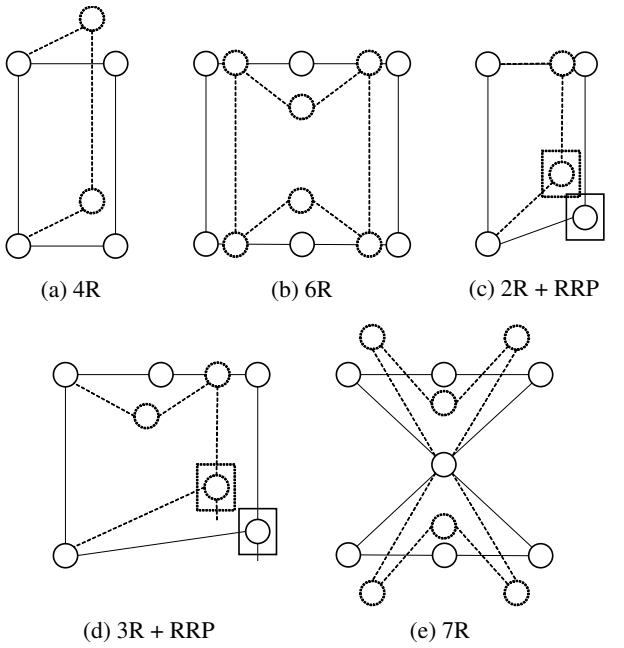



**図 2:** Five simple basic units for square deployable structure.

$$l = r \sin \frac{45^\circ}{2} \tag{1}$$

以降，この $l$ の部分に当たるリンクを回転リンク，上段と下段の正八角形を接続する支柱を縦リンクと呼ぶことにする．ここで本機構の自由度を求める．1 つの対偶の左右に 2 つの回転対偶があることに注意し，展開構造物の自由度 $F$ を求めると，式 2 のようになる．

$$F = 6N - J - 1 + \sum_{i=1}^{J} f_i = -6 \tag{2}$$

ここで，$N$ はリンクの数 ($N = 4 \times 8 + 8 = 40$)，$J$ は対偶の数 ($J = 2 \times 8 + 2 \times 2 \times 8 = 48$) であり，自由度は $-6$ であることがわかる．自由度が 0 以下である

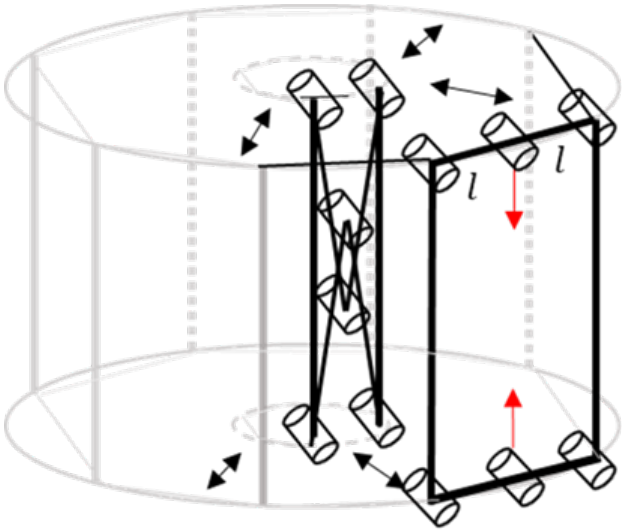



**図 3:** Designed regular octagonal deployable mechanism.

**図 4:** Designed joint with two revolute pairs.

ため，運動機構とならず構造物となる．しかし，いくつかのリンク長が等しい，いくつかの対偶の姿勢が等しいなどといった機構の寸法の特異性により運動可能となる構造物が存在し，それを過拘束機構と呼ぶ [15]．本稿で構築した機構においても，基本ユニットのすべての回転対偶の対偶軸が平行であり，回転リンクを接続している上下 2 つの回転対偶の対偶軸が展開前後の正八角の軸に垂直に交わること，2 本の縦リンクが並行であること，また，同一の基本ユニットを 8 個接続して軸対称構造をとるという機構寸法の特異性により運動可能な過拘束機構となる．本稿で構築した過拘束機構は自由度が 1 となるため，1 つの駆動源で展開でき，前述した求められる性能を満たすことができる．本機構は，高さを維持しつつ，正八角形の半径が拡大・縮小するような動作を行う．

### 2.2 基本ユニット間のジョイント部の検討

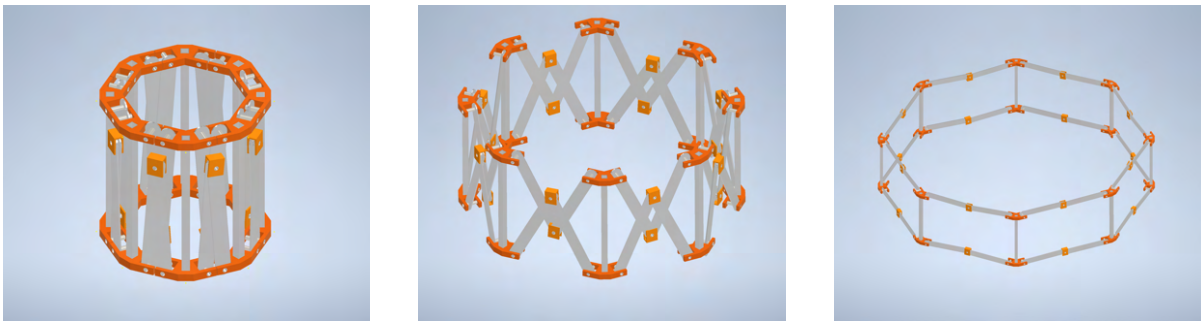
図 3 で示した展開構造物を実際に製作するにあたり，注目すべき点は正八角形の頂点にあるジョイントである．正八角形を維持するように，ジョイントの左右に回転リンクが接続され，このロッドの回転によって正八角形の一辺の長さが調整されることになる．従って，ジョイントの左右を回転対偶とする必要がある．このような条件から，図 4 のようなジョイントを設計した．

### 2.3 設計および試作

これらの展開機構およびジョイントに基づいて，図 5 に示す CAD モデルを作成し，実際に図 6 に示されるプロトタイプを製作した．本稿では，展開時に外径 900 mm となるマイクロホンアレイのための構造を考えた．式 1 で求めた $r$ と $l$ の関係から，$r = 450$ mm，$l = 172.2$ mm となる．しかし，ジョイント部の長さも考慮し，回転リンクの長さを 170 mm とした．縦リンクの長さは，回転リンクより長い必要があるため，200 mm とした．この場合，理論上では，格納時の直径が 170 mm となる．ジョイントは 3D プリンタにて ABS 材により作成し，縦リンクはアルミ角パイプ (A6063)，回転リンクは厚さ 2 mm のアルミ板 (A1100)，回転軸はアルミ丸パイプ (A6063) を用いた．試作した機構の質量は 1007.8 g となり，中型のドローンであれば十分搭載できる質量であると言える．

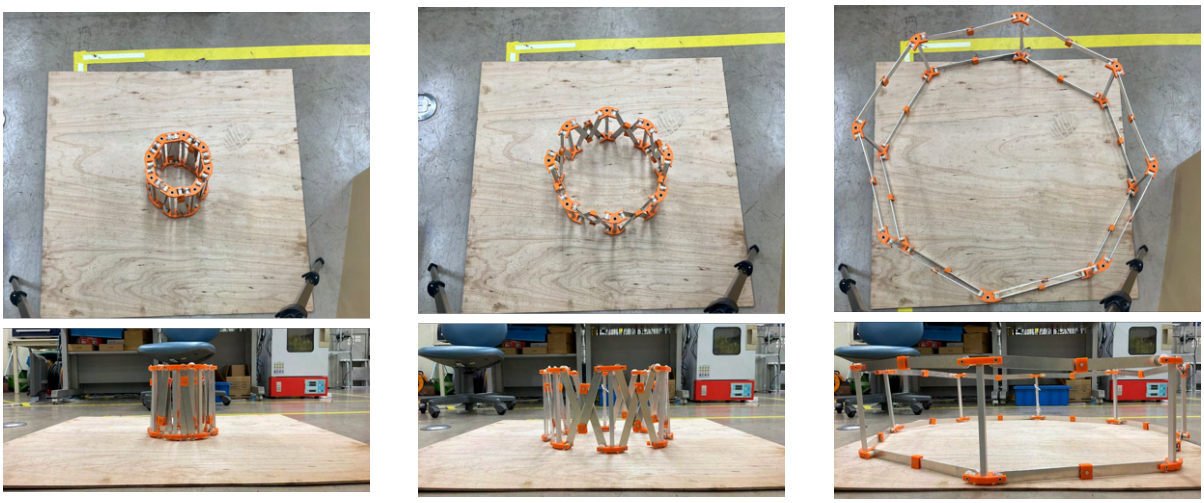
## 3 評価実験

試作した機構にマイクロホンを設置することでマイクロホンアレイを構築し，評価実験を行った．図 7 に示されるように，正八角形の頂点および縦リンク上に，計 12 個の MEMS マイクロホンを設置し，直径 900 mm の 12ch マイクロホンアレイとした．本実験は，マイクロホンアレイに対し，異なる 3 方向に存在する音源を定位する際の分解能を評価する．実験状況を図 8 に示す．マイクロホンアレイを机の上に固定し，音源 1〜3 の位置からスピーカーにより評価用信号を再生し，処理を行う．マイクロホンアレイに対する音源 1，2，3 の方向 (方位角 *theta*, 仰角 *phi*) はそれぞれ，(90°, 0°)，(80°, 35°)，(100°, 49°) であり，マイクロホンアレイからの距離は約 3〜5 m である．評価用信号には 0-8 kHz のアップチャープ信号を用いた．収録は，本稿で試作したマイクロホンアレイに加え，図 1b に示す，直径 100 mm の球形 16ch マイクロホンアレイを用いて，比較を行った．それぞれのマイクロホンアレイは，中心座標が一致するよう設置した．なお，球形 16ch マイクロホンアレイは，本稿で試作したマイクロホンアレイとチャンネル数を合わせるため，上半球に配置されている 12ch のみで収録を行った．音響信号はサンプリング周波数 16 kHz，量子化ビット数 24 bit で収録した．収録された音響信号は，音源定位でしばしば用いられる MUSIC (Multiple SIgnal Classification) 法 [16] を用い



(a) Fully Folded. (b) Half Deployed. (c) Fully Deployed.

**図 5:** CAD model.



(a) Fully Folded. (b) Half Deployed. (c) Fully Deployed.

**図 6:** Prototype.

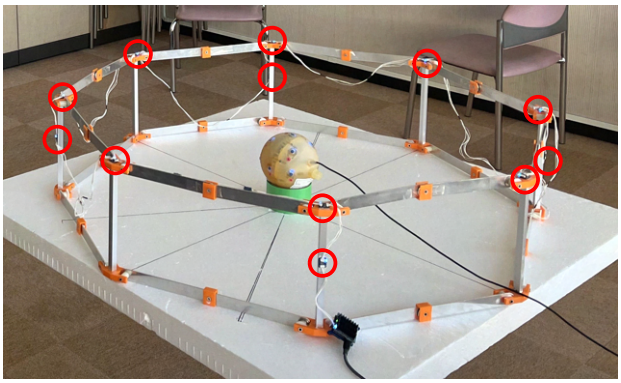


**図 7:** Deployable 12ch microphone array. Microphones are marked as red circles.

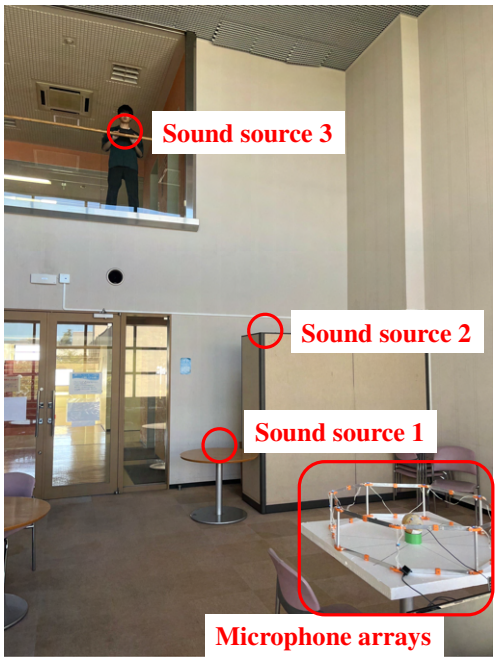



**図 8:** Experimental configuration.

て処理を行う．

## 4 結果

各マイクロホンアレイにて収録した音響信号を，MUSIC 法にて処理し，得られた空間スペクトルの結果を比較する．空間スペクトルは図 9 に従いプロットされ

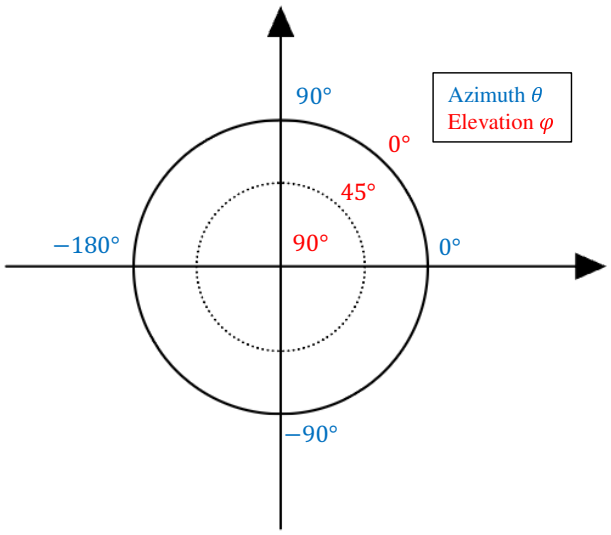



**図 9:** Setting of the coordinate system.

る．図 10 に球形マイクロホンアレイにて得られた空間スペクトル，図 11 に展開型マイクロホンアレイにて得られた空間スペクトルを示す．(a)〜(c) は音源 1〜3 に対する結果である．また，(i) は解析周波数が 500〜1000 Hz，(ii) は 2500〜3000 Hz，(iii) は 5500〜6000 Hz の結果である．空間スペクトルはカラーマップにて各方向から到来する音のパワーを示している．球形マイクロホンアレイの結果では，音源方向にピークが確認できるものの，幅の広い分布となっていることがわかる．特に低周波域で顕著にその傾向が確認できる．一方，展開型マイクロホンアレイでは，球形マイクロホンアレイと比較し，鋭いピークとなっており，高い分解能を持っていることがわかった．ただし，アレイを構成する各マイクロホン間の距離が大きいため，特に高周波域にてエイリアスとみられる成分が複雑に発生しており，音源とその他の成分の判別が難しくなることが予想される．よって，展開型マイクロホンアレイは，SNR (Signal-to-Noise Ratio) が大きく，音源とその他の成分の分離が容易である場合には高分解能での音源定位が期待できる．

## 5　考察

得られた空間スペクトルから分解能を算出し，考察する．分解能は，ピークの値から −3 dB となる範囲を抽出し，方位角と仰角の幅を算出した．図 12 に方位角方向の，図 13 に仰角方向の分解能を示す．これらは音源 1〜3 で平均されている．横軸が解析周波数，縦軸が分解能である．球形マイクロホンアレイの分解能は，方位角，仰角ともに，周波数が高くなるにしたがって分解能は小さくなるが，ほぼすべての周波数において 15° 以上であり，最大では 40° を超える．一方，提案した展開型マイクロホンアレイでは，方位角の分解能は球形マイクロホンアレイと比較すると 1/3 以下の値となっており，多くの周波数にて 5° 以下の分解能を得ることができた．仰角の分解能も球形マイクロホンアレイと比較すると 1/2 以下の値であり，優位性はあるものの方位角よりも分解能が大きくなっている．これは，展開型マイクロホンアレイに含まれるマイクロホンの多くは同一平面上に配置されており，方位角方向には分解能が小さくなるが，仰角方向には不利であることが原因と考えられる．また，展開機構の精度の影響で，展開時の各マイクロホンの位置と伝達関数との間に差異が発生してしまい，方位角，仰角ともの予想よりも分解能が大きくなった．

以上の結果から，提案した展開型マイクロホンアレイの音響的な有用性が確認された．

## 6　結言

本稿では，ドローン聴覚における，音源定位の分解能向上を目的に，ドローンに搭載可能な展開型マイクロホンアレイを提案した．航空宇宙工学分野で提案された展開構造物の基本ユニットから，最適なものを選定し，正八角形柱が高さを維持しつつ拡大・縮小する空間展開構造物を設計した．実際に 3D プリンタを用いて展開型マイクロホンアレイを試作し，音源定位の分解能を評価した．その結果，これまで用いられていた小型のマイクロホンアレイと比較し，分解能が向上することがわかり，提案した展開型マイクロホンアレイの有用性が確認された．今後は，展開の正確性やドローンに搭載した際の駆動などについて検討を行い，実際に実機に搭載し評価を行っていく予定である．

## 7　謝辞

本研究は，JSPS 科研費 22K14218 の助成を受けた．

## 参考文献

[1] 小谷稔, 飯塚敦, 河合克之: 急性期災害医療における DMAT 配置モデルに関する考察, 土木学会論文集 F6（安全問題）, Vol. 71, No. 1, pp. 32-45, 2015, DOI: 10.2208/jscejsp.71.32.

[2] Luo, C., Miao, W., Ullah, H., McClean, S., Parr, G., Min, G.: Unmanned aerial vehicles for disaster management, *Geological Disaster Monitoring Based on Sensor Networks*, pp. 83-107, 2019, DOI: 10.1007/978-981-13-0992-2-7.

[3] Meier, L., Tanskanen, P., Fraundorfer, F., Pollefeys, M.: PIXHAWK: A system for Autonomous Flight using Onboard Computer Vision, *Proceedings of IEEE International Conference on Robotics and Automation (ICRA2011)*, pp. 2992-2997, 2011, DOI: 10.1109/ICRA.2011.5980229.

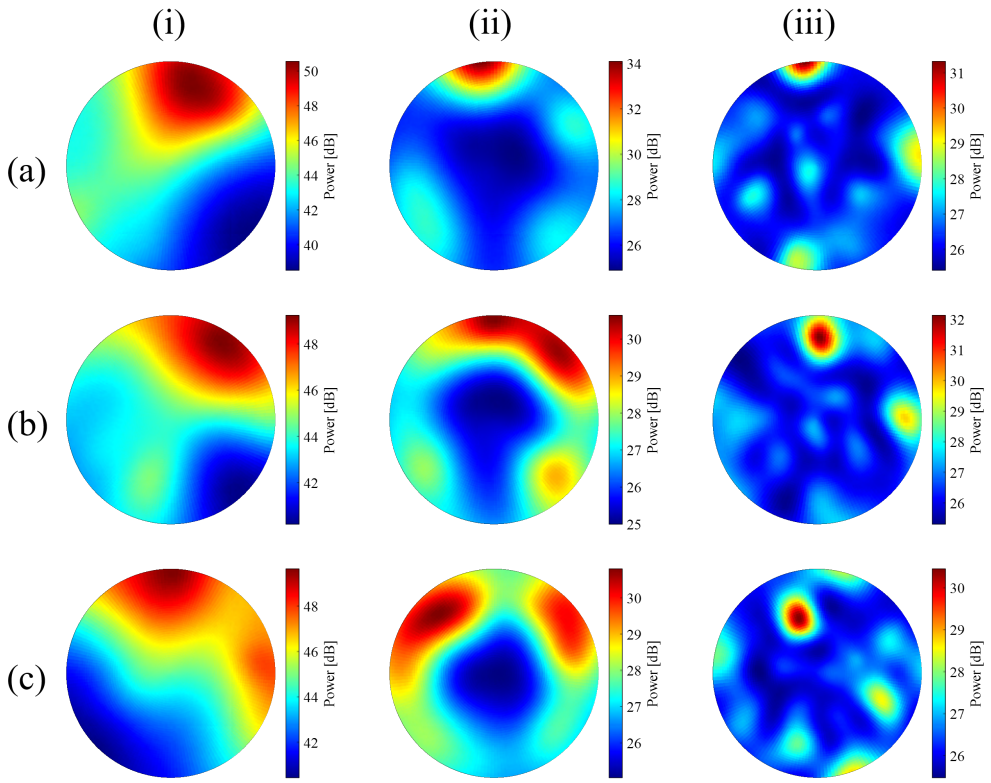



**図 10:** Spatial spectra obtained by small spherical microphone array. (a) Sound source 1, (b) Sound source 2, (c) Sound source 3. (i) 500-1000 Hz, (ii) 2500-3000 Hz, (iii) 5500-6000 Hz.

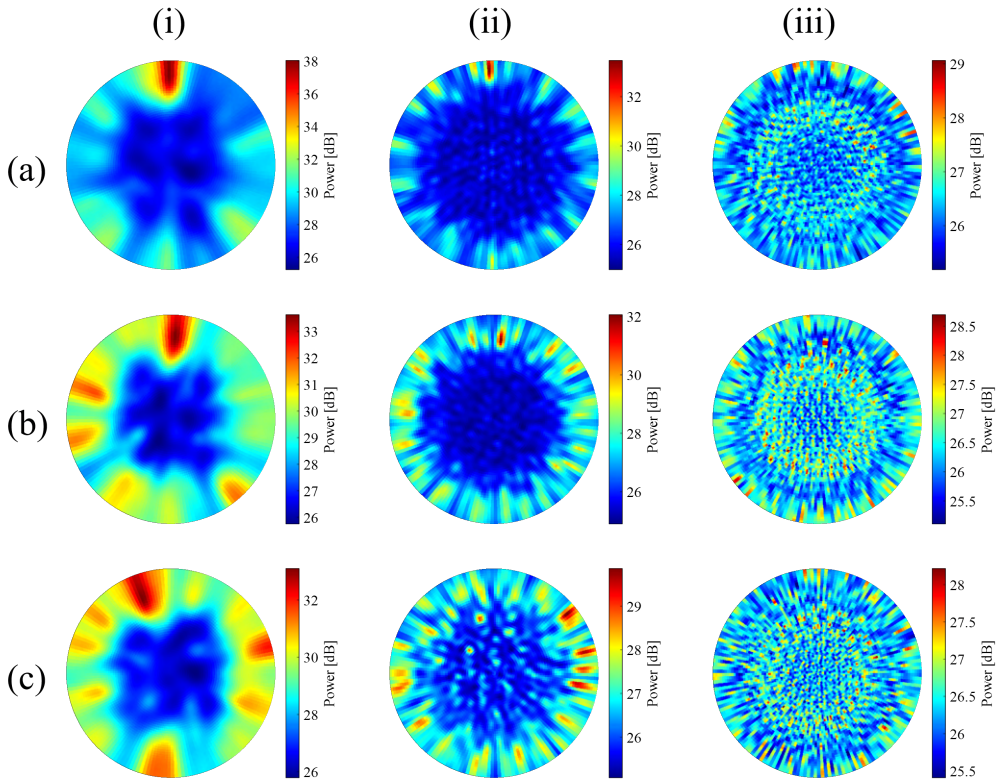



**図 11:** Spatial spectra obtained by proposed deployable microphone array. (a) Sound source 1, (b) Sound source 2, (c) Sound source 3. (i) 500-1000 Hz, (ii) 2500-3000 Hz, (iii) 5500-6000 Hz.

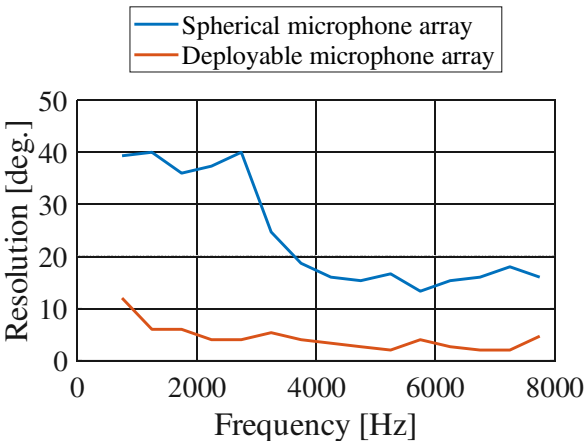



**図 12:** Resolution of azimuth direction.

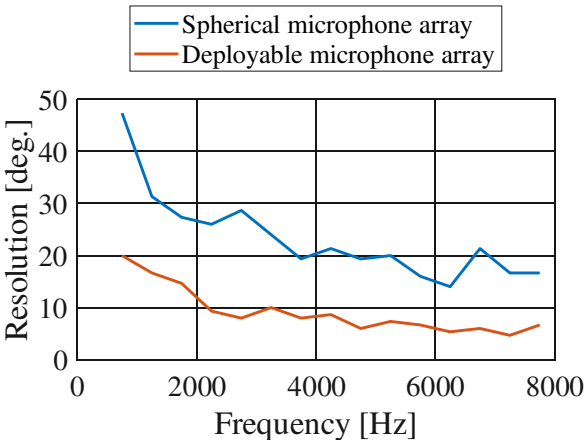



**図 13:** Resolution of elevation direction.


[4] Achtelik, W. M., Lynen, S., Weiss, S., Kneip, L., Chli, M., Siegwart, R.: Visual-Inertial SLAM for a Small Helicopter in Large Outdoor Environments, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2012)*, pp. 2651-2652, 2012, DOI: 10.1109/IROS.2012.6386270.

[5] Lee, S., Har, D., Kum, D.: Drone-Assisted Disaster Management: Finding Victims via Infrared Camera and Lidar Sensor Fusion, *Proceedings of 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE 2016)*, pp. 84-89, 2016, DOI: 10.1109/APWC-on-CSE.2016.025.

[6] Sandino, J., Maire, F., Caccetta, P., Sanderson, C., Gonzalez, F.: Drone-Based Autonomous Motion Planning System for Outdoor Environments under Object Detection Uncertainty, *Remote Sensing*, Vol. 13, No. 21, 4481, 2021, DOI: 10.3390/rs13214481.

[7] Basiri, M., Schill, F., Lima, P. U. and Floreano, D.: Robust acoustic source localization of emergency signals from Micro Air Vehicles, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2012)*, pp.4737-4742, 2012, DOI: 10.1109/IROS.2012.6385608.

[8] 干場功太郎, 岩附信行: ドローン聴覚による地表のアクティブ音響センシングに関する基礎検討, 第 40 回日本ロボット学会学術講演会講演予稿集, RSJ2022AC4J3-10, 2022.

[9] 浅野太, 音のアレイ信号処理 - 音源の定位・追跡と分離 -, コロナ社, 2011.

[10] Wakabayashi, M., Okuno, H. G. and Kumon, M.: Multiple Sound Source Position Estimation by Drone Audition Based on Data Association Between Sound Source Localization and Identification, *IEEE Robotics and Automation Letters*, Vol. 5, No. 2, pp.782-789, 2020, DOI: 10.1109/LRA.2020.2965417.

[11] Yamada, T., Itoyama, K., Nishida, K. and Nakadai, K.: Sound Source Tracking by Drones with Microphone Arrays, *Proceedings of IEEE/SICE International Symposium on System Integration (SII2020)*, pp. 796-801, 2020, DOI: 10.1109/SII46433.2020.9026185.

[12] Shah, S. I. H., Bashir, S., Ashfaq, M., Altaf, A., Rmili, H.: Lightweight and Low-Cost Deployable Origami Antennas—A Review, *IEEE Access*, Vol. 9, pp.86429-86448, 2021, DOI: 10.1109/ACCESS.2021.3088953.

[13] Wang, R., Sun, J., Dai, S. J.: Design analysis and type synthesis of a petal-inspired space deployable-foldable mechanism, *Mechanism and Machine Theory*, Vol. 141, pp. 151-170, 2019, DOI: 10.1016/j.mechmachtheory.2019.07.005.

[14] Han, B., Xu, Y., Yao, J., Zheng, D., Guo, L., Zhao, Y.: Type synthesis of deployable mechanisms for ring truss antenna based on constraint-synthesis method, *Chinese Journal of Aeronautics*, Vol. 33, Issue 9, pp.2445-2460, 2020, DOI: 10.1016/j.cja.2019.07.015.

[15] 日本機械学会, JSME テキストシリーズ 機構学 機械の仕組みと運動, 2007.

[16] Schmidt, R. O.: Multiple emitter location and signal parameter estimation, *IEEE Transactions on Antennas and Propagation*, Vol. 34, No. 3, pp. 276-280, 1986, DOI: 10.1109/TAP.1986.1143830.

人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-063-12

# 狭空間におけるスピーカー・マイクロホンアレイを複数用いた人の位置・姿勢の音響計測の検討

# Acoustic measurement of human position and posture using multiple speakers and microphone arrays

工藤康一郎[1*]　干場功太郎[1]　岩附信行[1]
Koichiro Kudo[1]　Kotaro Hoshiba[1]　Nobuyuki Iwatsuki[1]

[1] 東京工業大学
[1]Tokyo Institute of Technology

**Abstract:** トイレや浴室といった狭空間における室内残響を考慮した人の状態推定について，これまで，スピーカーとマイクロホンアレイを用い，MUSIC (MUltiple Signal Classification) 法に基づき，狭空間内の残響をマッピングし，その変化で推定を行う手法を提案した．しかし，計測系の狭い指向性により，一次反射成分の強い床面や便器といった箇所に分布が偏るという問題が判明した．本稿では，スピーカーとマイクロホンアレイの増設，および配置の工夫によって指向性の改善を試みることでこの問題に取り組んだ．トイレを模した実験室において，様々な状況で測定を行い，得られたマッピング結果を比較することで最適な実験配置の評価を行った．

## 1　はじめに

近年の高齢化社会の進行に伴い，高齢者家庭や介護施設において転倒事故が多発している [1]．こうした事故は迅速に検知し対応する必要があり，そのための転倒検知モニタリングの重要性が高まっている．高齢者転倒事故の半数以上が住宅等居住場所で発生している点 [2]，常時モニタリングを行うには人力では負担が大きい点を考慮すると，自動的にモニタリングを行うことが望ましい．リビングや廊下等の通常の生活空間での転倒検知には，監視カメラの画像情報から人を認識し分類することで検知する手法 [3]，ウェアラブル端末の加速度センサを利用する手法，天井に設置した温度センサで床面積に占める人部分の割合を測定することで検知する手法 [4]，等を利用することができる．しかし，トイレや浴室といった空間では，プライバシーの観点からカメラの設置は難しく，ウェアラブル端末を着用したままトイレや浴室を利用することは負担となりうる．温度センサによる測定についても，トイレや浴室では人の出入りや使用状況により温度変化が激しく，精度の劣化が見込まれる．以上の理由から，これらの手法はこのような空間に適用するのが難しい．そこで，このような空間でのモニタリング・転倒検知手法として，音響信号を用いることに着目した．

音響信号による人のモニタリング・転倒検知には，これまで幾つかの手法が提案されている．Li らは転倒音の定位と識別により転倒検知する手法 [5]，Alanwar らはスマートスピーカーから計測音を照射し，人からの反射波の解析により在室状態を検知する手法 [6]，川部らは壁面に設置したスピーカーから計測音を照射し，部屋の音響モードから人の位置を推定する手法 [7] を提案している．しかし，これらの手法では，トイレや浴室といった狭空間で発生する残響の影響で精度が低下する可能性や，限定された場所のみでしか計測できないといった問題がある．本研究ではこれらの課題のうち，特に残響下での精度低下の問題に取り組むこととした．残響の音響信号処理には，Atmoko らの一般化相互相関法 [8]，Liu らの残響のモデル化 [9]，Birnie らの残響自体を推定に用いる手法 [10] などが提案されているが，狭空間での適用には複雑な残響への対応が難しいといった課題が残っている．

これらの問題を解決するため，狭空間での残響そのものを利用した人の状態モニタリングについて提案し，検証実験により，提案手法の性能について基礎検討を行った [11]．提案手法では，スピーカーから信号を照射し，複数のマイクロホンアレイを用いて，MUSIC (MUltiple SIgnal Classification) 法 [12] に基づき空間内の残響源のマッピングを行う．これまでの報告からは，空間内の状況や時間経過に伴い，得られる残響マップが変化することがわかった．しかし，一次反射成分の強い床面や便器といった箇所の分布が支配的になる問題があり，その原因として計測系の機器が持つ指向

*連絡先：東京工業大学　工学院　機械系
〒 152-8552　東京都目黒区大岡山 2-12-1 I1-27
E-mail:kudo.k.ad@m.titech.ac.jp

性が考えられた．

本稿では，スピーカーとマイクロホンアレイの増設によって指向性の改善を試みることでこの問題に取り組んだ．スピーカーを 1 台から 3 台，マイクロホンアレイを 2 台から 4 台に増やすことで指向性を広くし，残響源の分布の偏りの軽減を図る．また，スピーカー，マイクロホンアレイの配置による分布の変化についても検討を行う．トイレを模した実験環境で検証実験を行い，その性能の評価を行った．

## 2　提案手法

### 2.1　残響マッピング手法

人の状態モニタリング手法として，狭空間で発生する残響を利用した手法を提案した [11]．本手法では，まずスピーカーから信号を照射し，複数のマイクロホンアレイで室内で発生する反射波および残響を収録する．各マイクロホンアレイで収録された残響信号に対して音源定位を行い，それらを統合することで空間内の残響マッピングを行う．これは，異なる位置のマイクロホンアレイによる測定のマッピング結果を統合し，壁面の騒音源の位置を特定する Castellini らの手法 [13] に着想を得たものである．本手法のコンセプト図を図 1 に示す．スピーカーから照射された信号は，様々な反射経路を経由し，各マイクロホンアレイに入射する (図 1(a))．このとき，各マイクロホンアレイはそれぞれの位置に応じた残響信号を収録する．室内で収録される信号は，

$$h_{rec} = h_{dir} + h_{ref} + h_{rev} + n \tag{1}$$

のようにモデル化できる．ここで，$h_{rec}$ はマイクロホンによる測定信号，$h_{dir}$ はスピーカーから直接マイクロホンに入射する直接波，$h_{ref}$ は任意の場所での 1 回のみの反射でマイクロホンに到達する一次反射信号，$h_{rev}$ は残響信号，$n$ はノイズである．本手法では，残響信号 $h_{rev}$ に対して音源定位を行うことで残響信号の到来方向を推定し，その分布を把握する．このようにして得られた定位情報を図 1(b) のように統合することで，どの位置からの残響が強いのかといった情報をマッピングすることができる．この残響マップやその変化から，狭空間内の人の位置や姿勢といった状態を推定する．

音源定位手法には MUSIC 法 [12] を用い，各マイクロホンアレイにおいて空間スペクトルを算出し，統合する．アルゴリズムを以下に示す．

$i$ 番目のマイクロホンアレイにて得られた $M$ チャンネル入力音響信号の $f$ フレーム目をフーリエ変換して得られる $Z_i(\omega, f)$ から，以下のように相関行列 $R_i(\omega)$

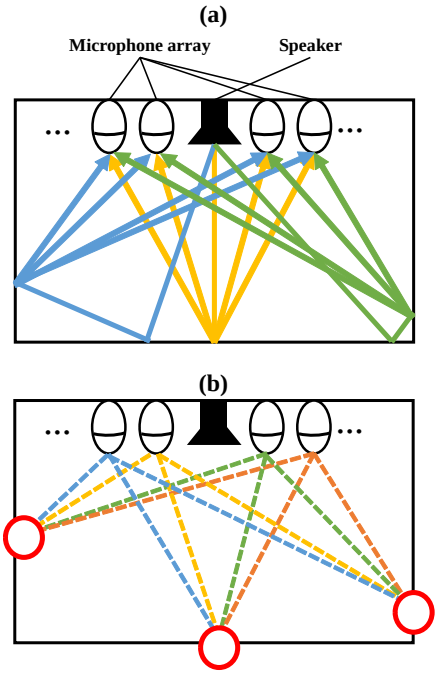



図 1: 提案手法のコンセプト図．(a) スピーカーから照射された信号の伝搬の様子，(b) 音源定位結果の統合による残響マッピング．

を得る．

$$R_i(\omega) = \frac{1}{f_2 - f_1 + 1} \sum_{f=f_1}^{f_2} Z(\omega, f) Z^*(\omega, f) \tag{2}$$

ここで，$\omega$ は周波数ビンのインデックス，$f_1$，$f_2$ は使用するフレームの開始と終了に対応したインデックス，$Z^*$ は $Z$ の共役転置である．得られた $R_i(\omega)$ を固有値展開して固有ベクトルを計算する．

$$R_i(\omega) = E_i(\omega) \Lambda_i(\omega) E_i^*(\omega) \tag{3}$$

$\Lambda_i(\omega)$ は降順に並べた固有値を対角成分に持つ行列，$E_i(\omega)$ は $\Lambda_i(\omega)$ に対応する固有ベクトルである．これと，マイクロホンアレイ座標系での方向 $\psi$ に対応した伝達関数 $G(\omega, \psi)$ を用いて，空間スペクトル $P_i(\omega, \psi)$ を計算する．

$$P_i(\omega, \psi) = \frac{|G^*(\omega, \psi) G(\omega, \psi)|}{\sum_{m=L+1}^{M} |G^*(\omega, \psi) e_{i,m}(\omega, \psi)|} \tag{4}$$

ただし，$L$ は目的音源数，$e_{i,m}$ は $E_i$ に含まれる $m$ 番目の固有ベクトルを表す．こうして得られた $P_i(\omega, \psi)$ を周波数 $\omega$ 方向に平均し，$\bar{P}_i(\psi)$ を得る．

$$\bar{P}_i(\psi) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} P_i(\omega, \psi) \tag{5}$$

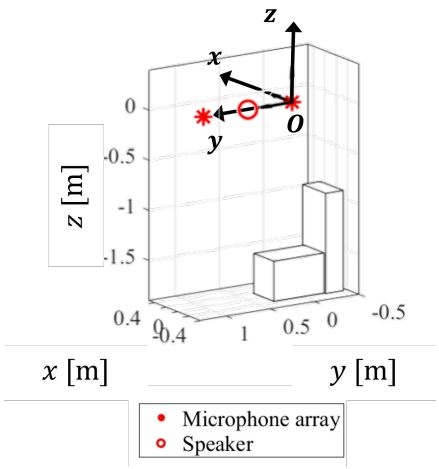



図 2: これまでの実験配置．スピーカーの座標は $(x,y,z)$=(0 m, 0.5 m, 0 m)，マイクロホンアレイの座標は $(x,y,z)$=(0 m, 0 m, 0 m), (0 m, 1 m, 0 m).

なお，$\omega_H$，$\omega_L$ は使用する周波数ビンの上限と下限に対応したインデックスである．得られた空間スペクトル $\bar{P}_i(\psi)$ を，三次元座標 $w=(x,y,z)$ に投影し，$\bar{P}_i(w)$ へと座標変換を行う．これらを $I$ 個のマイクロホンアレイに対して算出し，空間スペクトルを足し合わせて合成空間スペクトル $\bar{P}_{sum}(w)$ を得る．

$$\bar{P}_{sum}(w) = \sum_{i=1}^{I} \bar{P}_i(w) \tag{6}$$

得られた合成空間スペクトルは，各位置から到来する信号の強度を表しており，これにより空間内の残響源をマッピングすることができる．

## 2.2 問題点と解決法

これまでの報告 [11] では，基礎的な検証を行うため，トイレを模した狭空間実験室で，図 2 のように 1 台のスピーカーと 2 台のマイクロホンアレイを設置して実験を行った．実験室の寸法は 0.9×1.8×1.9 m となっており，上部は開放されている．スピーカーには Bang&Olufsen 社製の Beosound A1 (図 3(a))，マイクロホンアレイにはシステムインフロンティア社製の 8ch マイクロホンアレイ TAMAGO(図 3(b)) を使用した．基礎検証の結果，一次反射成分の強い床面や便器といった箇所の分布が支配的になる問題が判明した．この問題の原因のひとつとして，スピーカーの狭い指向性によって残響の到来方向が限られてしまっていることや，式 (6) において空間スペクトルを加算する際のマイクロホンアレイの個数が少ないために得られる残響マップもマイクロホンアレイの位置に偏った結果になっていること


図 3: 実験にて使用した機器．(a) スピーカー (Beosound A1)，(b) 8ch マイクロホンアレイ (TAMAGO).

表 1: 各実験配置における各スピーカー，マイクロホンアレイの $y$ 座標の値 [m]．位置．(a)〜(d) は図 5 の (a)〜(d) に対応.

| Device・Arrangement | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| Speaker 01 | 0.5 | 0.75 | 1 | 1.2 |
| Speaker 02 | - | 0.5 | 0.5 | 0.5 |
| Speaker 03 | - | 0.25 | 0 | -0.25 |
| Microphone Array 01 | 1.1 | 1.21 | 1.21 | 1 |
| Microphone Array 02 | 0.8 | 1 | 0.75 | 0.75 |
| Microphone Array 03 | 0.2 | 0 | 0.25 | 0.25 |
| Microphone Array 04 | -0.1 | -0.25 | -0.25 | 0 |

が考えられる．そこで，スピーカーとマイクロホンアレイを増設することで，照射される信号の指向性や空間スペクトルの偏りを解決することを試みた．

# 3 検証実験

## 3.1 実験環境の改良

本稿においても，これまでの報告 [11] と同様に，トイレを模した実験環境にて，計測用信号を照射するスピーカーと室内残響を収録するマイクロホンアレイからなる計測系で検証を行った．実験環境の寸法を図 4 に示す．判明した問題点を踏まえ，スピーカーとマイクロホンアレイの台数を増やした実験を行い，マッピング結果の変化を検討する．測定機器は図 3 に示したものを引き続き使用し，スピーカー 1 台あるいは 3 台とマイクロホンアレイ 4 台を室内で対称的になるように 4 通りに配置し実験を行った．

4 通りの実験配置を図 5 に示す．各スピーカー，マイクロホンアレイは，y 軸と一致する位置に設置したポールに沿って配置しているため，各機器の $x$ 座標，$z$ 座標はすべて 0 である．各機器の y 座標の値は表 1 に示す．

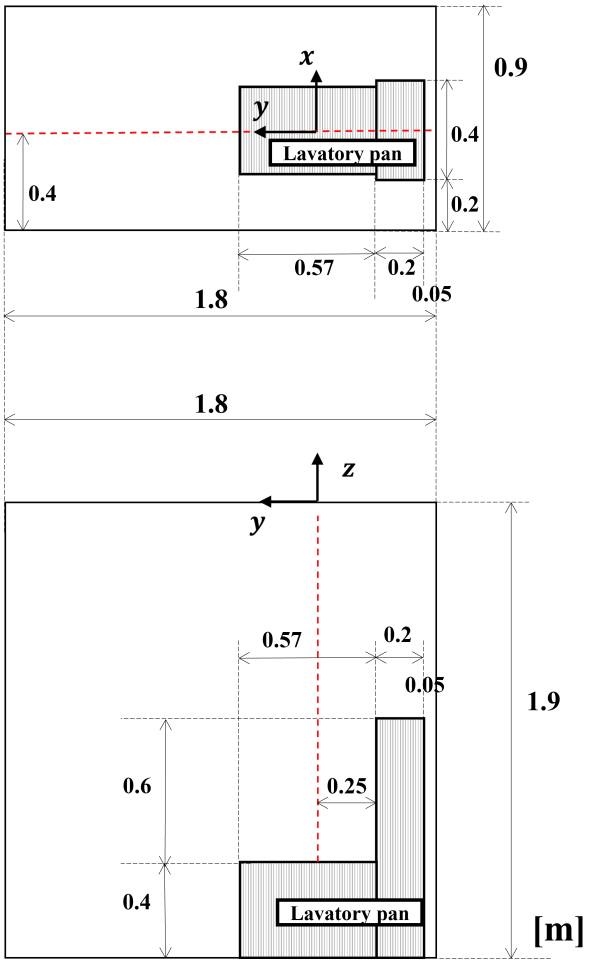



図 4: 実験環境の寸法.

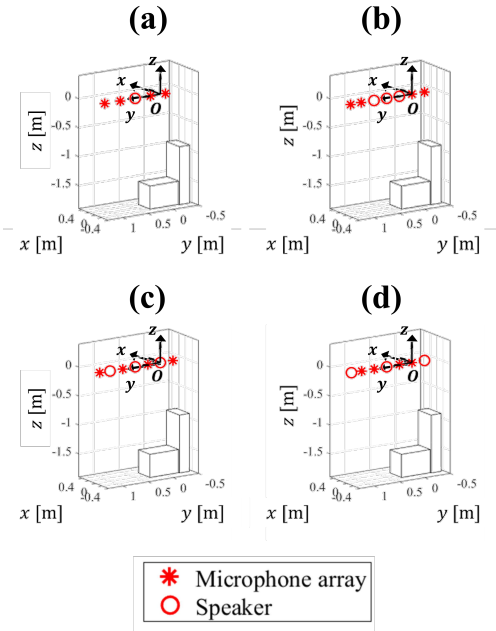



図 5: 4 通りの実験配置.

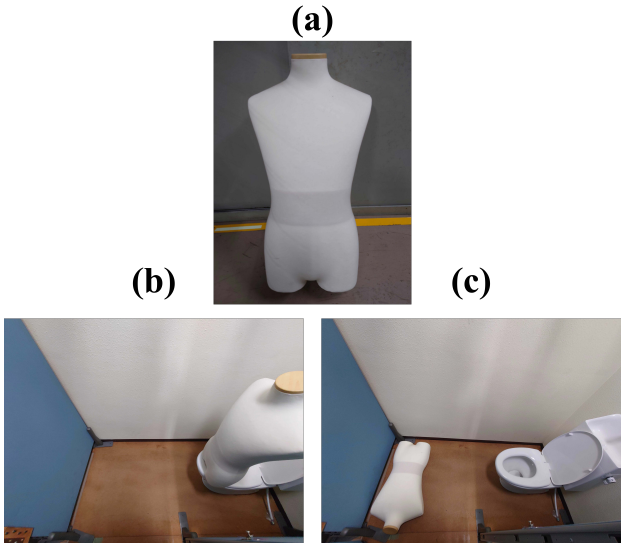



図 6: 実験時のトルソーの配置. (a) 使用したトルソー, (b) 着席模擬時の配置, (c) 転倒模擬時の配置.

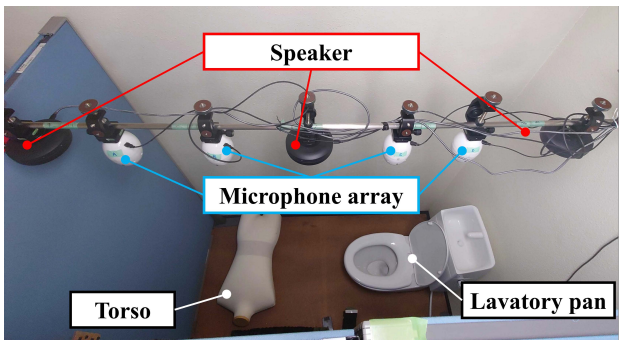



図 7: 転倒模擬時の実験状況の一例. 実験配置は図 5(d).

スピーカーから照射する信号には 0-8 kHz のチャープ信号を用いた. 各マイクロホンアレイでは, サンプリング周波数 16 kHz, 量子化ビット数 24 bit で音響信号が収録される. 収録された信号に対し, 音源信号の畳み込みによるパルス圧縮を行い, インパルス応答を得る. このインパルス信号に対し提案手法を適用し, 解析を行った. 式 (4) における伝達関数 $G$ は, 幾何計算により 1 deg. 刻みで算出したものを用いた. 部屋内部の状況に応じた残響マップの変化を観察するため, 人の代替ターゲットとして, 幅 0.3 m× 高さ 0.7 m× 奥行 0.2 m のトルソー (図 6(a)) を空間内に配置した. 空間内に便器のみがある空室時, 便器座面上にトルソーを配置した着席模擬時 (図 6(b)), 床面にトルソーを配置した転倒模擬時 (図 6(c)) の 3 つの室内状況を設定し, 音響信号を収録した. 実験状況の一例を図 7 に示す. それぞれの状態で 11 回の計測, 解析を行い, 外れ値を含む結果を除いた残響マップを平均化し評価を行った.

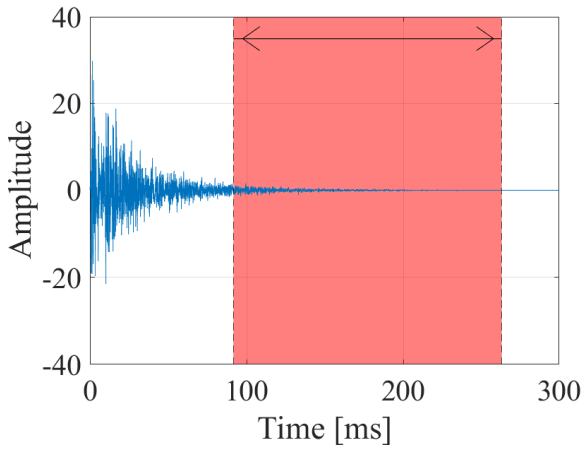



図 8: 得られたインパルス応答．赤い範囲が解析区間である．

### 3.2 実験結果

計測信号のパルス圧縮によって得られたインパルス応答の一例を図 8 に示す．0 ms 付近に見られるピークがスピーカーからマイクロホンに直接到達した直接波 $h_{dir}$ であり，10 ms 付近で床面からの一次反射 $h_{ref}$ が到達，その後残響 $h_{rev}$ が続いている．

残響のマッピングを行うため，得られたインパルス応答に対し，残響成分のみが観察できる区間を解析区間として設定する必要がある．これまでの報告から，直接音と一次反射の影響が残る区間で解析を行うと，これらの影響により残響の観測が難しくなるという問題点が判明している．そこで，直接波と一次反射の減衰を考慮し，解析区間を直接波到来から 90〜260 ms とした（図 8）．MUSIC 法における解析周波数については，予備実験により得られた空間スペクトルの平均値から，$\omega_L = 500$ Hz，$\omega_H = 1000$ Hz の範囲とした．周波数領域を低周波域に限定することで，直進性の強い高周波域における局所ピークの発生を低減できる．

残響マッピングの結果を図 9 に示す．マッピングの座標系は図 5 に示したものと等しく，図 5 における奥側の 2 壁面，及び底面のマッピングを行った．(a)〜(d) は図 5 の実験配置 (a)〜(d) に対応した結果，(e) は図 2 の実験配置の結果である．また，(i) は空室時，(ii) は着席模擬時，(iii) は転倒模擬時の結果である．カラーマップにて各位置から到来する信号の強度を示している．提案手法ではマイクロホンアレイ毎の空間スペクトルを加算してマップを作成しているが，(a)〜(d) と (e) ではマイクロホンアレイの数が異なり，加算した場合に絶対値での比較に影響が出るため，マイクロホンアレイの個数に応じた補正倍率を乗算している．

これまでの報告で判明した残響マップの傾向は，(i) 空室時は底面全域で強度の高い分布，(ii) 着席模擬時はトルソーにより残響が吸収され便器側の強度が低下することで相対的に床面側で強度が増加，(iii) 転倒模擬時はトルソーによる吸収で床面側の成分量が低下，便器側が相対的に強く観測されるが，全体としての強度は低下，といったものだった．

本稿ではマイクロホンアレイを 2 つから 4 つに増やしているため，誤った位置で空間スペクトルが強め合うこと (エイリアス) が低減し，より正しい残響マップになったと考えられる．

それぞれの結果について観察する．スピーカーの数はそのままにマイクロホンアレイ数を増やした (a) では，(e) と近い傾向が観測されたが，床面側の分布の強度が低下している．これは，スピーカーの位置が (e) と同じであり，得られる残響場も (e) に近くなるが，エイリアスが低減されたことで全体の強度が低下したと考えられる．しかし，照射する信号の指向性は改善されていないため，得られる残響源は偏ってしまう．

中心にスピーカーを増やした (b) では，音源中心の位置は変えず，照射レベルを上げる配置である．結果，従来の傾向を強調したような分布となった．床面側の分布強度が大きくなっていることが分布形状における (e) との差異である．本実験環境では，部屋の上部から信号を照射，収録しているため，差異が出やすいのは側面側よりも底面側である．そのため，照射レベルが上がった本配置において，従来と近い傾向を示しつつ，床面側で状況毎の差異が強調されるのは予想通りといえる．

スピーカーとマイクロホンアレイが交互に配置された (c) では，従来の傾向には沿っているが，空室時の分布が便器側に偏っており，着席模擬時に床面側で強い分布，転倒模擬時で再度便器側で強い分布，という結果だった．便器側の残響分布の影響が強まっていることから，(a)，(b) や (e) と比べ，照射する信号の指向性が広くなり，残響が全体に広がるような配置となったことが反映された結果であるといえる．

スピーカーを両端と中心に配置した (d) では，最も照射する信号の指向性が広いため，どの状況でも残響が全域で満遍なく観察されると予想されたが，それに反して状況毎の差異が大きいマッピング結果となった．特に，着席模擬時にトルソーの影響による相対的な床面側の強調の効果が非常に強い．理由としては，指向性の改善による影響について，残響が全域に広がって均一化される効果よりも，残響が発生する範囲の増加によって状況毎の差異が検出しやすくなる効果の方が強くなっており，差異が際立ったのではないかと考えられる．

以上のように，残響マッピング結果からは，スピーカーとマイクロホンアレイの増設はその位置に応じたマッピング結果の変化をもたらすことが観察でき，指向性の改善についても分布形状からその効果を確認す

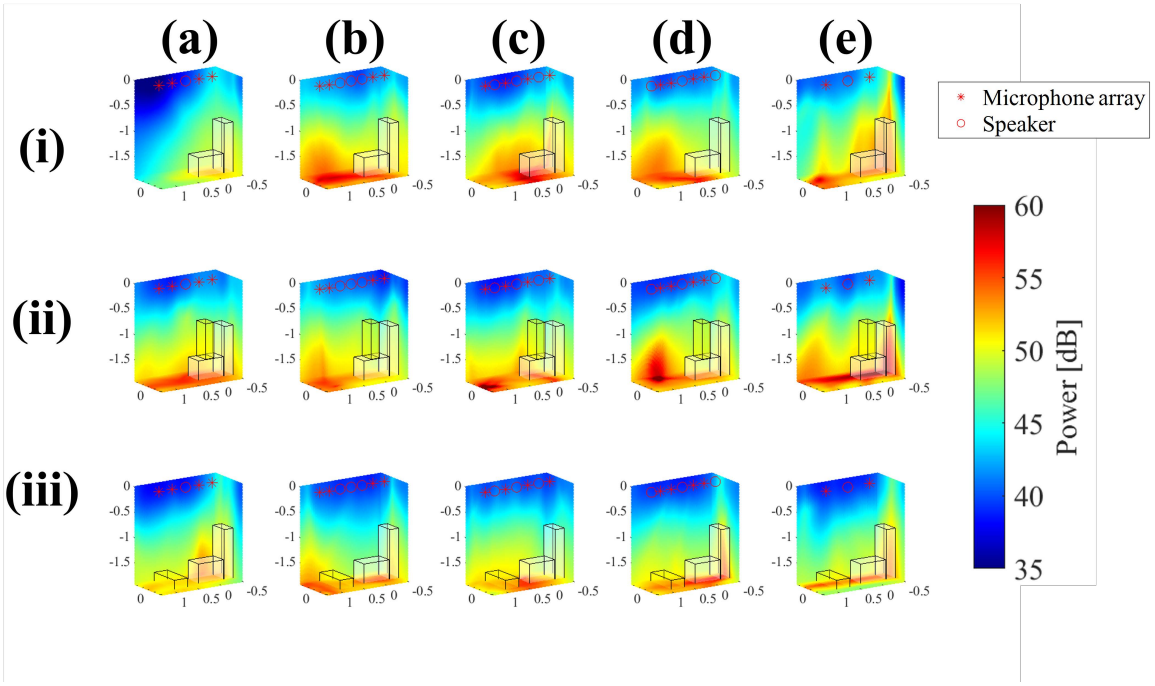



図 9: 得られた残響マップ．(a)〜(d) は図 5 の実験配置 (a)〜(d)，(e) は図 2 の実験配置に対応した結果である．(i) 空室時，(ii) 着席模擬時，(iii) 転倒模擬時．

ることができた．

## 4　考察

得られた残響マッピング結果について考察する．前章により，スピーカーおよびマイクロホンアレイ増設の効果を確認できたため，本章では図 9(b)〜(d) の比較を行う．室内状況を区別する性能を測るために，着席模擬時，及び転倒模擬時に対し，空室時との残響マップの差分を指標として評価する．前述の通り，今回の実験配置では状況毎の差異は底面に出やすく，壁面側には観測されづらいため，評価には底面側の差分のみを算出した．図 10 に算出した差分を示す．(b)〜(d) は図 5 の実験配置 (b)〜(d) に対応した結果である．また，(ii)-(i) は着席模擬時と空室時の，(iii)-(i) は転倒模擬時と空室時の差分である．カラーマップにて差分の強度を表している．単純な差分であるため，残響分布の強度が全体的に低下していた場合に，分布傾向が強調されていない場合もあるが，(d) の場合，(ii)-(i) では便器側のトルソーによる成分低下と床面側の相対的な上昇，(iii)-(i) では床面側のトルソーによる成分低下と便器側の相対的な上昇をどちらも抽出できている．以上から，差分においては (d) がもっとも良い結果であるといえる．

MUSIC スペクトルの差分を統計的な指標で評価する．図 10 に示した残響マップの差分のヒストグラムを

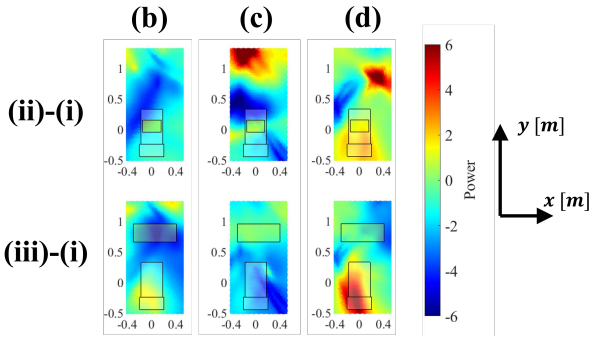



図 10: 底面の残響マップの差分．(b)〜(d) は図 5 の実験配置 (b)〜(d) に対応．(ii)-(i) 着席模擬時と空室時の差分，(iii)-(i) 転倒模擬時と空室時の差分．

図 11 に示す．(b)〜(d) は図 5 の実験配置 (b)〜(d) に対応した結果である．また，(ii)-(i) は着席模擬時と空室時の差分，(iii)-(i) 転倒模擬時と空室時の差分に対するヒストグラムである．横軸は差分の強度，縦軸は頻度である．さらに，得られたヒストグラムのパラメータを図 12 に示す．(A) は差分の絶対値の平均値，(B) が標準偏差，(C) が尖度，(D) が歪度である．(A) の差分の絶対値の平均は，全体的な差異の大きさを評価することができる．この結果は，スピーカー間の距離が離れるにつれ差分の値が低下していく傾向を概ね示しているといえる．これはスピーカーの距離が離れてい

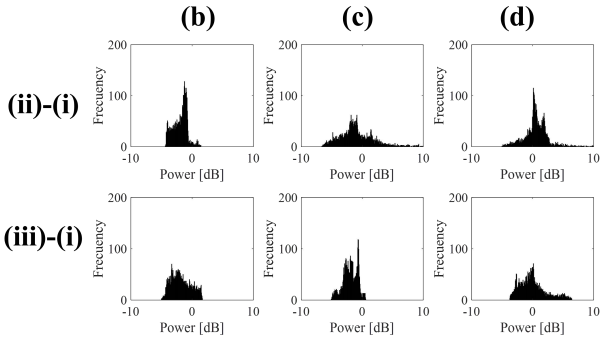



図 11: 残響マップの差分のヒストグラム．(b)〜(d) は図 5 の実験配置 (b)〜(d) に対応．(ii)-(i) 着席模擬時と空室時の差分，(iii)-(i) 転倒模擬時と空室時の差分．

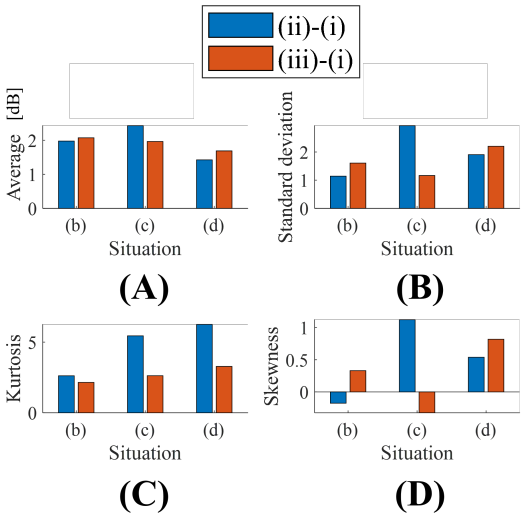



図 12: 図 11 の統計量を比較するグラフ．(b)〜(d) は配置 (b)〜(d) に対応しており，(ii)-(i) 着席模擬時と空室時の差分，(iii)-(i) 転倒模擬時と空室時の差分．残響マップの各点における MUSIC スペクトルの差分の（A）絶対値の平均値，（B）標準偏差，（C）尖度，（D）歪度

るほど指向性が改善され，様々な方向から残響が到来することで室内状況の差異が場所により偏りにくくなることを意味する．トルソーが配置されていない部分では，ランダムな残響がどの状況でもある程度到来するために差異が均されるはずであり，配置と指向性改善の関係に沿った結果であるといえ，改善の効果が確認できた．(B) の標準偏差は，差分のばらつきを示すため，正と負の差分がある (c) の (ii)-(i) と (d) で大きな値を示した．状況毎の差分はある程度高い値で，かつばらつきがあることが望ましいため，どちらの差分でも高い値を示した (d) が有利であると考えられる．(C) の尖度は，どちらの差分でも (b)<(c)<(d) となった．これは差異が少ない部分がどれだけ多いかを意味し，(A) から観察された指向性改善に伴う差が均される効果をこちらでも確認できる．(D) の歪度と併せて考慮すると，(b) は差分の分布が一部に集中しており室内状況の判別が難しく，(c)，(d) は差分の分布が広く，かつふたつの差分で違った傾向を示している為判別しやすいといえる．

以上の結果から，前回の課題である指向性を改善するとともに，状況判別を残響マップの差分によって行う上で最も有利なのは (d) の配置である，と結論付けた．今後の課題として，壁沿いに人が立っているなどの壁面側に人の影響が出ると思われるであろう状況の評価や，マッピング結果を解析的に位置と対応させることで状況推定の精度向上や位置推定につなげることが挙げられる．

## 5　結論

本稿では，残響の影響が強いトイレなどの狭空間における人の状態モニタリングを目的とし，空間内残響マッピングによって人の状態推定を行う手法の改良を行った．スピーカーとマイクロホンアレイを増設することで指向性を改善し，残響マッピングの精度の向上を試みた．検証実験の結果，スピーカーやマイクロホンアレイの配置の変更によるマッピング結果の差異やその特性を観察することができ，指向性の改善についても効果が確認された．マッピング結果の差分を用いた評価からは，3 台のスピーカー，4 台のマイクロホンアレイを用いた場合の最適な配置が明らかになった．今後は，室内状況のバリエーションを増やし，マッピング結果との対応関係をより明確化していくとともに，実際の室内状況推定を実施するための分類手法を模索していく．

## 参考文献

[1] 公益財団法人介護労働安定センター:”「介護サービスの利用に係る 事故の防止に関する調査研究事業」報告書”, https://www.kaigo-center.or.jp/report/pdf/h30_kaigojiko_houkoku_20180402.pdf

[2] 東京消防庁:　”救急搬送データからみる高齢者の事故”, https://www.tfd.metro.tokyo.lg.jp/lfe/topics/nichijou/kkhansoudeta.html

[3] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang and J. Chambers: ”A Posture Recognition-Based Fall Detection System for Monitoring an Elderly Person in a Smart Home Environment”, IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 6, pp. 1274–1286, 2012.

[4] Y. Ogawa, K. Naito: "Fall detection scheme based on temperature distribution with IR array sensor", Proceedings of IEEE International Conference on Consumer Electronics (ICCE), pp. 1-5, 2020.

[5] L. Yun, K. C. Ho, M. Popescu: "A microphone array system for automatic fall detection", IEEE Transactions on Biomedical Engineering, pp. 1291–1301, 2012.

[6] A. Alanwar, B. Balaji, Y. Tian, S. Yang, M. Srivastava: "EchoSafe: Sonar-based verifiable interaction with intelligent digital agents", Proceedings of the 1st ACM Workshop on the Internet of Safe Things, pp. 38–43, 2017.

[7] 川部, 和田, 中村: "浴室内における人の有無などの検知 －音響特性の計測による室内状態の検知 (IV) －", 日本音響学会 2021 年春季研究発表会講演論文集, pp. 581–582, 2021.

[8] H. Atmoko, D. C. Tan, G. Y. Tian, B. Fazenda: "Accurate sound source localization in a reverberant environment using multiple acoustic sensors", Measurement Science and Technology, vol. 19, no. 2, pp. 1–10, 2008.

[9] Z. Liu, R. Chen, F. Ye, G. Guo, Z. Li, L. Qian: "Improved TOA estimation method for acoustic ranging in a reverberant environment", IEEE Sensors Journal, vol. 22, no. 6, pp. 4844–4852, 2022.

[10] L. I. Birnie, T. D. Abhayapala, P. N. Samarasinghe: "Reflection Assisted Sound Source Localization Through a Harmonic Domain MUSIC Framework", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 279–293, 2020.

[11] 工藤，干場，岩附: "複数のマイクロホンアレイを用いた狭空間における人の位置・姿勢の音響計測に関する基礎検討", 第 41 回日本ロボット学会学術講演会講演予稿集, RSJ2023AC1D3-04, 2023.

[12] R. O. Schmidt: "Multiple Emitter Location and Signal Parameter Estimation", IEEE Transactions on Antennas and Propagation, vol. 34, no. 3, pp. 276–280, 1986.

[13] P. Castellini, A. Sassaroli, A. Paonessa, A. Peiffer, A. Roeder: "Average beamforming in reverberant fields: Application on helicopter and airplane cockpits", Applied Acoustics, vol. 74, issue 1, pp.198–210, 2013.